Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches

SARNAB BHATTACHARYA*, University of Texas at Austin, USA
REBECCA ADAIMI*, University of Texas at Austin, USA
EDISON THOMAZ, University of Texas at Austin, USA

Automatically recognizing a broad spectrum of human activities is key to realizing many compelling applications in health, personal assistance, human-computer interaction and smart environments. However, in real-world settings, approaches to human action perception have been largely constrained to detecting mobility states, e.g., walking, running, standing. In this work, we explore the use of inertial-acoustic sensing provided by off-the-shelf commodity smartwatches for detecting activities of daily living (ADLs). We conduct a semi-naturalistic study with a diverse set of 15 participants in their own homes and show that acoustic and inertial sensor data can be combined to recognize 23 activities such as writing, cooking, and cleaning with high accuracy. We further conduct a completely in-the-wild study with 5 participants to better evaluate the feasibility of our system in practical unconstrained scenarios. We comprehensively studied various baseline machine learning and deep learning models with three different fusion strategies, demonstrating the benefit of combining inertial and acoustic data for ADL recognition. Our analysis underscores the feasibility of high-performing recognition of daily activities using inertial-acoustic data from practical off-the-shelf wrist-worn devices while also uncovering challenges faced in unconstrained settings. We encourage researchers to use our public dataset to further push the boundary of ADL recognition *in-the-wild*.

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

Additional Key Words and Phrases: Multimodal classification, Audio Classification, Sound Sensing, Motion sensing, Gesture Recognition, Wearable, In-the-wild, Dataset, Smartwatch, Activities of Daily Living, Human Activity Recognition

ACM Reference Format:
Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 42 (June 2022), 28 pages. https://doi.org/10.1145/3534582

1 INTRODUCTION

Detecting human behaviors in the home has been one of the driving forces of the field of human activity recognition (HAR) for many years [4, 34, 36, 50, 53]. Indeed, HAR is key to realizing ubiquitous computing applications, such as home automation and personal assistance, and provides a foundation for new types of human-computer interactions that hinge on anticipating and responding to people's needs. Due to their relevance to health and well-being, Activities of Daily Living (ADLs) have gained special attention from HAR researchers;

*Both authors contributed equally to this work.

Authors' addresses: Sarnab Bhattacharya, sarnab2008@utexas.edu, University of Texas at Austin, Austin, Texas, USA; Rebecca Adaimi, rebecca.adaimi@utexas.edu, University of Texas at Austin, Austin, Texas, USA; Edison Thomaz, ethomaz@utexas.edu, University of Texas at Austin, Austin, Texas, USA.

2474-9567/2022/6-ART42 $15.00
https://doi.org/10.1145/3534582

ADLs represent essential tasks and skills necessary for an individual to meet their basic physical needs, such as cooking, cleaning, and grooming. Objectively measuring how ADLs are performed can be an effective means of health assessment, with the potential to extend aging in place and as a way to monitor the progress of debilitating medical conditions such as cognitive impairment. However, the recognition of these types of activities in real-world environments continues to be a challenge; so far, HAR methods that generalize beyond the laboratory have been largely constrained to detecting mobility states, e.g., walking, running, standing [16, 17].

Over the last decade, the popularization of smartphones has brought sensing and computation closer to the everyday human experience, enabling new HAR methods for ADL recognition that complement existing approaches based on environmental sensors. Underlying most of this work has been the assumption that a user's phone is always at-hand, serving both as a conduit for data collection and proxy for human behavior recognition. However, empirical investigations conducted by Patel *et al.* in 2006 [42] and then again by Dey *et al.* in 2011 [15] showed that this assumption is only true 50% of the time, indicating that smartphones are not always as close to individuals as widely believed. This finding, and a new wave of emerging wearable technologies has encouraged researchers to look beyond the smartphone.

In this work, we explore the potential of stand-alone, commodity smartwatches in recognizing ADLs in naturalistic settings with multimodal sensing. Smartwatches open up exciting new opportunities in this field. Many ADLs such as eating, writing, and cooking are characterized by specific hand and wrist gestures, so the ability to collect sensor data from a location close to the hand is highly desirable. Additionally, smartwatches can capture ambient sounds that are in close proximity to the body and thus highly contextualized to people's activity. And from a human-centered perspective, smartwatches do not carry the stigma of specialized sensing devices; they are comfortable to wear, are socially acceptable and have become increasingly popular in recent years.

While prior studies have featured smartwatches in HAR research, our research adds to the body of work in the field in important ways. First and foremost, we place the smartwatch front and center, leveraging its capabilities as a stand-alone device; this is different from previous work where the smartwatch played a supporting role to the smartphone [20, 49, 55]. Secondly, this paper targets ADL recognition (i.e., complex human behaviors) using multimodal sensing, i.e., inertial and acoustic; prior research has been largely focused on the recognition of specific gestures and activities [9, 13, 48, 59], typically leveraging only the inertial sensors on the watch to track movement and motion. And crucially, our work emphasizes ADL recognition in the real-world. We captured two datasets, one in people's own homes and another in completely naturalistic settings in order to develop and validate approaches for smartwatch-centric HAR. The specific contributions of this work are:

- Two annotated datasets[1] capturing synchronized inertial and acoustic data collected from an off-the-shelf smartwatch. One dataset consists of data captured as 15 participants performed various activities of daily living in their own homes; the other dataset was compiled from 5 participants performing activities completely *in-the-wild* and without any supervision; ground truth was established from video evidence captured with a wearable camera.
- A comprehensive quantitative evaluation with various modeling approaches and fusion methods that demonstrates how fusion of inertial and acoustic smartwatch data is beneficial for recognizing ADLs, especially for activities that have minimal motion pattern but unique sound patterns and vice versa. A baseline framework is presented and shown to recognize these activities with an average macro F1-score of 89.7% in a Leave-One-Participant-Out (LOPO) performance evaluation, and up to 94.3% with a personalized model fine-tuned for each participant on the *semi-naturalistic* dataset. Evaluation *in-the-wild* achieved 30.0% macro F1-score and 55.8% weighted F1-score, demonstrating the difficulty of recognizing ADLs in

[1] https://doi.org/10.18738/T8/NNDFQD

unconstrained environments. We characterize the various challenges that arise when dealing with real-world settings and encourage researchers to use our dataset to develop and explore methods to mitigate these difficulties and improve ADL recognition *in-the-wild.*

2 RELATED WORK

A variety of sensing modalities have been explored in human activity recognition. For example, vision-based recognition was employed in early behavior monitoring systems in the home [7] and more recently with egocentric imaging and depth cameras [30, 35, 45]. Electromagnetic sensing is another modality that has been explored with promising results [5, 62]. In a recent study by Patterson *et al.*, objects with RFID tags and a glove with an RFID reader were used to track object usage to determine fine-grained activities being performed [43]. Obtrusiveness and excessive instrumentation are major drawbacks in such methods, because the required sensors are not available in common consumer devices like smartwatches. In this work, we focus on leveraging unmodified commodity smartwatches using inertial and acoustic data. Therefore, we highlight relevant prior research that explores these modalities and explain what sets apart our work from prior research.

2.1 Inertial Sensing

Inertial sensors have been used in human activity research for quite some time. Before the advent of ubiquitous sensing platforms like smartphones and smartwatches, custom designed systems were used. RecoFit [37] used a platform which used a inertial sensor on the forearm to detect repetitive activities for exercise tracking. The work by Moschetti *et al.* [38, 39] used rings and bracelets with inertial sensors to detect activities ranging from mobility states to more complex activities like eating, teeth brushing, etc.

Activity recognition using inertial data from a smartwatch has been explored recently by many researchers. In the work conducted by Weiss *et al.*, activity recognition performance between a smartwatch and a smartphone was compared using motion data in a study with 17 participants and 18 activities [59]. The data collected using smartwatches resulted in higher activity recognition performance compared to the data collected using smartphones, concluding that smartwatches are more proficient in collecting data for activity recognition. Further research in activity recognition by Filippoupolitis *et al.* enhanced motion data from smartwatches by fusing location information from BLE beacons in indoor environments [18]. The addition of location information lent context to the activities, but required instrumentation of the environment. In a study by Reiss *et al.*, inertial data was collected from 9 participants performing 12 acivities[47]. This dataset was released as the PAMAP2 dataset. While activity recognition was achieved with greater accuracy, the activities were more geared towards movement and physical activities rather than ADLs. In a multimodal model created by Ma *et al.* [33], an attention-based layer was implemented to learn the weight of each data stream for different activities. Public datasets that do not explicitly deal with ADLs were used, such as Skoda and PAMAP2. Bhattacharya *et al.*[10] focused on investigating the benefits of integrating RBM based models on a bleeding edge smartwatch platform, highlighting the limits of model complexity possible while maintaining acceptable energy and execution time. Finally, in research by Laput *et al.*, a smartwatch with a modified software kernel was used to sense 25 fine grained hand activities [29]. The motion data was sampled at 4KHz and inference was performed offline using a CNN trained on the spectrograms of the motion data. Our method is similar in that we wanted to push the limits of commercial smartwatches but without any major modification in its operating system. Instead, we augmented motion data with audio data captured in close proximity to the activity being performed.

2.2 Acoustic Sensing

Audio has often been used as a high-fidelity source of data for context understanding [14]. Despite drawbacks like the incoherence caused by background noise and other acoustic artifacts, audio is a rich source of data for

HAR. Notably, it offers several advantages over inertial sensors. Audio data is often directly comprehensible by humans, greatly simplifying the labeling process; annotations can be obtained by simply listening to the audio. Indeed prior research has often used audio as an input to recognize activities and corresponding contexts. In work by Stork *et al.*, Non-Markovian Ensemble Voting was used to recognize 22 activities from audio[52]. In further work by Ubicoustics, training on public available datasets and testing on real world sounds performed well across multiple sound capture devices [28]. ADLs were not emphasized in the analyses but some of the activity recognition classes included common household tasks. Encouraging results were demonstrated by Liang *et al.* using transfer learning for ADL recognition in homes with large scale audio embeddings from YouTube videos [32]. More recently, Adaimi *et al.* investigated how background sounds captured in voice interactions with conversational assistants can be a rich source of context and be used for ADL recognition [4].

2.3 Multimodal Sensing

The idea of fusing motion and audio data in human activity recognition is not new, as seen, for example, in earlier efforts by Minnen et al [36]. In this work, multiple accelerometers and microphones were placed at specific locations on the body and were focused on detecting repetitive periodic activities. More recently, research by GestEar[9] demonstrated motion and audio data fusion from smartwatches to classify gestures. This research was limited by a narrow set of activities, which included only plain gestures and simplified the recognition pipeline. Kim *et al.* also investigated combining accelerometer and acoustic data collected from a single off-the-shelf smartwatch for recognizing a set of 5 activities (sleeping, eating, vacuuming, TV-watching, and showering) [25]. Their approach required a three-stage collaborative classifier that integrated an ensemble of classifiers and a ground-truth mapping table for reliable prediction. While shown effective, their approach was evaluated on a limited data of 3 participants performing only 5 activities. Other work by TapSkin [63] also used motion and sound data to recognize taps on 11 different locations around the wrist. The work by Ward et al [58] also explored using motion and audio from multiple sources to recognize activities done in the workshop. Apart from fusing inertial and acoustic data, multiple prior work has focused on fusing a variety of other sensor data streams in order to increase HAR performance. Fusion of inertial data from multiple IMUs has been studied in [40, 61]. Radu *et al* explored multiple methods of sensor data fusion accross multiple datasets containing both inertial data as well as physiological data [46].

In other closely related work, the ExtraSensory dataset collected by Vaizman *et al.* [55] and subsequent work on multimodal, multi-label classification [56] used data streams from both smartwatches and smartphones, drawing from a large data set collected from 60 participants. However, this work did not focus on ADLs and audio was not collected from a smartwatch. Additionally, because location was among the sensor data used in the project, many of the activities were location-based and listed as "at work" or "at home". Moreover, relying on self-reporting for data annotation resulted in missing labels which can hinder model training. On the other hand, our label acquisition process for both *semi-naturalistic* and *in-the-wild* datasets provides more reliable annotations. In very recent work by Siddiqui and Chan [51], an array of 10 microphones and an inertial motion unit (IMU) in a wrist-mounted setup were evaluated to recognise hand gestures, achieving good performance without using deep models for classification.

Compared to prior work, our work is different in several important ways. First, our approach is not reliant on placing obtrusive sensors in the environment or on the human body. Second, we explore existing capabilities of an unmodified commodity smartwatch to capture inertial and acoustic data streams. Finally, our work includes a large set of activity classes (23 activities) captured in *semi-naturalistic* environments in people's own homes rather than a controlled lab setting as well as *in-the-wild.* Our label acquisition strategy provides a more accurate data annotation and synchronization of sensors with remote supervision and action cues in the *semi-naturalistic* study and first-person snapshots using a phone around the neck in the *in-the-wild* study.

Fig. 1. 23 activity classes of daily living captured using a commodity smartwatch. Each activity is labeled with a letter (top right) for easier reference.

3 DATA COLLECTION

This section presents the dataset creation that was done via two IRB-approved user studies: (1) a *semi-naturalistic* supervised study with 15 participants performing a set of daily activities in their own homes and (2) an *in-the-wild* free-living study with 5 participants. We first present the hardware setup used for data collection and then describe the study protocol followed for each study. The datasets are publicly available for researchers to use for further exploration and developing future ADL recognition systems.

3.1 Hardware Setup

Fossil Gen 4, an off-the-shelf smartwatch running the Android Wear OS 2.12, was used to collect data from participants. It was fitted with the Snapdragon 3100 processor, 4GB of storage, and 786MB of RAM. Accelerometer, gyroscope, and microphone sensors already present on the watch were used. A custom Android application was designed to collect acoustic and inertial data synchronously and store it locally on the watch. The maximum inertial sensor sampling rate supported by the smartwatch was 50Hz, while acoustic data was sampled at 22.05KHz. Given that the aim is to capture, in addition to acoustic data, hand-based motion patterns of daily activities, the watch was worn on the wrist of the dominant arm.

3.2 Activities Set

To reiterate, our main focus was to recognize complex activities of daily living using inertial and acoustic data collected using a smartwatch. The following criteria were used to select the set of activities:(1) what activities do people do with their hands?, (2) do these activities generate characteristic inertial signals, acoustic signals, or both? (i.e., Are the activities distinct and separable?), and (3) does a commodity smartwatch provide sufficient fidelity to capture such activities? To that end, we selected 23 classes that cover typical activities one performs daily (Figure 1). The activities chosen include daily activities as well as recreational activities to better capture the general population.

To validate our experimental design protocol and instrumentation, we conducted a formative controlled pilot study with two participants. Our laboratory was equipped with the tools and appliances necessary to carry out the activities, such as microwave, sink, and stove. The pilot study helped us ensure the internal validity

Fig. 2. Screenshots from Zoom video call sessions from different participants showing activities performed in their own homes. This demonstrates the natural environment in which the activities were performed.

of our study by addressing issues in our experimental procedure. For instance, we were able to verify that our smartwatch could capture the required sensor data for the duration of the study in a consistent manner as well as improve the user interface of the data collection application. The pilot was used to establish the order of activities participants had to follow for the supervised study, and to evaluate our data annotation and synchronization scheme. Finally, the pilot also provided us with the preliminary data needed to test our hypothesis about the benefit of combining acoustic and inertial modalities for ADL recognition.

3.3 Semi-Naturalistic Data Collection

We ran a field study with participants in their homes to capture realistic data that was used to build and evaluate our recognition models. A diverse group of 15 individuals (9 females and 6 males) with ages varying from 23 to 64 (mean 43.6) and from varying professions and socioeconomic status were recruited through an agency. The studies were conducted remotely via video call due to social distancing regulations. Participants were made sure to have all necessary equipment for the study, mainly tools needed to perform the activities listed in Figure 1. They were only provided with potatoes for the kitchen related activities, such as chopping, grating, and frying. Figure 2 depicts screenshots from the video calls of participants performing activities in their homes.

In the study, participants performed various activities around the house in succession. Two sessions of data collection were conducted, and in each session all 23 activities were performed once. Data collection was continuous from the beginning of a session to the end of a session, capturing all activities and any in-between movements. Once the first session was done, participants were asked to remove the smartwatch and take a 15 minute break before replacing the watch and beginning the second session. This procedure was designed to introduce variability in-between sessions and test wearable placement sensitivity. The watch was always worn on the arm that performed the activities, each of which lasted for a minimum of 30 seconds each. In order to facilitate the annotation process, participants were asked to knock on a surface to indicate the start and end of an activity. The knocking introduced distinct inertial and acoustic markers that helped in segmenting the activities and syncing both sensor data. All the activities were done over video call to monitor and guide participants. Data annotation was manually carried out by the researchers at the end of the study. It was observed that there was a lag in the knocks that participants performed at the beginning and end of activities and the actual starting and

Fig. 3. *In-the-wild* data annotation setup using a mobile phone sitting on a strap on the chest to capture egocentric snapshots while users wore the smartwatch on the wrist.

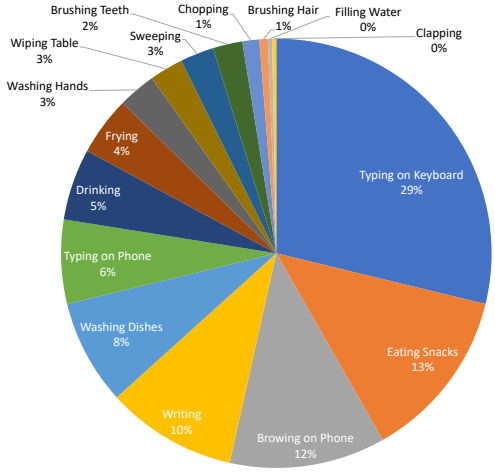


Fig. 4. Activity distribution captured in the *in-the-wild* study.

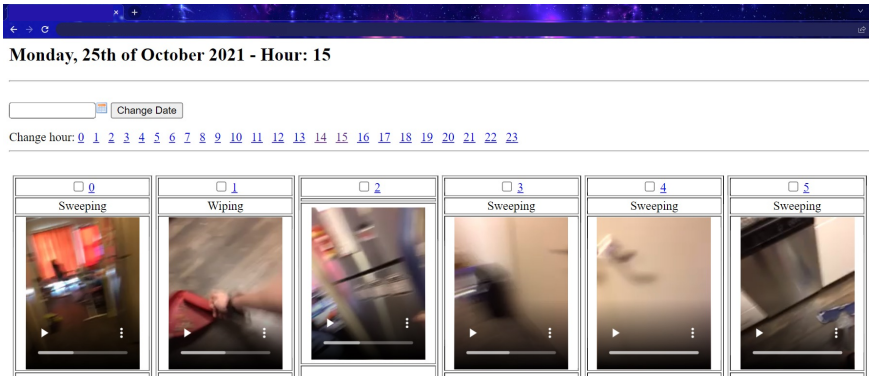


Fig. 5. A screenshot of the web interface used for annotating the video clips collected during the wild study.

ending of the activities. So during the annotation process the video that was captured was also used for precise annotation of activity start and stop events. For activities that required doing repetitive but spaced actions, like drinking or eating snacks, the whole duration of the activity was annotated and individual actions were not singled out.

3.4 In-the-Wild Data Collection

While the *semi-naturalistic* data is valuable for obtaining reliable training data with accurate ground truth labeling, it is still limiting in fairly representing real-world natural behavior. Thus, in order to better evaluate natural behavior, we further conducted an *in-the-wild* study where participants did not follow a script and performed activities in their natural environment and on their own free time.

3.4.1 Data Collection. Equipped with the same smartwatch, we collected data from 5 new participants (4 males and 1 female) with ages varying from 24 to 29 (mean 27) recruited from the university student pool. To ensure natural behavior, the participants did not follow any script but rather were only told to perform their natural daily activities while wearing the smartwatch. Attaining in-the-wild data typically trades off with other aspects of the data collection process, resulting in fewer data samples and/or a smaller range of target activities. More specifically, it is hard to ensure that some target activities are performed and captured during the study especially when the data collection device has a limited battery life. Moreover, acquiring labels of human behavior when unsupervised in their natural environments is challenging. While some rely on self-reporting [55], this method is not always accurate as some participants can fail to report leading to missing labels. Thus, to generate more reliable labels, participants were provided with a smartphone that takes egocentric video clips that were later used to label the activity. The smartphone ran a mobile application that captured a 25-second video every minute and uploaded it to a remote server. The smartphone was worn on the chest with a strap as shown in Figure 3. It is important to note that participants did not follow a specific protocol about which activities to perform, how, or when to perform them. However, to ensure that at least a subset of the target activities are captured, participants were asked to wear the data collection setup when they would be normally performing some of the activities under consideration. Before starting data collection, a video call was scheduled and the participants were instructed on how to start the smartwatch and smartphone application. After verifying that both the applications were running properly and the camera was oriented properly such that the hands were in focus, the participants were asked to keep wearing this setup until the watch battery ran out and perform their daily tasks naturally without any constraints. After the call, no contact was maintained with the participants until the end of the study. This process was conducted twice for every participant on separate days resulting in 2 sessions of data collected per subject for a total of 10 sessions of *in-the-wild* data.

3.4.2 Data Annotation. At the end of the study, data annotation was performed by manually checking the video clips and assigning an appropriate activity label. Synchronizing the timestamp of the data files and the video clips ensured proper segmentation of the inertial-acoustic data. Activiome, the data collection system used for collecting ground truth in-the-wild, is comprised of three components: a mobile phone application, a web application, and a web back-end infrastructure [54]. The mobile application was configured to repeatedly capture 25-second egocentric video clips every 1 minute, which are then uploaded in real-time to the web back-end infrastructure organized around a web server and database. It is important to note that the capture frequency and (video) capture duration affect the battery life and annotation task. Increasing the video capture frequency increases battery consumption, while decreasing the capture duration increases the annotation effort [54]. As such, the application was configured to provide a reasonable compromise between these challenges. The web application was used to review and annotate the video clips at the end of the study. The web interface offers a view of all first-person videos taken on a given day, and provides a tagging interface to annotate the videos with their corresponding activity label (Figure 5). Every video clip was given a label from the list of target activity classes (Figure 1). Some activities don't last the whole 25-second clip. For example, people type on the keyboard for around 10-15 seconds at a stretch. For some other activities, the video either starts or stops in the middle of the activity which results in the early half or later half of the video not capturing the activity. Thus, a video clip was labelled as an activity if it captured that specific activity for the majority of its duration. This same rule was followed when a clip captured two activities. Annotation was conducted by one of the main authors of this work, and the first-person perspective video captures rich contextual detail of the individual's everyday activities eliminating any ambiguity in the labeling process. While the web application interface certainly simplifies the annotation process, it still required researchers to manually check every video clip for reliable labeling.

Since there was no set protocol followed, the set of activities captured at the end of the study as well as the total duration recorded varied significantly across participants. In total, we were able to obtain ~ 33 hours of data.

We analyze the activity distribution in Figure 4. We observe a wide distribution across a total of 16 activities. Since participants were left to perform any activities freely, it can be seen that the dataset is imbalanced with people spending more time typing on keyboard or browsing on phone as well as washing dishes, writing, and eating snacks, which is consistent with naturalistic daily living. Other less frequent shorter activities were also captured, such as drinking, chopping, brushing teeth etc.

4 DATA PROCESSING

Three-axis accelerometer and gyroscope data were sampled at 50Hz, while acoustic data was sampled using the microphone on the smartwatch at a sampling frequency of 22.05KHz. Human speech and other environmental sounds typically contain frequencies up to 8KHz [6]. Thus, this sampling rate permitted frequencies of up to 11KHz to be sampled according to the Shannon-Nyquist sampling theorem [44]. Both audio and motion data are segmented using a frame size of 10 seconds with 50% overlap. The frame size was chosen empirically, and its effect on activity recognition performance on the *semi-naturalistic* dataset is further discussed in Section 7.2. Based on this segmentation and the varying sampling rate of each modality, the input size was 220,500 samples for acoustic data and was 500 samples for inertial data across all six axes.

5 ACTIVITIES OF DAILY LIVING RECOGNITION

Given the *semi-naturalistic* dataset collected from studies described in Section 3, we examine various machine learning methods to build an activity recognition model able to recognize 23 activities of daily living. Our goal is to promote ADL recognition in natural settings with the use of inertial and acoustic sensors from commodity smartwatches, showing the advantage of fusing the two modalities to improve recognition of complex everyday activities. To that end, we explore several classical machine learning models as well as multimodal deep learning frameworks with different sensor fusion methods to establish a set of baselines.

5.1 Classical Machine Learning Models

We employ simple standard machine leaning models built around commonly used measures. After data processing and frame extraction as described in Section 4, we computed statistical inertial features and Mel Frequency Cepstral Coefficients (MFCC) for each frame extracted from inertial and acoustic data respectively. Eight inertial statistical features—mean, median, variance, maximum, minimum, root mean square, skewness, and kurtosis—were extracted from each axis separately for both 3-axis-acceleration and 3-axis-gyroscope data, totalling 48 frame-level features. These features comprise standard commonly used representation for the underlying inertial sensor data. Within the acoustic data, the segmented frames were further broken up into 1-second clips from which 30 MFCCs were extracted and averaged across all 10 clips per frame.

The extracted features serve as input for classification. We explored three machine-learning models and two fusion techniques to establish baseline performance: (1) a Random Forest(RF) classifier with 50 trees, (2) a Naive Bayes classifier(NB), and (3) an AdaBoost classifier that uses a 30-tree Random Forest as the base estimator. For the data fusion, two commonly-used techniques are early-fusion and late-fusion. In the early-fusion technique, features from both modalities were concatenated to form a joint representation before applying a single model that learned the correlation and interactions between the low-level features of each modality. As for the late-fusion technique, models were trained for each of the acoustic and inertial data separately, then the per-modality class probabilities were averaged and used for final prediction.

5.2 Multimodal Deep Learning Framework

While classical models have been shown effective in HAR, those methods typically heavily rely on heuristic handcrafted feature extraction, which is usually limited by human domain knowledge. Deep neural network

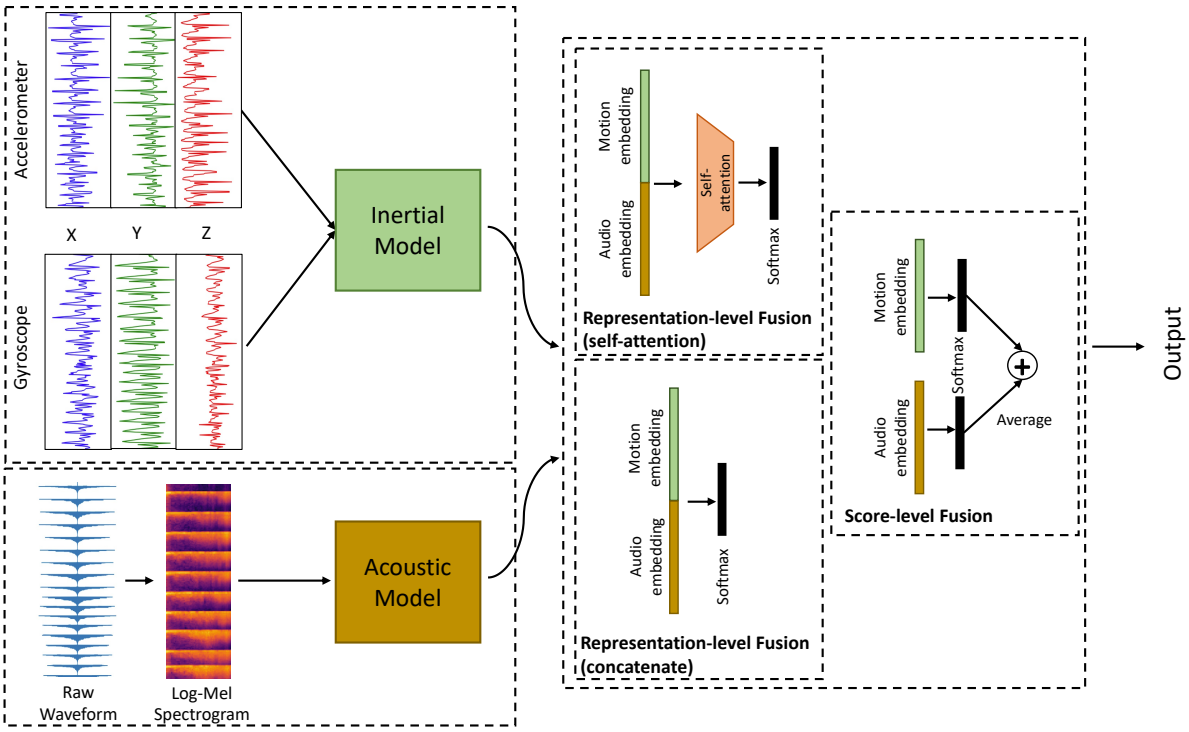
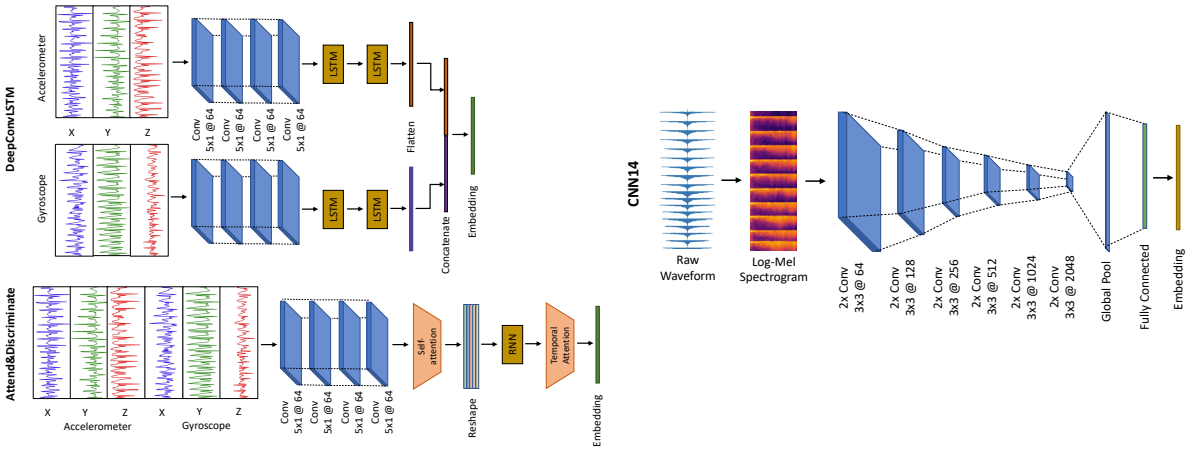


Fig. 6. An overview of the multimodal deep learning activity recognition framework. Our framework includes two single-modality feature extractors, the inertial and acoustic models. Score-level and representation-level late fusion methods were explored for combining audio and inertial information. Representation-level fusion corresponds to either (1) concatenating the inertial and acoustic embeddings or (2) applying a self-attention and then applying joint-training of both models with a single classifier head. Score-level fusion corresponds to training each model with a separate classifier head and averaging the predicted probabilities.

models have been explored and have demonstrated state-of-the-art results for human activity recognition. To that end, we explored several multimodal deep learning (DL) frameworks inspired to push activity recognition performance even further. The goal is to leverage a deep HAR model that directly consumes raw sensory data captured by wearables and outputs precise activity classification decisions. In this section, we describe the state-of-the-art models used for inertial and acoustic activity recognition separately and explore several fusion approaches demonstrating the gain in performance when leveraging both modalities (Figure 6).

5.2.1 Inertial Model. Taking as input the raw inertial sensory data, we explored 2 models to establish a baseline performance on our dataset: (1) *DeepConvLSTM* [41] and (2) *Attend&Discriminate* [1].

DeepConvLSTM: *DeepConvLSTM* is a DNN framework for wearable activity recognition based on convolutional and LSTM recurrent units. After some experimentation, we modified the architecture to extract separate features for accelerometer and gyroscope data by implementing separate *DeepConvLSTM* feature extractors as shown in Figure 7a. This essentially allows the model to better capture intra-modal information, which we observed to improve performance by around 2% compared to the original *DeepConvLSTM*. Thus, every modality input is processed by a convolutional network operating along the temporal dimension followed by recurrent



(a) Inertial Models. (b) Acoustic Models.

Fig. 7. Architecture of Inertial and Acoustic Models.

layers that model the temporal dynamics of the activation of the feature maps. The embeddings of each inertial modality are then concatenated at the last layer before passing through a softmax classifier.

Attend&Discriminate: The second model we investigated is a HAR framework that incorporates cross-channel self-attention and temporal attention [1]. This model, which we will refer to as *Attend&Discriminate*, was shown to achieve state-of-the-art performance on public HAR datasets, outperforming *DeepConvLSTM*. *Attend&Discriminate* applies early fusion by concatenating the sensor channels of the accelerometer and gyroscope data and processes the data through a similar convolutional backbone as *DeepConvLSTM*. Then, a cross-channel self-attention module takes as input the initial convolutional feature-maps at each time-step and learns the interactions between any two sensor channels within the feature-maps. The resulting feature maps are now contextualized with the underlying cross-channel interactions. The feature maps are then passed through a recurrent neural network to model the temporal dynamics. Given that not all time-steps equally contribute to recognizing activities, a temporal attention module is added to learn the relative importance of the time-steps. Figure 7a illustrates the model architecture.

5.2.2 Acoustic Model. For audio classification, we used log-mel spectrograms as input to our deep learning model because they have been successfully employed in prior work [22]. We extracted the spectrograms of our audio clips by computing the short-time Fourier transform (STFT) for each segment, using a Hanning window of 1024 samples and a hop size of 320 samples. The linear spectrogram was then converted into a 64-bin log-scaled Mel spectrogram. Example log mel spectrograms of each activity class are depicted in Figure 8.

Convolutional Neural Networks (CNNs), inspired from VGG-like network, have been proven effective in audio classification when applied to the log-mel spectrograms of the acoustic data [27]. Experimenting with several CNN architectures for recognizing ADLs, we built upon the CNN architecture depicted in Figure 7b, referred to as *CNN14*. The architecture comprises of 6 convolutional blocks, each consisting of 2 convolutional layers (3x3 kernel) and intermediary average pooling layers. Global pooling is applied after the last convolutional layer to summarize the feature maps into vector embeddings which are finally passed through a fully connected layer.

5.2.3 Fusion Methods. Traditional fusion strategies include representation-level fusion [9] and score-level fusion [19]. These strategies can be carried out at different stages, such as early fusion and late fusion.

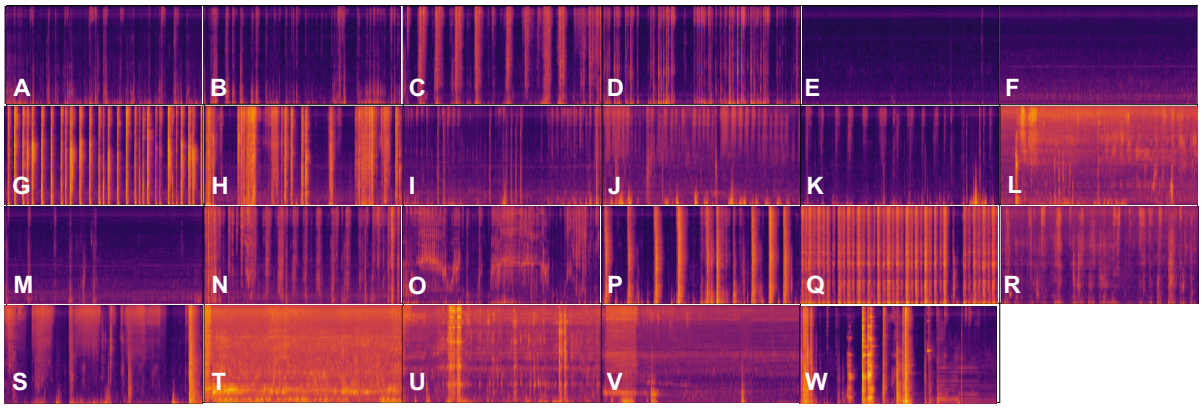


Fig. 8. Example Log Mel spectrograms of 10-second audio clips for 23 activity classes. Photos of these activities and the activity names corresponding to each letter are shown in Figure 1.

Early Fusion: In early fusion, data from different modalities are concatenated (stacked) at the input stage. In our analysis, early fusion of accelerometer and gyroscope data is explored in *Attend&Discriminate* model. Since inertial and acoustic data are processed and sampled differently, concatenation of the two modalities at the input stage is not possible. Moreover, different feature extractors are suitable for each modality. Therefore, late-fusion is instead applied for inertial-acoustic fusion.

Late Fusion: Late fusion processes each modality with a separate network and then combines all their high-level representation via an aggregation operation. As mentioned, the aggregation method can be at the representation-level or the score-level. The representation-level fusion can be (1) a simple concatenation of feature maps or (2) a cross-modality self-attention operation that captures the inter-modality relationship between the acoustic and inertial representations. This is then followed by a single classification head and joint training of both inertial and acoustic networks is applied. The score-level fusion, on the other hand, applies an ensemble method that combines the predictions of each modality-specific network. As such, each network is followed by classification head and is trained separately. The predicted class probabilities are then averaged to obtain a final probability value. The various methods are illustrated in Figure 6.

5.3 Training Implementation

5.3.1 Loss Function. Standard training of deep learning HAR models usually relies on the supervision signal provided by the cross-entropy loss. This directs the model towards yielding inter-class separable activity features. Abedin *et al.* [1] proposed a center-loss criterion that minimizes intra-class variation while maximizing inter-class differences, thus achieving more discriminative feature representations. Given the *semi-naturalistic* aspect of our dataset that introduces more variability across participants, we investigated this additional loss objective on our dataset and its effect on performance. Thus, the final loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{ent} + \gamma \frac{1}{2} \sum_{i=1} \left\| z_i - c_{y_i} \right\|_2^2 \tag{1}$$

where $\mathcal{L}_{ent}$ is the entropy loss, $z_i \in \mathbb{R}^z$ denotes the deep representation for sensory segment x_i, $c_{y_i} \in \mathbb{R}^z$ the y_ith activity class center, and γ the weight coefficient which we set empirically to 0.003. The models were trained for 100 epochs by minimizing the loss function using the Adam optimizer. The learning rate was set to 0.001 and decayed every 10 epochs by a factor of 0.9.

Table 1. Table showing average F1-score for each of motion and audio single-modality classification and each combination of fusion strategies and methods using three different evaluations (LOPO, LOSO, P-LOPO) on the *semi-naturalistic* dataset. LOPO denotes training on all data from other participants and testing on data from the target participant, LOSO denotes training on one session and testing on the other session per participant, and P-LOPO refers to personalized-LOPO with training on all data from other participants and one session from the target participant and then testing on the remaining session from the target participant.

	Model	**Fusion Method**	**LOPO**	**LOSO**	**P-LOPO**
Motion	Random Forest	–	68.4	67.5	74.2
	Naive Bayes	–	60.4	42.3	62.3
	AdaBoost	–	67.8	65.6	73.1
	DeepConvLSTM	–	72.0	44.5	77.1
	Attend&Discriminate	–	**84.0**	**67.9**	**90.8**
Audio	Random Forest	–	41.7	**53.5**	51.5
	Naive Bayes	–	41.4	36.8	42.5
	AdaBoost	–	40.0	49.4	47.7
	CNN14	–	**74.5**	41.2	**82.4**
Early Fusion	Random Forest	Concatenation	77.3	74.8	83.9
	Naive Bayes	Concatenation	72.3	47.8	75.2
	AdaBoost	Concatenation	75.3	74.6	81.9
Late Fusion	Random Forest	Softmax Averaging	64.4	**75.7**	81.2
	Naive Bayes	Softmax Averaging	71.6	45.1	66.2
	AdaBoost	Softmax Averaging	63.8	72.9	80.9
	DeepConvLSTM-CNN14	Softmax Averaging	83.6	55.6	88.5
		Concatenation	78.5	49.2	83.7
		Self-Attention	84.2	68.1	88.7
	Attend&Discriminate-CNN14	Softmax Averaging	88.8	72.6	92.7
		Concatenation	**89.7**	65.9	**94.3**
		Self-Attention	89.1	63.1	94.2

5.3.2 Data Augmentation. A difficulty in human activity recognition with wearable sensors is the acquisition of large amounts of annotated data. This limitation hinders the generalizability and effectiveness of HAR benchmarks. Recently, a data-agnostic augmentation strategy, referred to as *mixup*, was shown effective for time-series HAR data [1]. We incorporate the same on our dataset by implementing the same augmentation strategy as explained in [1]. Essentially, at every batch, *mixup* applies linear interpolation of randomly sampled pairs of data points which are then used for training.

6 EVALUATION AND RESULTS

Our goal is to evaluate and analyze our inertial-acoustic data collected from a smartwatch for the ADL recognition task. Using the data collected, described in Section 3, we aimed to explore the following questions:

- How informative is each modality for ADL recognition in a *semi-naturalistic* setting?
- How beneficial is inertial-acoustic fusion for ADL recognition?
- How does a model pre-trained on *semi-naturalistic* data perform on unseen *in-the-wild* data?

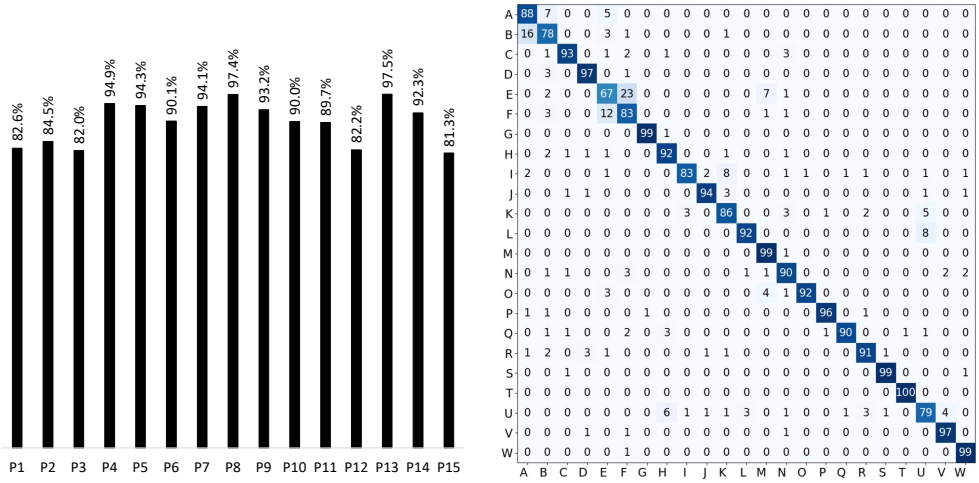


Fig. 9. Results of LOPO evaluation using our multimodal recognition framework: (left) Bar plot of F1-score per user and (right) normalized confusion matrix showing how well each activity is classified. We can see an improvement in some activities (e.g. browsing on mobile phone (F)) when combining acoustic and inertial data compared to single-modality classification (Figure 10).

As previously mentioned , for all our evaluations, a frame size of 10 seconds with 50% overlap was used. We compute the macro F1-score averaged over the classes as the evaluation metric (Equation 2). The macro F1-score is given by:

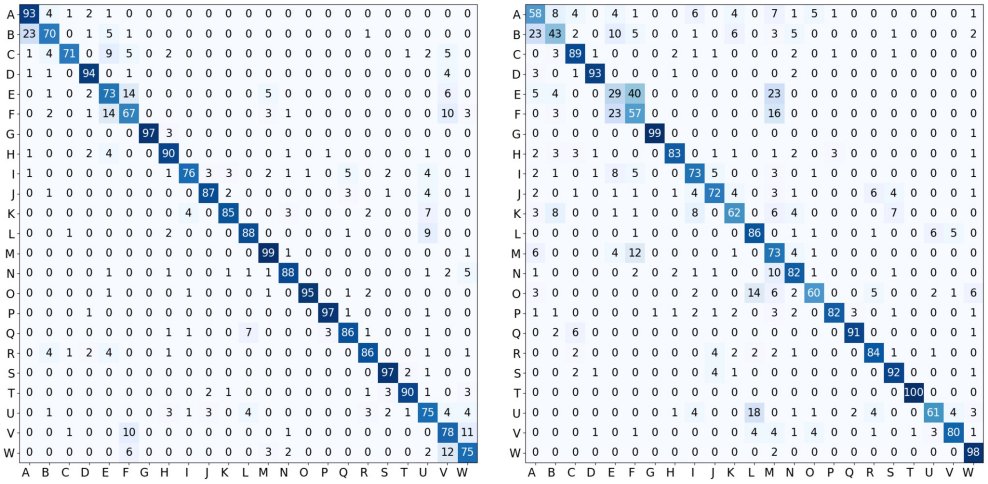
$$F1\text{-}score = \frac{2}{C}\sum_{i=1}^{C}\frac{prec_i \times rec_i}{prec_i + rec_i} \tag{2}$$

where C denotes the number of classes while $prec_i$ and rec_i correspond to the precision and recall for every class i. We compute the metric for the acoustic and inertial model individually and demonstrate the performance improvement with multimodal fusion. Source code of the analysis is made available to the community at https://github.com/Human-Signals-Lab/Sound-and-Wrist-Motion-for-Activities-of-Daily-Living-with-Smartwatches.

6.1 Semi-Naturalistic Evaluation

Using the data collected during the *semi-naturalistic* study, we conducted two evaluations to better analyze our data for ADL recognition: (1) a user-independent evaluation using the leave-one-participant-out (LOPO) strategy and (2) a user-dependent personalized evaluation.

6.1.1 User-Independent Evaluation. Conducting our study in the participants' homes lent our data a significant amount of variability in the living environment and the way activities were performed. To evaluate how well the model generalized across the participants, we employed a Leave-One-Participant-Out (LOPO) evaluation. To quantify the benefits of combining inertial and acoustic data, we examined the acoustic and inertial model performance separately (Table 1). For acoustic-based classification, the deep learning *CNN14* framework achieved a significant improvement in performance compared to the other models (RF, NB, and AdaBoost) reaching an F1-score of 74.5%. As for motion-based classification, *Attend&Discriminate* achieved the highest performance reaching an F1-score of 84% while *DeepConvLSTM* only achieved 72%. We observe that deep learning methods



(a) Motion. (b) Audio.

Fig. 10. Normalized confusion matrices of LOPO evaluation of single-modality classification using acoustic (*CNN14*) and inertial (*Attend&Discriminate*) models. These matrices show how the classification performance of different activities vary between both modalities. For example, vacuuming (T) or using microwave (W) are captured better using acoustic data as opposed to inertial data.

outperform classical models with standard hand-crafted features. While further investigation can potentially further improve performance, in this work our aim is to provide a baseline characterization of our dataset for the task at hand.

Combining acoustic and inertial sensing, we observed an overall increase in performance across all models for both early and late fusion (Table 1). Applying concatenation-based late fusion using the *Attend&Discriminate* and *CNN14* architectures for each of the inertial and acoustic modalities respectively achieved the highest performance of 89.7% F1-score, an improvement of 5.7% over the top single-modality classifier. The other fusion methods (softmax averaging and self-attention) using the same architectures achieved comparable performance. When using the *DeepConvLSTM* architecture instead for the inertial stream, there was still an improvement when compared to the single-modal classifier with self-attention fusion method achieving 84.2%. The confusion matrix and the per-participant performance for the best performing model (*Attend&Discriminate-CNN14 with Concatenation*) is shown in Figure 9. The confusion matrices for the single-modal models, *Attend&Discriminate* for motion and *CNN14* for audio, are also shown in Figure 10.

6.1.2 User-Dependent Evaluation. Every user's environment varies significantly, whether from background noise provided by the ambient hum of HVAC or bustling traffic, or in terms of appliance sounds such as microwaves, blenders, and device placement. Furthermore, it was observed when visualizing the data that the same activity across different participants exhibited different patterns, especially when visualizing user-specific inertial data. Thus, developing personalized models has the potential to further improve performance. To explore this potential, we tested our model using an individualized Leave-One-Session-Out (LOSO) evaluation. The data associated with each participant included 2 sessions of 23 activities and for every participant, we trained on one session and tested on the other (Table 1). While Random Forest and Adaboost showed an increase in performance compared to LOPO when using audio, motion did not exhibit a significant change. Combining both modalities using the

Table 2. Table showing average macro (f_m) and weighted (f_w) F1-score for each of motion and audio single-modality classification and each combination of fusion strategies on the *in-the-wild* dataset.

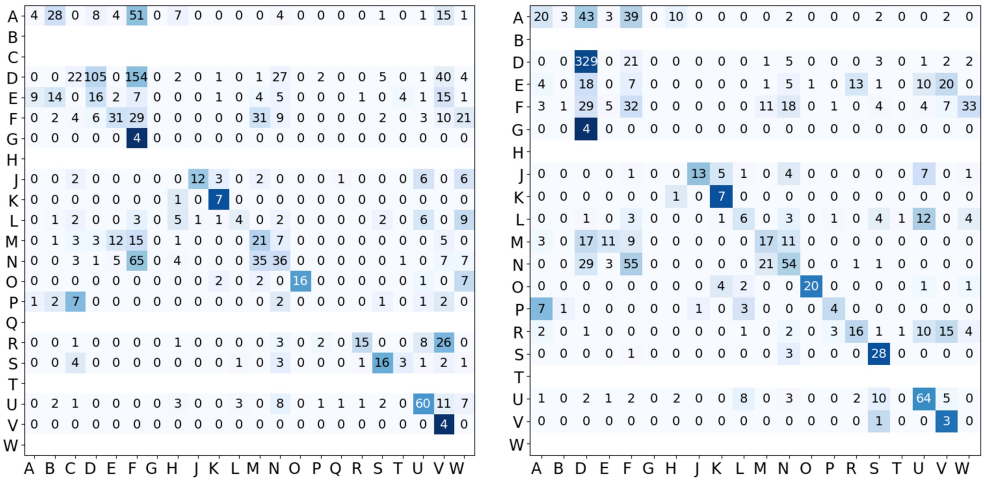
	Model	Fusion Method	Inference (f_m/f_w)	Fine-tuning (f_m/f_w)
Motion	DeepConvLSTM	–	14.7 / 21.3	20.8 / 51.2
	AttendDiscriminate	–	17.7 / 24.4	23.1 / 51.8
Audio	CNN14	–	10.9 / 21.0	20.0 / 45.8
Late Fusion	DeepConvLSTM-CNN14	Softmax Averaging	14.6 / 23.9	23.5 / 54.9
		Concatenation	16.5 / 23.7	22.5 / 50.3
		Self-Attention	16.8 / 23.8	23.8 / 50.5
	Attend&Discriminate-CNN14	Softmax Averaging	16.1 / 25.5	24.2 / 54.2
		Concatenation	**20.1 / 26.8**	**30.0 / 55.8**
		Self-Attention	19.2 / 27.4	27.6 / 55.4

aforementioned classifiers resulted in comparable performance. Moving towards deep learning frameworks, performance for the acoustic *CNN14* model as well as both *DeepConvLSTM* and *Attend&Discriminate* inertial models exhibited a significant drop. Applying late fusion in both *DeepConvLSTM-CNN14* and *Attend&Discriminate-CNN14*, we observed a $\sim 20-30\%$ drop in performance for concatenation-based and self-attention based fusion methods, while a $\sim 16\%$ drop in performance with *Attend&Discriminate-CNN14* with softmax averaging.

An important shortcoming of LOSO evaluation is the limited amount of data available for model training which is especially challenging for deep learning frameworks, which can lead to overfitting and in turn cause a drop in performance as observed. Therefore, in another evaluation strategy, we trained a model for a target participant with data from all other users, similar to LOPO, but with an additional session from the target participant, effectively allowing us to increase our training data while still personalizing our model. We refer to this evaluation as personalized-LOPO (P-LOPO). As expected, this yielded better results for both single-modal and multimodal deep learning frameworks (Table 1). The acoustic model achieved an F1-score of 82.4%, an increase of 7.9% compared to LOPO. The inertial *Attend&Discriminate* model achieved an F1-score of 90.8%, an increase of 6.8% compared to LOPO. Fusion also resulted in a further increase in performance reaching 94.3% with concatenation-based *Attend&Discriminate-CNN14*, an increase of 4.6% over LOPO evaluation. Similarly, Random Forest and AdaBoost resulted in a $\sim 6\%$ and a $\sim 16\%$ increase in performance with early and late fusion respectively compared to LOPO. It is evident from all three evaluation strategies that inertial-acoustic fusion is beneficial for improving ADL recognition compared to single-modal classification. To motivate this evaluation strategy in an uncontrolled environment, the smartphone user would have to annotate some of their own data or do a specific activity, by request of the system, to make the detector more robust to personal variations. In similar settings, voice assistants request repeating the trigger phrase a few times during setup to personalise the voice recognition model to a specific user.

6.2 In-the-Wild Evaluation

Another critical topic we explore in this work is the challenge of dealing with *unseen in-the-wild* data collected in a completely free-living setting. It is important to note, participants from both studies are mutually exclusive. Given the highly imbalanced nature of our *in-the-wild* dataset as observed in Figure 4, we report both macro F1-score (f_m) and weighted F1-score (f_w). Macro F1-score is highly sensitive to rare labels, which can unfairly dominate the evaluation metric. On the other hand, weighted F1-score deals with this imbalance by considering the number of samples per class in the data.



(a) Inference. (b) Fine-tune.

Fig. 11. Confusion matrices of *in-the-wild* (a) inference and (b) fine-tune evaluations using inertial-acoustic *Attend&Discriminate-CNN14* model with concatenation-based fusion. Note that due to the imbalance nature of the dataset, the values reported correspond to the number of samples (i.e. un-normalized matrix) to better observe the majority and rare classes, while the color mapping corresponds to the normalized matrix to better visualize how accurately the model predicts a specific class. Missing classes in the dataset are left blank or eliminated to maintain a square matrix.

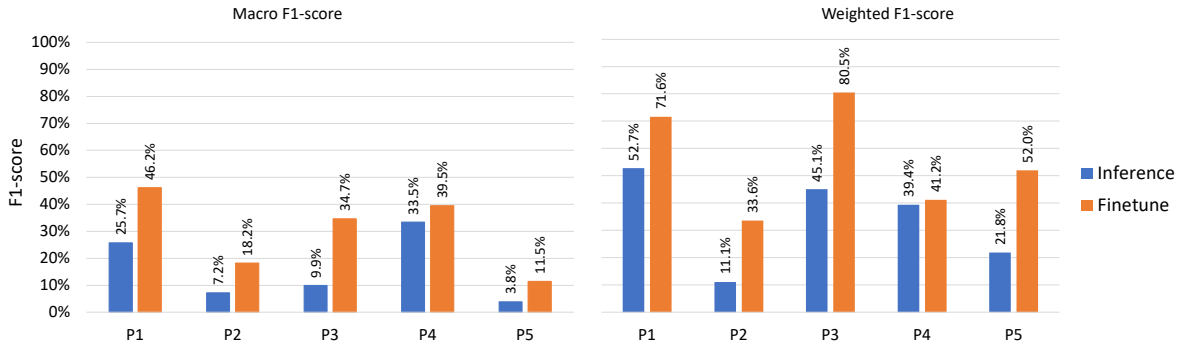


Fig. 12. Bar plots of macro (left) and weighted (right) F1-scores per user in the *in-the-wild* study for both Inference and Fine-tune evaluations.

6.2.1 Inference. To evaluate the ecological validity of our acoustic-inertial ADL recognition frameworks, we conducted inference on the *in-the-wild* data using the models trained on the *semi-naturalistic* data. This essentially enabled us to better understand the applicability of such models when deployed and tested in real-world settings. As expected, we observed that ADL recognition on *unseen in-the-wild* data is indeed a challenging problem, resulting in a drastic drop in performance compared to *semi-naturalistic* settings (Table 2). Examining the single-modality performances, we observed the *Attend&Discriminate* model achieving the highest performance when using only the motion data, an average of 17.7% f_m and 24.4% f_w across all participants. Using only audio data,

CNN14 achieved an f_m of 10.9% and an f_w of 21.0%. Applying late fusion for inertial-acoustic modeling, the *Attend&Discriminate-CNN14* with concatenation achieved the highest performance of 20.1% f_m and 26.8% f_w, representing a ~ 2% gain compared to single-modal inertial classification. Looking at the confusion matrix (Figure 11a), we can observe how well the model can accurately predict a majority class as opposed to a rare class and where the confusion occurs. For example, some rare classes such as *filling water* (V) and *brushing hair* (K) were accurately predicted while others such as *clapping* (G) were mis-predicted. The majority class, *typing on keyboard* (D), was predicted correctly around 30% of the time. Further examining the variability in performance across the different participants, we plot the macro and weighted F1-scores per participant (Figure 12).

6.2.2 Fine-tuning. The inference results point to the fact that *in-the-wild* data, being completely naturalistic, exhibits distribution shifts, wherein the training distribution differs from the test distribution [26]. Motivated by this observation, we performed a fine-tuning analysis, wherein a model pre-trained on the *semi-naturalistic* data is fine-tuned on the *in-the-wild* data. Due to the limited amount of data collected as well as the imbalanced and incomplete nature of the classes included, we performed a leave-one-session-out evaluation with sampling and augmentation from the *semi-naturalistic* data. At every iteration, we fine-tune the pre-trained model with data from the *in-the-wild* dataset. Specifically, we train on 9 sessions of the *in-the-wild* dataset, test on a hold-out session and then iterate for all sessions, averaging the result. To help reduce the class imbalance observed in the *in-the-wild* dataset, and resulting bias towards certain classes, we also added randomly sampled data from the *semi-naturalistic* dataset, focusing on missing and rare classes.

With such evaluation, we observed an improvement in performance across all models and fusion methods (Table 2), with the *Attend&Discriminate-CNN14* model with concatenation reaching 30.0% f_m and 55.8% f_w. For single-modality classification, *CNN14* achieved a performance of 20.0% f_m and 51.2% f_w on the audio data, and *Attend&Discriminate* reaching 23.1% f_m and 51.8% f_w. Further analyzing the per-participant performance, we observed significant variability as shown in Figure 12. While fine-tuning improved results, P2 and P5 still exhibited low macro F1-scores. This variation can be attributed largely to the different context and environments of the participants in their natural settings, e.g., some had more background audio in the form of news, music and conversation than others, as well as differences in the activities captured. Weighted F1-scores improved drastically for some participants, reaching 80.5% for P3 and 71.6% for P1, a gain of ~ 25-45%. To further understand the change in the model's predictive capabilities per-class after fine-tuning, we plot the confusion matrix (Figure 11b) which shows a definite improvement in predicting the majority class, typing on keyboard (D), compared to simple inference. This explains the large improvement in the weighted F1-score. The overall performance improvement compared to inference further validates the distribution difference between real *in-the-wild* and structured data.

7 DISCUSSION

7.1 NULL Class and False Positives

The evaluation procedures presented in the paper so far were performed with a closed set of activities. Examples of these activities were captured in the human-subjects study and used to train the classifier. However, real-world deployments are subjected to unknown or out-of-scope activities, never before encountered by the classifier. These types of activities are often referred to as members of a NULL class. Evaluating models on out-of-scope activities is useful because it can help assess their performance with respect to false positives. To compose this NULL class, we leveraged sensor data from the *semi-naturalistic* study collected from participants as they transitioned from one activity to another during the study, which included varying motion patterns, as participants grabbed and moved objects while chatting with the experimenter. We evaluated our LOPO models on this NULL class by applying the *Attend&Discriminate-CNN14* model with concatenation-based fusion to the out-of-scope data, the average prediction confidence for all instances was 45.4% (± 19.8%). When using only audio data, the average confidence prediction was 34.0% (±16.2%). When using only motion data, a higher average prediction confidence

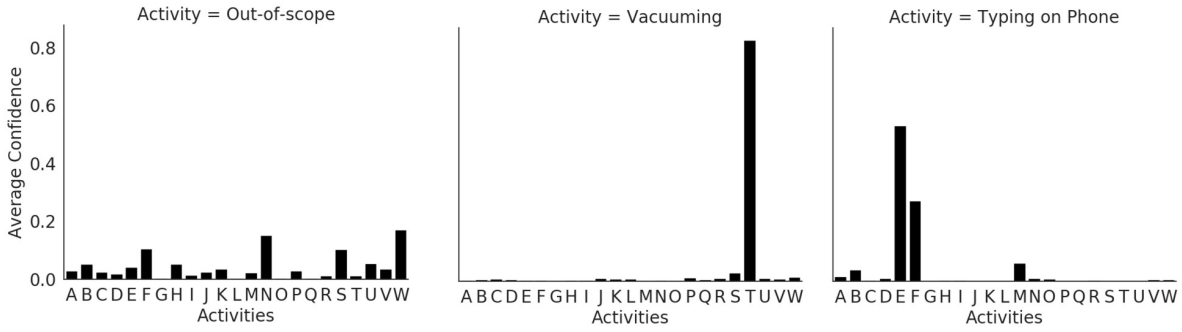


Fig. 13. Bar plots showing the average probability confidence (softmax output) distributed across the 23 activities from data corresponding to three activities. Every bar plot corresponds to data from a specific activity. The out-of-scope bar plot shows the average confidence is low and distributed across all activities. On the other hand, vacuuming (T), which the model had 100% F1-score in (see Figure 9) depicts a peak average confidence > 0.8 on the target activity. Looking at an activity where the model was less confident (e.g. typing on phone (E) was confused with browsing on phone (F))

, the average confidence dropped to ∼ 0.5 but was still more skewed compared to out-of-scope data.

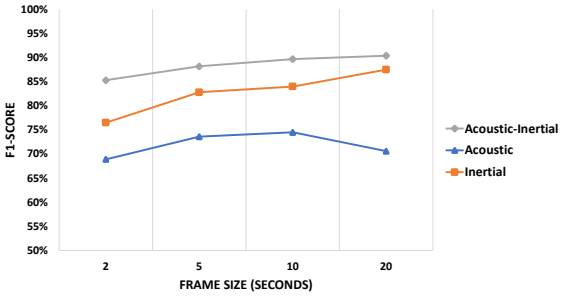


Fig. 14. F1-score with variable frame size. Performance increased with increasing frame size and no significant increase occurred beyond a 10-second frame size.

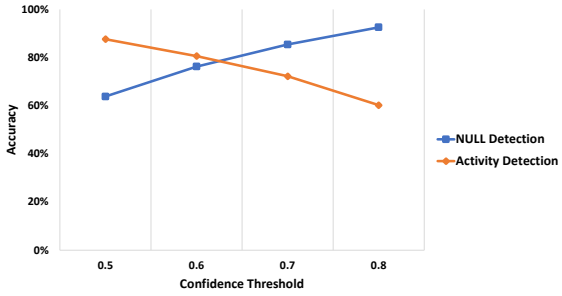
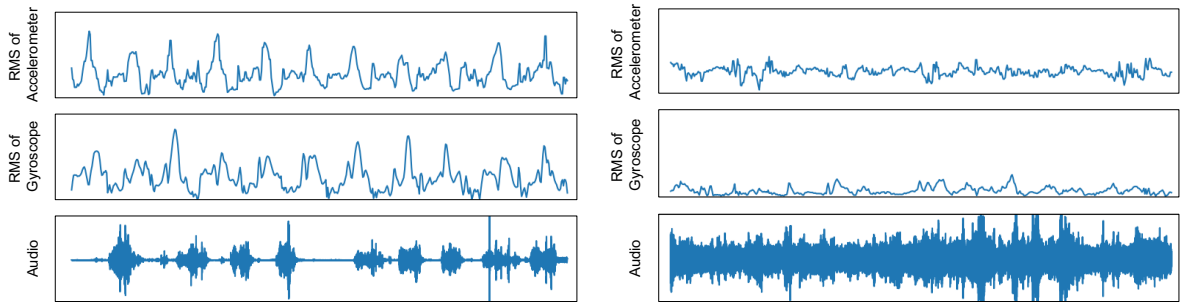


Fig. 15. Accuracy performance with variable confidence threshold. Trade-off between NULL detection and Activity detection is observed with changing threshold.

on out-of-scope activities was observed with the *Attend&Discriminate* model reaching 48.2% (±21.2%). From our previous LOPO evaluation on our target activities set, the prediction confidence value averaged around 78.2% (± 19.5%), when fusing both modalities, which exceeds the average prediction confidence for the "unknown" instances. Thus, this value can be used to define a confidence threshold for classifying and ignoring "unknown" instances, i.e. an instance is classified as "unknown" if the top predicted class does not exceed a certain confidence threshold. When using the single-modal models, modeling only audio data with *CNN14* resulted in an average prediction confidence of 57.5% (±23.2%) while motion resulted in 74.1% (±21.2%) average prediction confidence. Looking at the predicted probability distribution across the 23 activities that the softmax outputs and averaging across all data instances belonging out-of-scope data, we can see that the model's average confidence in each activity is low and spread out across all activities, which indicates the model's high uncertainty in its prediction (Figure 13a). On the other hand, visualizing the same for a target activity that the model is highly confident in (e.g. vacuuming), the model's average peak confidence is > 0.8 and is centered at the corresponding target activity



(a) Brushing hair (b) Vacuuming

Fig. 16. Example raw acoustic and inertial data from two different activities: (a) brushing hair and (b) vacuuming. Inertial data is more relevant for brushing hair as it exhibits a unique motion pattern and contains little acoustic data. Conversely, vacuuming contains more acoustic than inertial information.

(Figure 13b). When dealing with a target activity that the model is less confident in (e.g. typing on phone was confused with browsing on phone), the average peak confidence drops to ~ 0.5 which still shows a lower model uncertainty compared to out-of-scope data (Figure 13c).

To identify a reasonable confidence threshold, we perform NULL and Activity detection in a LOPO evaluation while varying the confidence threshold. NULL detection depicts the model's ability to reject "unknown" frames from the out-of-scope data while activity detection shows its ability to correctly detect activity frames. Varying the confidence threshold, we observe a trade-off with NULL detection accuracy increasing as the threshold increases while that of activity detection decreases (Figure 15). Setting a confidence threshold of $\sim$ 60%, the best trade-off was achieved with NULL detection reaching 76.2% and activity detection 80.6%. Using a model trained on all participant's activity data in the *semi-naturalistic* dataset, we again evaluated the NULL detection performance, with 60% confidence threshold, and observed an improvement reaching 98.2% accuracy.

7.2 Frame Size Sensitivity

In human activity recognition systems, frame size plays an important role in feature extraction and classification, ultimately affecting overall performance. To gauge how sensitive ADL recognition is to this parameter, we evaluated the best performing frameworks for single-modal (*Attend&Discriminate* and *CNN14*) and multimodal (*Attend&Discriminate-CNN14* with Concatenation) activity recognition while extending the frame size from 2 seconds to 20 seconds in non-uniform jumps in a LOPO evaluation. Figure 14 illustrates a gradual increase in performance with increasing frame size for both inertial data and acoustic data. Increasing the frame size to 20 seconds, we observed a drop in performance for acoustic data, while an improvement of $\sim$ 3% was observed for inertial data. When applying the multimodal analysis, increasing the frame size did not yield significant improvement. Additionally, further increasing the frame size increases the detection latency, which is critical when deploying this system in real-time. Finally, some activities might also be short, and so having a large frame size may hinder their detection. Therefore, we found a 10-second frame size to be optimal for our application.

7.3 Multimodal Sensor Fusion

Fusion of the acoustic and inertial data does increase the overall performance of the system. However, some activities do not exhibit a drastic improvement which include activities lacking in either acoustic or inertial data. For example, hair brushing has a distinct hand motion associated with it while containing negligible sound.

Conversely, vacuuming has insubstantial motion involved with it and it has a distinct loud sound while in use. Figure 16 shows the significant difference in information captured between modalities for these two activities. Moreover, investigating where misclassification occurs in the confusion matrix helps to understand the predictive ability of the system. Figure 10 supports this observation as the acoustic and inertial models are more accurate in detecting respective activities.

For fusing inertial and acoustic data, we explored several late fusion methods applied to deep learning frameworks: (1) score-level fusion via softmax averaging and (2) representation-level fusion via concatenation or self-attention. The representation-level fusion method performs joint training of a multi-branch architecture via a single classification head and a single loss. This couples the inertial and acoustic branches together, sharing information from both modalities to learn a suitable data representation. On the other hand, score-level fusion via softmax averaging essentially trains an ensemble of classifiers separately before aggregating their predictions. As such, each classifier only learns from their corresponding modality, thus capturing the diversity across modalities. In our evaluations, concatenation with the *Attend&Discriminate-CNN14* model outperformed all methods in LOPO and P-LOPO evaluations. However, the performance difference between concatenation, self-attention, and softmax averaging is insignificant, suggesting that not one method is more effective than the other. In LOSO evaluation on the other hand, representation-level fusion, whether via concatenation or self-attention, suffered a drop in performance, with softmax averaging achieving better performance. This observation can be attributed to the effect of limited training data when joint training a complex multi-branch model architecture with a large number of training parameters. Thus, separately training classifiers, in this case, resulted in better performance, with even low-complexity classical models, such as Random Forest, achieving the best performance in LOSO.

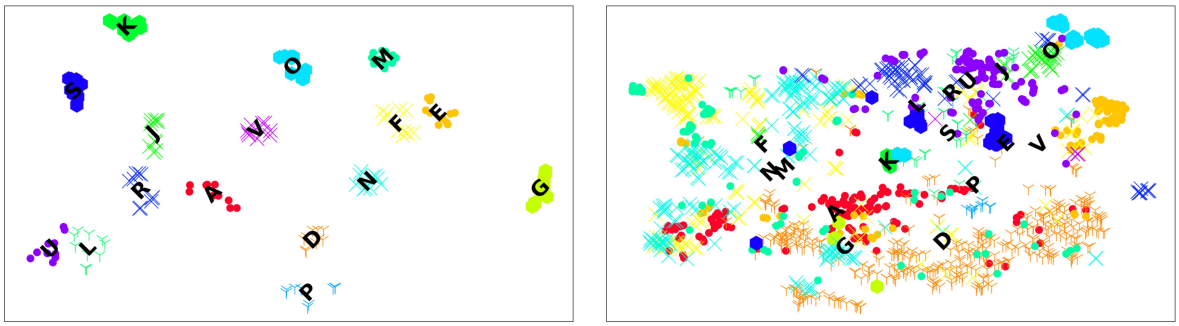
With regards to *in-the-wild* evaluations, inertial-acoustic fusion improves performance compared to single-modal classifiers, with *Attend&Discriminate-CNN14* with concatenation-based fusion outperforming other methods. Interestingly, a larger performance difference between score-level fusion via softmax averaging and representation-level fusion was observed. This suggests that representation-level fusion with joint training of a multi-modal architecture can potentially provide a more robust model that is more effective on real-world unstructured data. A more extensive exploration and analysis of this topic is needed to better characterize its effectiveness.

7.4 Smartwatch Power Consumption

In terms of data collection, one area of concern when using a smartwatch for continuous data capture is power consumption. For our current implementation and the purposes of our study, data processing was not performed on the watch. But, sampling both the microphone and the IMU continuously affects battery life during data collection. During the in-the-wild study, participants wore a fully charged smartwatch until the battery was completely exhausted, which lasted on average for 3.5 hours when sampling both audio and motion data. When sampling only motion data, a fully charged smartwatch lasted for 8 hours, whereas sampling only audio data resulted in a battery life of 4.5 hours. As expected, sampling audio has a larger effect on power consumption since sampling occurs at a much higher frequency. Moreover, compression and encoding of the data is needed for storage. It is important to note that our smartwatch data collection application was not optimized for power consumption. We earmark power consumption optimization as well as on-device data processing and model prediction as an important area of research we plan to explore in future work.

7.5 In-The-Wild Challenges

In this section, we discuss and characterize the specific challenges that we encountered when processing and analyzing the *in-the-wild* real-world dataset.



(a) *Semi-naturalistic* data. (b) *In-the-wild* data.

Fig. 17. t-SNE visualization of data embeddings showing distribution shift between (a) *semi-naturalistic* and (b) *in-the-wild* test data. Activity classes are clearly clustered in *semi-naturalistic* data while not clearly separable in the *in-the-wild* data.

7.5.1 Noisy Data: Recognizing daily activities in-the-wild is more challenging, compared to controlled or semi-controlled settings, largely due to the variability in real-life. Variability in the way people perform different activities as well as the overlap in multiple activities being performed simultaneously was visible during annotation. For example, in several cases, participants had TV or music on while cooking or washing dishes which the acoustic model appears to not generalize well to, causing performance degradation. Noise from activities being performed by other members of a household were also captured. In terms of inertial data, some activities were not performed using the dominant hand which would not be captured and thus can be misleading to the motion model. Finally, some participants also performed multiple activities at the same time like browsing on the phone while cooking/eating, drinking while eating etc. In such cases, the label of the activity being performed by the dominant hand on which the watch was worn was used. Thus, while inertial sensing was able to capture the target activity, acoustic data was less accurate since it was exposed to other background noise. In light of these challenges, an evident future direction is to investigate better modelling and fusion approaches to improve ADL recognition in the wild.

7.5.2 Data Annotation: Another difficulty in-the-wild was the resolution of annotation. The video clips we obtained for establishing ground truth were 25 seconds long. Continuous capture of video was not feasible due to smartphone battery life considerations as well as the data collection software functionality. In our ground truth labelling setup, it was not possible to label only a fraction of the video clips, and the whole clip had to be labeled as a specific activity. To clarify, this is a limitation of the labelling software which did not have the provision to accurately mark the start or end of an activity. During annotation, it was noted that some activities don't last the whole 25 seconds. For example, people type on the keyboard for around 10-15 seconds at a stretch. For some other activities, the video either starts or stops in the middle of the activity which results in the early half or later half of the video not capturing the activity. A video clip was labelled as an activity if it captured that specific activity for the majority of its duration. As a consequence, when the frames are created by segmenting the data following the video clips, the corresponding ground truth assigned might not reflect the activities captured in a small number of frames. In the future, we hope to improve the validity of the labels by enabling further segmentation of the video clips, and supporting low-burden forms of annotation [2], including methods that rely on active learning [3].

7.5.3 Distribution Shifts: As observed from the evaluation in Section 6.2, *in-the-wild* data exhibits critical domain and distribution shifts compared to the *semi-naturalistic* data. Distribution shifts, due to variability among participants, device types being used, environments, context, and activity patterns, are ubiquitous in

real-world settings, resulting in a need for domain generalization and adaptation. While several prior work have acknowledged this challenge and attempted to tackle this problem [12, 21], analysis has been limited to either open datasets collected in controlled settings or retrofitting such datasets with cleanly characterized distribution shifts and augmentations that are not always likely to arise in real-world deployments. In this work, we study this problem by collecting real-world data in completely naturalistic uncontrolled settings and observe the distribution shift numerically through performance drops as well as visually by applying t-SNE visualization [57]. In Figure 17, using the *Attend&Discriminate-CNN14* model with concatenation fusion trained on P1-14 participants from the *semi-naturalistic* data, we visualize the data embeddings of test data from P15 and *in-the-wild* data. We clearly observe that classes are well clustered in data collected in the *semi-naturalistic* setting, similar to the training data, despite it being collected from a participant in their own home whose data was excluded from the training data. On the other hand, data collected *in-the-wild* are less separable demonstrating the domain shift problem that arises when moving from controlled or semi-controlled settings to real-world environments.

7.5.4 Imbalanced Data and Missing Classes: While collecting data in uncontrolled naturalistic environments enables capturing real-world data and evaluating the robustness and effectiveness of the recognition frameworks, it trades off with other aspects of the data collection process, resulting in limited imbalanced data with missing activity classes (Figure 4). As a result, assessing the model's predictive abilities across different activities is a challenge, especially for rare classes. Moreover, fine-tuning the models on this data, to mitigate the distribution shift discussed in Section 7.5.3, causes the model to become biased towards the majority class and leads to forgetting of missing classes. In an effort to mitigate this issue, we sampled remaining classes from the *semi-naturalistic* study, ultimately showing some performance improvement. We believe this analysis is the first step towards better understanding the gap between structured and real-world data and the challenges faced when deploying standard recognition models in real-life. Due to the challenge of collecting and annotating real-world data, research towards developing methods for real-world fine-tuning and domain adaptation [24], especially with imbalanced data or missing target classes, is needed to be able to achieve reliable deployment of DL models in the wild.

8 APPLICATION SCENARIOS

Similarly to how coarse-grained activity recognition opened up new possibilities in health tracking, we believe practical ADL recognition will enable a wide range of new applications in a variety of domains. We describe some of these applications below.

- **Personal Diary**. ADL recognition and tracking allows maintenance of more detailed personal logs of a user's daily activities. Instead of only monitoring coarse mobility states such as running, driving, walking or sitting, a user could capture more granular information such as time spent cooking, time spent in recreational activities, etc., providing a more holistic view of their lifestyle. By being presented with such information, users would also be able to reflect on their day-to-day activities and better manage their time.
- **Smart Environments**. Knowing the context of the user is important in developing truly smart environments that can seamlessly integrate with their lifestyle. For example, sitting at a desk to write and sitting at a desk to work on a computer may require different lighting conditions—writing might require more focused lighting, whereas working on a computer might require a brightened room to prevent eye-strain. Automatically detecting and adjusting to these scenarios can improve user satisfaction. While context and activity awareness are useful for the enhancement of smart homes, we believe these systems are even more powerful when driving accessibility applications, facilitating the life of individuals with disabilities.
- **Health and Hygiene**. Apart from health metrics that are provided by the available fitness trackers, our system can provide monitoring of complex daily activities for a more comprehensive approach to health tracking. For instance, drinking detection can track user hydration, while monitoring eating can help with

controlling diabetes and obesity [11]. Tracking activities like teeth-brushing and hand washing can increase user attention towards personal hygiene. Recognition of other activities like washing dishes, sweeping, and vacuuming can nudge users to keep their environments clean thus promoting a healthier lifestyle.

- **Aging, Skill Degradation, and Rehabilitation**. Longitudinal activity tracking in the home can provide insights into skill degradation, which is generally a sign of mobility impairment, and potentially detect symptoms that signal the onset of neuro-muscular and neural diseases such as Parkinson's and myasthenia gravis [8, 23]. Moreover, people undergoing physical therapy after an accident often have scheduled clinical visits for tracking their rehabilitation progress. Having a system that can unobtrusively and continuously monitor activities can help physiotherapists better assess their patient's progress and provide interventions when needed.

8.1 Privacy Considerations

Despite the myriad of ways in which applications can benefit from activity and context recognition, deployment in real-world settings demand special consideration when it comes to human-centered issues such as privacy. While audio has been extensively used as a sensing modality in activity recognition [28], the privacy risks it poses cannot be underestimated. Although mitigating privacy risk was not a focus of this work, there are many steps we would take to render our approach viable. First of all, we see our system being expanded to not only capture sensor data with the smartwatch but also process audio data on the device. As such, pre-processing, data segmentation, and inference would run locally on the smartwatch, eliminating the need for any captured audio to be saved, thus extending its risk surface. Incorporating speech-filtering or voice-masking algorithms into the framework would be another way to strengthen the privacy protection of our approach [60]. Finally, with a small trade-off between privacy mitigation and recognition performance, frame-level audio degradation can also be used to reduce audio intelligibility while maintaining the necessary information for audio classification [31].

9 LIMITATIONS

While our work demonstrated the capabilities and limitations of smartwatches for ADL recognition in naturalistic settings, it is important to highlight the shortcomings of our study and approach. Firstly, our participants, while diverse in age and gender, were all right-handed. This characteristic of our population may have impacted our results. Many individuals choose to wear a watch on the passive arm, which makes it challenging to capture relevant inertial data for many activities, especially for tasks like writing, brushing teeth or hair, chopping, and wiping. Leveraging audio data in our system can compensate for the first limitation, but it also leads to the second limitation of our approach: Our model leverages acoustic data for enhancing activity recognition performance, and one of the major disadvantages of using sound as an input is its high susceptibility to external noise. As such, for our *semi-naturalistic* dataset, we attempted to minimize the noise in our collected data. However, the data being collected in peoples' own homes was still subject to some common household background noise, such as HVAC systems, pets, babies crying, and other people moving around the house. For the *in-the-wild* dataset, background audio interference resulted in performance degradation and misclassifications.

Regarding our framework, a limitation was its inability to recognize simultaneous activities. Real world settings are often chaotic as people multitask, with multiple activities occurring at the same time, as was observed during the in-the-wild study. Our *semi-naturalistic* dataset focused on one activity at a time, and thus, did not include non-overlapping sounds and motion patterns originating from multiple activities. Our in-the-wild dataset included overlapping activities that caused confusion, especially when inertial data captured one activity while audio data captured another. For example, one of our study participants browsed the web on the phone while cooking.

Lastly, we recognize that additional participants would enhance the external validity of our results. Having said this, our experimental approach was very rigorous. For example, we chose to establish ground truth with video evidence from a wearable camera instead of relying on self-reported surveys. We feel that we gained important and useful insights from data collected in both *semi-naturalistic* and fully in-the-wild settings.

10 CONCLUSION

In conclusion, this paper explores inertial-acoustic sensing from off-the-shelf commodity smartwatches for recognizing activities of daily living. We demonstrated that an off-the-shelf commodity smartwatch can be used without any modifications, hardware or software, to collect synchronous acoustic and inertial data. We validated this claim by collecting two datasets: (1) a *semi-naturalistic* dataset with 15 participants naturally performing 23 activities supervised in their homes and (2) an *in-the-wild* datasets with 5 participants. Through a comprehensive set of evaluations, we showed the benefit of leveraging multi-sensor data for ADL recognition as well the challenges that arise in real-world settings. This work represents a step forward in building high-performing activity recognition systems that leverage both acoustic and inertial data captured using practical off-the-shelf, wrist-worn devices.

REFERENCES

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 1 (March 2021), 22 pages. https://doi.org/10.1145/3448083

[2] Rebecca Adaimi, Ka Tai Ho, and Edison Thomaz. 2020. Usability of a Hands-Free Voice Input Interface for Ecological Momentary Assessment. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 1–5.

[3] Rebecca Adaimi and Edison Thomaz. 2019. Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–23.

[4] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–24.

[5] S. Arshad, C. Feng, Y. Liu, Y. Hu, R. Yu, S. Zhou, and H. Li. 2017. Wi-chase: A WiFi based human activity recognition system for sensorless environments. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. 1–6.

[6] Rohima Badri, Jon Siegel, and Beverly Wright. 2011. Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms. *The Journal of the Acoustical Society of America* 129 (02 2011), 852–63. https://doi.org/10.1121/1.3523476

[7] Tanvi Banerjee, James M. Keller, Mihail Popescu, and Marjorie Skubic. 2015. Recognizing complex instrumental activities of daily living using scene information and fuzzy logic. *Computer Vision and Image Understanding* 140 (2015), 68 – 82. https://doi.org/10.1016/j.cviu.2015.04.005

[8] Carolina Barnett, Vera Bril, Moira Kapral, Abhaya Kulkarni, and Aileen M. Davis. 2014. A Conceptual Framework for Evaluating Impairments in Myasthenia Gravis. *PLOS ONE* 9, 5 (05 2014), 1–9. https://doi.org/10.1371/journal.pone.0098089

[9] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: Combining Audio and Motion Sensing for Gesture Recognition on Smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19)*. ACM, New York, NY, USA, 10–19. https://doi.org/10.1145/3341163.3347735

[10] Sourav Bhattacharya and Nicholas D Lane. 2016. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *2016 IEEE International conference on pervasive computing and communication workshops (PerCom Workshops)*. IEEE, 1–6.

[11] Lora E Burke, Jing Wang, and Mary Ann Sevick. 2011. Self-monitoring in weight loss: a systematic review of the literature. *Journal of the American Dietetic Association* 111, 1 (2011), 92–102.

[12] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 39 (mar 2020), 30 pages. https://doi.org/10.1145/3380985

[13] Keum San Chun, Ashley B. Sanders, Rebecca Adaimi, Necole Streeper, David E. Conroy, and Edison Thomaz. 2019. Towards a Generalizable Method for Detecting Fluid Intake with Wrist-Mounted Sensors and Adaptive Segmentation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 80–85.

https://doi.org/10.1145/3301275.3302315
[14] B. Clarkson and A. Pentland. 1998. Extracting context from environmental audio. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*. 154–155.
[15] Anind K Dey, Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos. 2011. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*. 163–172.
[16] Katherine Ellis, Suneeta Godbole, Jacqueline Chen, Simon Marshall, Gert Lanckriet, and Jacqueline Kerr. 2013. Physical activity recognition in free-living from body-worn sensors. *ACM International Conference Proceeding Series*, 88–89. https://doi.org/10.1145/2526667.2526685
[17] Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, and Gert Lanckriet. 2014. Multi-sensor physical activity recognition in free-living. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 431–440.
[18] Avgoustinos Filippoupolitis, William Oliff, Babak Takand, and George Loukas. 2017. Location-enhanced activity recognition in indoor environments using off the shelf smart watch technology and BLE beacons. *Sensors* 17, 6 (2017), 1230.
[19] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.
[20] John J Guiry, Pepijn Van de Ven, and John Nelson. 2014. Multi-sensor fusion for enhanced contextual awareness of everyday activities with ubiquitous devices. *Sensors* 14, 3 (2014), 5687–5701.
[21] Yujiao Hao, Rong Zheng, and Boyu Wang. 2021. Invariant Feature Learning for Sensor-based Human Activity Recognition. *IEEE Transactions on Mobile Computing* (2021).
[22] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135.
[23] D. Iakovakis, R. E. Mastoras, S. Hadjidimitriou, V. Charisis, S. Bostanjopoulou, Z. Katsarou, L. Klingelhoefer, H. Reichmann, D. Trivedi, R. K. Chaudhuri, L. J. Hadjileontiadis, S. D, and J. D. 2020. Smartwatch-based Activity Analysis During Sleep for Early Parkinson's Disease Detection. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 4326–4329.
[24] Masato Ishii, Takashi Takenouchi, and Masashi Sugiyama. 2020. Partially zero-shot domain adaptation from incomplete target data with missing classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3052–3060.
[25] Hyunchoong Kim, Jonghoon Shin, Soohwan Kim, Yohan Ko, Kyoungwoo Lee, Hojung Cha, Seong-il Hahm, and TaeJun Kwon. 2016. Collaborative classification for daily activity recognition with a smartwatch. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 003707–003712.
[26] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.
[27] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark Plumbley. 2019. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition.
[28] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609
[29] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 338, 13 pages. https://doi.org/10.1145/3290605.3300568
[30] Jinna Lei, Xiaofeng Ren, and Dieter Fox. 2012. Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 208–211.
[31] Dawei Liang, Wenting Song, and Edison Thomaz. 2020. Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–10.
[32] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (March 2019), 18 pages. https://doi.org/10.1145/3314404
[33] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition.. In *IJCAI*. 3109–3115.
[34] A. Madabhushi and J. K. Aggarwal. 1999. A Bayesian approach to human activity recognition. In *Proceedings Second IEEE Workshop on Visual Surveillance (VS'99) (Cat. No.98-89223)*. 25–32. https://doi.org/10.1109/VS.1999.780265
[35] T. Maekawa, Y. Kishino, Y. Yanagisawa, and Y. Sakurai. 2012. WristSense: Wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 510–512.

[36] D. Minnen, T. Starner, J. A. Ward, P. Lukowicz, and G. Troster. 2005. Recognizing and Discovering Human Actions from On-Body Sensor Data. In *2005 IEEE International Conference on Multimedia and Expo*. 1545–1548. https://doi.org/10.1109/ICME.2005.1521728

[37] Dan Morris, T Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3225–3234.

[38] Alessandra Moschetti, Laura Fiorini, Dario Esposito, Paolo Dario, and Filippo Cavallo. 2016. Recognition of daily gestures with wearable inertial rings and bracelets. *Sensors* 16, 8 (2016), 1341.

[39] Alessandra Moschetti, Laura Fiorini, Dario Esposito, Paolo Dario, and Filippo Cavallo. 2017. Daily activity recognition with inertial ring and bracelet: An unsupervised approach. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3250–3255.

[40] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 158–165.

[41] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[42] Shwetak N Patel, Julie A Kientz, Gillian R Hayes, Sooraj Bhat, and Gregory D Abowd. 2006. Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In *International conference on ubiquitous computing*. Springer, 123–140.

[43] Donald J Patterson, Dieter Fox, Henry Kautz, and Matthai Philipose. 2005. Fine-grained activity recognition by aggregating abstract object usage. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. IEEE, 44–51.

[44] A. Physics. 2010. Nyquist–Shannon Sampling Theorem.

[45] H. Pirsiavash and D. Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2847–2854.

[46] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–27.

[47] Attila Reiss and Didier Stricker. 2012. Creating and benchmarking a new dataset for physical activity monitoring. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*. 1–8.

[48] Keum San Chun, Hyoyoung Jeong, Rebecca Adaimi, and Edison Thomaz. 2020. Eating episode detection with jawbone-mounted inertial sensing. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 4361–4364.

[49] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga, and Ozlem Durmaz Incel. 2015. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 591–596.

[50] Shuangquan Wang, Jie Yang, Ningjiang Chen, Xin Chen, and Qinfeng Zhang. 2005. Human activity recognition with user-free accelerometers in the sensor networks. In *2005 International Conference on Neural Networks and Brain*, Vol. 2. 1212–1217. https://doi.org/10.1109/ICNNB.2005.1614831

[51] Nabeel Siddiqui and Rosa HM Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *Plos one* 15, 1 (2020), e0227039.

[52] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras. 2012. Audio-based human activity recognition using Non-Markovian Ensemble Voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 509–514.

[53] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. 2004. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In *Pervasive Computing*, Alois Ferscha and Friedemann Mattern (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 158–175.

[54] Edison Thomaz. 2020. Activiome: A System for Annotating First-Person Photos and Multimodal Activity Sensor Data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 1–6.

[55] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing* 16, 4 (2017), 62–74.

[56] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–22.

[57] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[58] Jamie A Ward, Paul Lukowicz, Gerhard Troster, and Thad E Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1553–1567.

[59] Gary M. Weiss, Jessica L. Timko, Catherine M. Gallagher, Kenichi Yoneda, and Andrew J. Schreiber. 2016. Smartwatch-based activity recognition: A machine learning approach. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. 426–429. https://doi.org/10.1109/BHI.2016.7455925

[60] Stephen Xia and Xiaofan Jiang. 2020. PAMS: Improving Privacy in Audio-Based Mobile Systems. In *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things (AIChallengeIoT '20)*. Association for Computing Machinery, New York, NY, USA, 41–47. https://doi.org/10.1145/3417313.3429383

[61] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. Deepfusion: A deep learning framework for the fusion of heterogeneous sensory data. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 151–160.

[62] Hui-Shyong Yeo, Gergely Flamich, Patrick Schrempf, David Harris-Birtill, and Aaron Quigley. 2016. RadarCat: Radar Categorization for Input & Interaction. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 833–841. https://doi.org/10.1145/2984511.2984515

[63] Cheng Zhang, AbdelKareem Bedri, Gabriel Reyes, Bailey Bercik, Omer T Inan, Thad E Starner, and Gregory D Abowd. 2016. Tapskin: Recognizing on-skin input for smartwatches. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*. 13–22.