VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors

PRASOON PATIDAR, Carnegie Mellon University, USA
MAYANK GOEL, Carnegie Mellon University, USA
YUVRAJ AGARWAL, Carnegie Mellon University, USA

The use of audio and video modalities for Human Activity Recognition (HAR) is common, given the richness of the data and the availability of pre-trained ML models using a large corpus of labeled training data. However, audio and video sensors also lead to significant consumer privacy concerns. Researchers have thus explored alternate modalities that are less privacy-invasive such as mmWave doppler radars, IMUs, motion sensors. However, the key limitation of these approaches is that most of them do not readily generalize across environments and require significant in-situ training data. Recent work has proposed cross-modality transfer learning approaches to alleviate the lack of trained labeled data with some success. In this paper, we generalize this concept to create a novel system called VAX (Video/Audio to ‘X’), where training labels acquired from existing Video/Audio ML models are used to train ML models for a wide range of ‘X’ privacy-sensitive sensors. Notably, in VAX, once the ML models for the privacy-sensitive sensors are trained, with little to no user involvement, the Audio/Video sensors can be removed altogether to protect the user’s privacy better. We built and deployed VAX in ten participants’ homes while they performed 17 common activities of daily living. Our evaluation results show that after training, VAX can use its onboard camera and microphone to detect approximately 15 out of 17 activities with an average accuracy of 90%. For these activities that can be detected using a camera and a microphone, VAX trains a per-home model for the privacy-preserving sensors. These models (average accuracy = 84%) require no in-situ user input. In addition, when VAX is augmented with just one labeled instance for the activities not detected by the VAX A/V pipeline (~2 out of 17), it can detect all 17 activities with an average accuracy of 84%. Our results show that VAX is significantly better than a baseline supervised-learning approach of using one labeled instance per activity in each home (average accuracy of 79%) since VAX reduces the user burden of providing activity labels by 8x (~2 labels vs. 17 labels).

CCS Concepts: • Human-centered computing → Ambient intelligence; • Computer systems organization → Sensors and actuators.

Additional Key Words and Phrases: ubiquitous sensing, privacy first design, human activity recognition

ACM Reference Format:

1 INTRODUCTION

Human Activity Recognition (HAR) within home environments enables compelling applications around Active and Assisted Living (AAL), healthcare monitoring, security and surveillance, and tele-immersion applications [85]. As a result, over the last few decades, researchers built more accurate and practical HAR systems driven by...
inexpensive sensors [14, 56] and rapid advances in ML algorithms [84]. Given the advances in computer vision and audio processing and the presence of several large annotated datasets (e.g., [19], [40], [2]), researchers have built ML models that can perform audio-based [29, 54], and video-based activity and body pose estimation [66, 94]. Notably, these pre-built ML models for Audio/Video (A/V) data are meant to be generalizable and work reasonably well in new outside-the-lab settings without requiring any in-situ training data. Consumers, however, are increasingly finding IoT sensors, particularly cameras and microphones in their personal spaces invasive [1, 16, 32, 33, 48]. The problem is further compounded, given numerous high-profile incidents of compromised IoT devices [6, 7, 39, 87].

Apart from microphones and cameras, researchers have also explored other sensing modalities, such as IMUs (accelerometers, gyroscopes), low-fidelity thermal arrays, high-fidelity thermal image sensors, WiFi signals, Doppler RADARs, and a slate of environmental sensors such as temperature, humidity, and PIR for movement [9, 10, 55, 56, 58, 91]. The key advantage of these approaches over video and audio-based HAR is that they are more privacy sensitive, and the data captured is often not personally identifiable, alleviating many privacy concerns. However, these approaches still suffer from (a) requiring in-situ training data for building supervised learning models which comes with a significant user burden; and (b) not enough training data to create generalized models that work in the real world without training. Most recently, researchers have explored domain adaptation [20], i.e., converting training data from a richer (but more invasive) sensing modality to a more privacy-sensitive modality. For example, Ahuja et al. created synthetic Doppler RADAR data from videos captured to train a model for a set of motion-related activities [5]. Such approaches are still limited to a small set of activities (e.g., activities with large body motion or activities that generate descriptive sound or motion signals). To build an approach that generalizes to a wider range of activities, we often have to rely on multimodal sensing but generating synthetic training data for all these modalities is not yet possible.

In this paper, instead of generating synthetic signals, we propose VAX, a hybrid approach that uses output labels from off-the-shelf models to perform in-situ training of specific sensors. VAX allows to quickly learn in-situ ML models for a variety of ‘X’ privacy-sensitive sensors (e.g., Lidar, mmWave radar, low fidelity thermal cameras) using training labels obtained from models built on existing Audio and Video (A/V) datasets. This approach can go beyond the set of sensors explored in this paper and can generalize to a wide range of sensors without relying on mathematical or empirical modeling across different signals. We envision when starting to use VAX, users would augment it with a camera and microphone for a small duration (on the order of days). The camera and microphone, during this period, will use existing audio/video-based ML models to generate training labels for activities performed in the user’s environment and train the rest of the ‘X’ sensors on VAX. Once the privacy-sensitive sensors are trained, VAX can inform the user to remove the microphone and camera altogether. The design of VAX is based on the following key insights: (1) current generation pre-built A/V models are reasonably good at identifying a variety of HAR activity patterns [29, 30]; (2) these A/V models are usually generalizable and work across environments; (3) combining multiple privacy sensitive sensors allows us to leverage the sensor data for each modality to build accurate per-home models for a diverse set of activities; (4) a limited set of activities, where the A/V models are inaccurate, need a small amount of in-situ training data to improve accuracy further.

In designing and implementing VAX¹, we had to solve three major technical challenges. (1) while there exist a number of off-the-shelf A/V-based ML models, they are trained with a diverse (and non-exhaustive) set of labels and are not uniformly accurate across all activities [30, 62], which makes the selection of appropriate pre-built A/V models and combining their outputs non-trivial. (2) overall accuracy of a wide range of A/V models on native datasets is less than 75-80% [69], and as a result, incorrect labels from sub-optimal A/V models will lead to inaccuracies in training the models on the privacy-sensitive sensors, and (3) data collected from different privacy

¹ [www.github.com/synergylabs/vax](http://www.github.com/synergylabs/vax)
sensitive modalities is heterogeneous, thus, how to train user level activity recognition models for any set of ‘X’ privacy sensitive modalities.

To address these challenges, VAX bootstraps a set of off-the-shelf A/V-based ML models with some labeled data from a set of starter homes and then uses this ensemble of models to predict activities in new homes. We call this part of VAX the A/V pipeline. We then propose an unsupervised learning approach that utilizes unlabelled data from ‘X’ sensors to increase our activity detection rate and reduce the impact of erroneous labels from the A/V pipeline. In addition, we propose an approach to train activity recognition models across privacy-preserving modalities using the labels provided by the A/V pipeline. We collected data for 17 activities in 10 participant’s homes and show that VAX can detect activities with an accuracy of 74% across all activities, and 84% across activities detected by our A/V pipeline (15 out of 17). Given that existing A/V models cannot yet accurately detect all activities across all homes, we propose that the user provides one labeled example for each “undetected” activity (approximately 2 activities per home). With such minimal user input, VAX’s accuracy improves to approximately 84% (as compared to a baseline case of 79% where the user provides an input for each of the 17 activities and does not benefit from the bootstrapping provided by A/V pipeline). Note, the baseline also incurs 8× user burden for providing training labels (~2 labels vs. 17 labels).

In summary, we make the following contributions:

- We present VAX, an end-to-end framework for training activity recognition models for a set of ‘X’ privacy-sensitive sensing modalities using labels from off-the-shelf A/V models.
- We present a novel method to bootstrap and combine these off-the-shelf A/V models with the unlabelled data from ‘X’ modalities to detect activities of daily living in a new home.
- We evaluated VAX with 10 participants performing 17 activities in three different locations (Kitchen, Living Room, and Bathroom) in their homes. We show that VAX can detect activities with an average accuracy of 84% using labels generated by VAX’s A/V pipeline and only one input for activities which are not present in our A/V labels (~2 labels vs. 17 labels).
- We show that without any user input, VAX performs considerably better (74% vs. 38%) when compared with baseline approach (i.e., directly training models for activity recognition on privacy-sensitive sensors). We also show that with user input, VAX has comparable performance (84% vs. 79%) with significantly less labeling effort (2 labels/home vs. 17 labels/home) from users.

2 RELATED WORK

We have organized the related work into three categories. First, we review prior work on Human Activity Recognition (HAR) using audio and visual data as well as other privacy-sensitive modalities. Next, we discuss domain adaptation techniques across heterogeneous sensing modalities to address the scarcity of labeled data in the target domain. Finally, we discuss self-supervised learning (SSL) approaches in HAR, which focus on learning informative representations from unlabelled data, and augmenting it with a small amount of labeled data to build accurate ML models.

2.1 Human Activity Recognition (HAR) using Different Sensing Modalities

HAR is a well-researched problem, with researchers proposing different sensing modalities and various ML methods to detect activities. With rapid advances in deep learning and the availability of large-scale video datasets with 100s of annotated labels [19, 42, 92], the use of video data for HAR has become very popular (See [77] for a comprehensive survey). Notably, platforms such as OpenMMLab have emerged for video understanding (i.e. MMAction2 [69], MMPose [70], MMDetection [23], etc.) which leverage pre-trained, and open-source, state-of-the-art action recognition models using video data. Similarly, acoustic (sound-based) activity recognition has also been explored for generalized HAR across various settings [29, 54, 71]. The approaches use microphone data to
detect activities with unique sound signatures. Often these audio-based approaches convert sound signals into a 2-D image using Fast Fourier Transforms (FFTs) and then generate a feature representation using pre-trained deep CNN-based architectures. Large-scale datasets like AudioSet [37] and YouTube-8M [2] consist of labels for hundreds of diverse activity classes and are used to train generic audio-based models. While using audio and video data is indeed promising for HAR, their widespread adaptability is limited due to significant privacy and security concerns [16, 79]. In addition, the real-world accuracy for these models individually for common HAR tasks is still not that high [69].

Researchers have also proposed other sensing modalities such as mmWave Doppler Radar, IMUs, Lidars, thermal arrays, pressure, humidity, temperature etc. for HAR [5, 12, 46, 53]. While these sensing modalities may seem less capable than video and audio based HAR, they have a distinct advantage in terms of reducing privacy concerns since the data sensed is generally not personally identifiable [83]. Building upon these approaches, multi-modal sensors and sensor fusion has also been proposed for more accurate HAR [4, 68, 73, 75, 78]. Some works also explore generative and adversarial models to generate synthetic data to reduce privacy burden by hiding user identity while supporting a particular class of applications [64, 102]. Most of these supervised learning-based approaches rely on significant in-situ labeled data to provide high-accuracy HAR. In addition, they are tested in controlled settings only and do not generalize well across different environments [18].

2.2 Domain Adaption Methods for Cross-Modality Learning

To reduce the labeling burden, researchers have explored domain adaption to transform raw data from one domain (source) to another (target) domain [21]. Relevant to our work, some of these domain adaptation techniques use video or audio data as the source domain [5, 8, 15, 53, 82, 97]. Vid2Dopler [5], for example, created a synthetic training data set for Doppler sensors using videos of exercises as examples. IMU2Doppler [12] explored domain adaptation for privacy-sensitive modalities to learn models for the Doppler sensor data using the IMU data as a source domain. While these cross-modality approaches have inspired our work, they are limited to specific kinds of activities, as the target sensing modality can only sense motion. An approach that can detect a wider swath of activities would most probably require multimodal sensing (e.g., [56]), but with a wider set of sensors generating synthetic training data becomes harder. Thus, there is a need for an approach that can train a target sensing modality even if a direct physical relationship does not exist between the source and target sensing domains.

2.3 Self-Supervised Approaches to Reduce Labeling Effort

Researchers have also explored self-supervised learning (SSL) methods to minimize reliance on labeled data (See [26] for a comprehensive survey on SSL methods for temporal and multimodal data). ColloSSL [47], COCOA [27], CPC [41], FedCLR [81] are recent techniques to employ SSL for the HAR domain. The two works closest to us in terms of their vision are ColloSSL [47], which focuses on SSL across a time-synchronous multi-device setting, and FedCLR [81], which uses a combination of active learning and label propagation to provide pseudo labels to large amounts of unlabeled data. These approaches utilize deep learning architectures for label propagation and representation learning in a multi-device setting. However, these approaches target single sensing modalities for SSL or assume a correlation across different sensing modalities. VAX used an alternative approach in contrast to self-supervised learning, where unlabeled data is processed to obtain useful representations to get pseudo labels. In VAX, we deploy rich (but privacy-invasive) sensing modalities like video and audio sensing with other privacy-sensitive modalities for a brief amount of time and learn (pseudo) labels from global models built on these (audio and video) sensors. Our approach generates labels independent of what privacy-sensitive sensing is being used. Thus, we do not need to make assumptions about or restrict the selection of privacy-sensitive modalities.
In this section, we describe the design of our VAX hardware apparatus which includes both audio/video sensors as well as a selection of privacy-sensitive sensors. For the audio and video sensing modalities, we used a Yeti USB microphone and a Logitech USB web camera with a 720p resolution. In selecting the suite of privacy-sensitive sensors, we looked at prior works in HAR [43, 49, 51, 57, 84, 98] and considered aspects such as sensing range, a breath of signals captured, fidelity and sensitivity of the sensed signals. We describe below the different sensing modalities that are present on our VAX apparatus and our rationale for including them in our testing.

- **Movement sensing using Doppler radars**: mmWave Doppler radars are promising for HAR, especially now as they become more affordable [24, 34, 59, 67, 72, 106, 107]. These radar sensors emit a reference RF signal, and measure the reflected signal back from objects in the vicinity. Whenever there is movement, such as by an object or by a person doing some activities, the reflected signals are Doppler-shifted, which can be used to create a 1-D Doppler plot. With frequency modulation (FMCW), a 2-D plot of range vs. the Doppler plot can be produced to characterize the type of motion at different ranges (See Figure 1).

- **Position sensing using a Lidar**: Lidar (light detection and ranging) uses the Time of Flight (ToF) or parallax of a laser beam to perform range finding. Most Lidar sensors are designed to sense range across one axis. However, electromechanical (spinning on a single axis) Lidars can detect a range of objects in 360 degrees. Lidar is common in robotics to detect and avoid obstacles and on mobile embedded systems for localization at a given timestamp for various kitchen activities.

![VAX HARDWARE DESIGN](image)

![VAX HARDWARE DESIGN](image)

Fig. 1. Visualizing information across various activities from various privacy-sensitive sensors. (a) shows the movement range and radial velocity of moving objects using a Doppler sensor and object distance and angle using a Lidar sensor in the sensor plane across a sequence of activities happening in the living room. (b) shows snapshots of Thermal and Sound Localization at a given timestamp for various kitchen activities.

3 VAX HARDWARE DESIGN

In this section, we describe the design of our VAX hardware apparatus which includes both audio/video sensors as well as a selection of privacy-sensitive sensors. For the audio and video sensing modalities, we used a Yeti USB microphone and a Logitech USB web camera with a 720p resolution. In selecting the suite of privacy-sensitive sensors, we looked at prior works in HAR [43, 49, 51, 57, 84, 98] and considered aspects such as sensing range, a breath of signals captured, fidelity and sensitivity of the sensed signals. We describe below the different sensing modalities that are present on our VAX apparatus and our rationale for including them in our testing.

- **Movement sensing using Doppler radars**: mmWave Doppler radars are promising for HAR, especially now as they become more affordable [24, 34, 59, 67, 72, 106, 107]. These radar sensors emit a reference RF signal, and measure the reflected signal back from objects in the vicinity. Whenever there is movement, such as by an object or by a person doing some activities, the reflected signals are Doppler-shifted, which can be used to create a 1-D Doppler plot. With frequency modulation (FMCW), a 2-D plot of range vs. the Doppler plot can be produced to characterize the type of motion at different ranges (See Figure 1).

- **Position sensing using a Lidar**: Lidar (light detection and ranging) uses the Time of Flight (ToF) or parallax of a laser beam to perform range finding. Most Lidar sensors are designed to sense range across one axis. However, electromechanical (spinning on a single axis) Lidars can detect a range of objects in 360 degrees. Lidar is common in robotics to detect and avoid obstacles and on mobile embedded systems for localization at a given timestamp for various kitchen activities.
guidance. However, Lidar can be repurposed to detect activities with positional relevance (i.e., sitting on the sofa could imply relaxing or sitting at the dining table could imply having a meal). SurfaceSight [55], for example, utilizes Lidar to enable object detection and user interactions. Figure 1 shows an example of how Lidar signal can vary based on different human activities.

- **Thermal sensing (Infrared Sensors):** Thermal/infrared cameras are widely used for low light scenarios since they can detect IR (serves as a heat signature) from all kinds of objects. Various versions exist to support numerous applications such as in-home security, personal safety, energy efficiency, pest control, home maintenance, and automotive care. However, with sufficiently high resolution, thermal sensors can also raise privacy concerns even if the data is not personally identifiable [86]. In VAX, we used a low resolution (10x8) thermal sensor, which does not reveal any Personally Identifiable Information (PII), but still can provide signals on activities with a thermal signature of appliances such as a fridge or a stove (see Figure 1).

- **Sound localization (Micarrays):** Unlike full fidelity audio, being able to detect the direction from which sounds are coming (i.e. localize the sound) and their intensity (i.e. dB level) is much more privacy-preserving while being helpful to detect activities around appliances. This can be done using an array of microphones spatially separated on a Printed Circuit Board (PCB), and measuring the difference in arrival rate to infer the 3-D location of sound sources (see Figure 1). However, even the mere presence of a microphone, even if it is only inferring the location and the intensity of sound may still raise privacy concerns in case it is hacked to capture full-fidelity audio. We believe that in the future, sensors that are guaranteed to be limited to certain functionality in hardware are going to be possible [100] and hence we have included a commercial off-the-shelf sound localization sensor called the MicArray [105] on our current VAX rig.

- **Sensing ambient vibrations:** Inertial Measurement Units (IMUs), which includes an accelerometer, a gyroscope, and a magnetometer are often used for HAR on wearable devices [60, 71] or even in the ambient environment such as a wall socket [56]. We include an IMU on our VAX hardware to similarly sense ambient vibrations, coupled through the wall and other structural elements when performing different activities such as exercising or operating appliances such as a vacuum cleaner or washer/dryer. Other people have also looked at finer-grained activities of daily living through vibrational sensing [76].

- **Environmental sensors (Temperature, Humidity, Pressure, Light, Color, RSSI, Passive Infra-Red):** Some common human activities result in changes to the environment that can be picked up by ambient sensors such as temperature (e.g. using a space heater) or humidity (activities in the bathroom) or a light sensor (watching TV, turning on lights) [28, 56]. As a result, we utilize the Mites.io [13, 68] multi-modal sensor hardware which incorporates twelve different sensing dimensions. Notably, while the Mites device does provide highly featurized data from an on-board microphone, we do not use that data in VAX.

Figure 2 shows our current VAX hardware prototype. While all of our chosen privacy-sensitive sensors can be integrated onto a single board, we chose to build our prototype using off-the-shelf hardware modules and sensor evaluation boards. We limit ourselves to ambient sensors to reduce the user burden of using any smart (or sensor instrumented) devices or wearable sensors. However, utilizing additional sensing modalities can bring interesting opportunities and challenges which is beyond the scope of this paper. We used a Logitech USB Webcam (C910, 720p resolution) as well as an iPhone (as a backup data source) for both audio and video data. We use the AWR1642BOOST-ODS [44], a 77GHz mmWave evaluation board with an integrated DSP and an ARM Cortex-R4 Processor from Texas Instruments. This evaluation board incorporates a patch-antenna structure that provides a wide field of view (FOV) in both azimuth and elevation planes. The integrated microcontroller on this board is, however able to provide data at a relatively low frequency of 5Hz, which we overcome by adding a separate DCA1000 FPGA board [45]. This board connects directly to the ADCs of the radar board and extracts the samples of the doppler signals generated in a binary format. Samples collected with the DCA 1000 EVM board are sent via
VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors

Fig. 2. VAX hardware apparatus. Doppler, Thermal, and the Mites are fixed vertically, and Lidar and Micarray are kept horizontally. The setup is mounted on a box with a combined USB and Ethernet switch enabling data transfer.

Fig. 3. High-level architecture of our VAX training and inference engine, highlighting the key components. VAX recognizes environmental context during the training phase using a set of off-the-shelf audio/video models and generates labels to train privacy-preserving sensors for recognizing activities. A/V sensors are removed and the trained VAX models for privacy-preserving sensors are deployed in the environment.

UDP to a small form factor PC (Intel NUC). This raw data is further processed to create a range-doppler heatmap over time with a sampling rate of 15 Hz. For thermal sensing, we use a FLIR Lepton 3.5 thermal camera with FLIR Systems 3.2a radiometry [3], which is interfaced with a Purethermal Mini Development board to enable communication via USB. This sensor captures high-fidelity thermal data (160x120 pixels) at 8Hz. We observed that the high spatial resolution of this sensor can visibly identify and differentiate between two users based on their physical features (i.e., height, weight, facial features, etc.). To alleviate such privacy concerns, we reduced the resolution of the thermal sensor by 16x to just 10x8 pixels. We attempted to use the Panasonic Grid-Eye sensor, which provides an 8x8 thermal resolution, but the data quality and range were not great so we used the FLIR sensor instead, and down-sampled it as a proxy of a very low-resolution thermal sensor. We use the Slamtec RPLIDAR A1M8 [50], a mechanical LiDAR to capture object distance in a single horizontal plane. It is a two-dimensional Lidar sensor with a maximum range of 6m. The horizontal scanning range is 0°-360°. The scanning frequency is 5.5 Hz when sampling 360 points on each revolution with an angular resolution of 1°. We also tried to use a 3D-Lidar to capture object distances; however, we chose not to use it due to the significant latency for it to provide stable readings and intermittent data during periods of high movement. For sound
localization, we use a ReSpeaker 4-Mic Array with a Raspberry Pi [105], which is a quad-microphone expansion board for Raspberry Pi designed for AI and voice applications. We only capture output for sound localization and intensity at very low resolution (1Hz) from this sensor. Finally, we use a multi-modal sensor board, the Mites [13, 68], to sense vibrations (using its IMU) and a range of environmental data (i.e. temperature, humidity, pressure, light intensity, and color, movement based on PIR, low-resolution IR). Our VAX rig is wall-powered, and all the sensors are either connected over USB or Ethernet with a small form factor PC (Intel NUC, 8-core, 16 GB RAM) to collect and store processed data.

4 VAX MACHINE LEARNING AND SOFTWARE SETUP

In Figure 3, we show the high-level software architecture for VAX. We have separated the ML components into the training and deployment phases. In our vision, for the initial training phase, we deploy a camera with a microphone along with privacy-sensitive sensors in a user’s home. Our Audio-Video (A/V) pipeline uses off-the-shelf pre-trained ML models to detect activities and generates training labels. Then, our privacy-sensitive pipeline uses these generated labels from the A/V pipeline to train ML models for privacy-sensitive sensors. Subsequently, during the deployment phase, the camera is physically removed and the A/V is not used anymore, and the trained models on the privacy-sensitive sensors are used for activity detection. We now go into the details for each of these phases.

4.1 Deployment and Data Collection

We deployed our VAX prototype and collected data for participants in their own homes. For each home, we deployed our prototype in three locations (kitchen, bathroom, and living room) to capture the diversity of activities. We started with a union set of 30+ activities from recently published activity recognition literature [54, 56, 71]. We filtered activities from these sets based on two criteria, (i) we removed activities that do not belong to the three location contexts we have considered, i.e., activities like drilling, using screwdrivers, etc., and (ii) we removed activities that do not have any movement or audio signature, like sitting, or sleeping. We intentionally kept a wide range of high-level activities (i.e., Cooking, Exercising, etc.) to allow for diverse movement and sound patterns. We finalized a list of 17 activities. For the kitchen, we collected samples for activities such as using the oven (Baking), using a blender/mixer (Blender), heating food in the microwave (Microwave), using a fridge (FridgeOpen), chopping and grating vegetables/fruits (Chopping+Grating), cooking a meal on the stove (CookingOnStove), washing dishes in the sink (WashingDishes). For the bathroom, we collected data on activities like using a hair brush (HairBrush), using an electric hair dryer (HairDryer), washing hands (HandWash), using an electric shaver (ShaverInUse), and flushing the toilet (ToiletFlushing). Finally, we collected samples for activities that can happen anywhere in a home, such as eating or drinking (Eating/Drinking), exercising (Exercising), coughing (Coughing), and operating a vacuum cleaner (Vacuum). We asked the participant to perform these activities in their living room. All activities are performed by a single participant in their own home. We collected data from 10 participants, and the sample size is based on a review of similar participant studies in HAR literature [12, 25]. Each study took around 4-5 hours, including the time to carry the VAX rig to different homes and to set it up at multiple locations in the home, which resulted in approximately 10 hours of data collection across all participants. Our study design was approved by our Institutional Review Board (IRB).

4.2 Activity Recognition Using Off-the-shelf A/V Models

In this section, we present a novel approach for discovering good label-activity matches from any set of off-the-shelf A/V models. A straightforward approach to collect activity labels from A/V modalities is to use the best-performing models for each or do a confidence vote across multiple A/V-based ML models. We shortlisted twenty popular off-the-shelf A/V models [11, 29, 31, 35, 36, 38, 40, 61, 62, 65, 71, 89, 90, 93, 95, 96, 99, 101, 103, 104, 108]
based on open-source availability of the models, including their pre-trained weights, as well as their performance benchmark against publicly available datasets [19, 42, 92]. However, our empirical evaluation of these off-the-shelf A/V models, used as-is, on the data we collected for our chosen 17 activities of daily living did not provide accurate results. Some of the reasons for this inaccuracy include: (i) the models are trained to provide labels for activities that have overlapping signatures (e.g., blenders in a kitchen or tools like a power drill). (ii) the models are trained for HAR in different settings and thus do not have a clear one-to-one mapping for common activities of daily living (e.g., atomic actions like sitting can attribute to multiple human activities). Merely combining all the outputs of different models is also non-trivial since it needs to consider that different models may be better suited to certain activities vs. others. Based on our experiments with these twenty A/V models, we make the following observations:

- Skeleton-based action recognition models work consistently across different settings as their input (i.e., human pose) is environment agnostic.
- No single A/V model provides consistent labels across the 17 activities we tested on.
- The same skeleton-based action recognition model trained on different types of large-scale datasets can learn multiple activity signatures and perform well on different categories of activities.
- Many audio-based models perform consistently well to identify a similar set of activities (e.g., using an electric shaver, blender, microwave), i.e., we do not need more than one audio model for our ensemble of A/V models.
- The capability of video models to detect activities and confidence depends on the vantage point (e.g., front or side view) and occlusions.
- The labels provided by existing models may not have a 1-to-1 correspondence with each other or even with our chosen activities, requiring reconciliation. For example, “Exercising” might be recognized by one video model and as “Doing Jumping Jacks” by another model.

For building our A/V pipeline, we finalized 2 off-the-shelf models for activity recognition using human pose estimation, POSEc3D [31] and STGCN [103] trained on the NTU60 [88] dataset with 60 activity classes. We further use 3 more copies of POSEc3D, each trained on different datasets, i.e., NTU120 [65], UCF101 [92] and HMDB51 [52] with 120, 101 and 51 different activity classes. For the off-the-shelf model for audio-based HAR, we use YAMNet [29], a deep network architecture that predicts 521 audio event classes based on the AudioSet-Youtube [37] corpus.

### 4.2.1 Building an A/V model from a set of training homes:

For our chosen set of off-the-shelf A/V models, we collect the top-k predictions for a set of activities across a set of homes (i.e., reference homes) that we use for training data. Using a shallow neural network classifier, we learn a non-linear mapping for each off-the-shelf A/V model from raw A/V labels to VAX activity labels. The intuition behind this approach is as follows: the top-k predictions from raw A/V labels might not have a static mapping to VAX activity labels, but they are weakly consistent across different home environments. This is true as audio models are agnostic to (home) environments, and video models are trained on large datasets like Youtube-8m [2], which contains observations for the same activity across a wide variety of home environments. We also observe that audio models show a more consistent set of top labels for single VAX activity across environments than video. Thus, we designed two ensembles comprising of different classifiers. The first ensemble (audio-only ensemble) consists of classifiers built from audio models, which works well on various activities with distinct audio signature (i.e., Blender, HairDryer, Vacuum, etc.). The second ensemble (audio-video ensemble) combines classifiers from both audio and video models to detect all the activities (see Figure 4). Combining the two (audio-only and audio-video) works better than using a single audio-video ensemble or two separate ensembles for audio-only and video-only off-the-shelf A/V models. These two ensembles are combined together to create a final VAX A/V model. To predict labels for activity instances in new homes, we run our set of top 5 (YAMnet, STGCN, POSEc3D+NTU120, POSEc3D+UCF, and
Off-the-shelf models deployed locally at each home and outputs raw labels \( (L_k) \) from off-the-shelf models with VAX activity labels \( (y^h \in A) \) are used as input to train global VAX A/V Model.

Posec3D+Hmdb51) off-the-shelf A/V models and collect the top-k labels for all models. These models are then passed through the audio-only and the audio-video ensemble to generate activity predictions, and the activity with the highest confidence value is chosen. To reduce the error in the VAX A/V pipeline, we generate confidence thresholds for each ensemble, and an activity instance is marked undetected if predictions from both ensembles are below these confidence values.

### 4.2.2 Improving A/V labels with data from the privacy-preserving sensing modalities

One challenge with our A/V pipeline is to set a cutoff threshold that optimizes both accuracy and the detection rate (i.e., the fraction of activities with a confidence score higher than the A/V model cutoff thresholds) for activities. Optimizing for a high-accuracy model with a low detection rate could miss detection on all instances for a particular activity, thus leading to no training labels for the next phase for VAX. In contrast, optimizing for high detection rates at the cost of accuracy would lead to more noisy generated labels from the A/V pipeline, affecting the accuracy of the privacy-sensitive sensor pipeline. To alleviate the issue of scarcity of good labels, we opportunistically utilize information from raw data collected from privacy-sensitive ‘X’ sensing modalities to increase the detection rate for activities. For example, (i) activities like FridgeOpen and Baking (or using an oven) will have clear separability in thermal signatures (see Figure 1), (ii) activities like Walking and Drinking/Eating will have a clear separability in their Doppler signatures. We can find similar pairs (or sets) of activities that provide a clean separability of activity signatures for a given sensing modality. We can leverage this information to increase the confidence score for an instance of an activity that did not receive any label from the A/V pipeline but looks very similar to another labeled instance. However, this would not work for activities that do get classified even once with high confidence from the A/V pipeline. For each ‘X’ privacy-sensitive modality, we identify activities, which are separable in that specific modalities feature space based on cluster membership of A/V predictions, i.e., if two or more high-confidence activity labels from A/V fall in the same cluster. Finally, we use these clustering models to enhance confidence scores for highly separable activities.

Figure 6 and 7 shows a 2-D embedding of clusters created using data from the Lidar and Thermal sensors, respectively. The different colored dark dots represent points clustered together into different clusters without...
Fig. 6. tSNE embedding (2-D) for supporting clusters using Lidar data to improve A/V pipeline labels.

Fig. 7. tSNE embedding (2-D) for supporting clusters using Thermal data to improve A/V pipeline labels.

Fig. 8. Increasing the detection rate of activity labels using support models from Lidar, Thermal, Doppler, and when using all three (Lidar+Thermal+Doppler) sensors.

any activity information, highlighted areas that correspond to dense regions for data from different activities, and grey points represent non-clustered points (or outliers). We observe that some activities like Walking (green shade region), Drinking/Eating (purple shaded regions) are clearly separable with Lidar data, whereas activities...
like Coughing (red shaded region) and Vacuum (blue shaded region) are closer to each other, and even confusing among the single cluster. We can see similar patterns in thermal data (i.e., good separation for Coughing and Vacuum and Drinking/Eating). This figure illustrates that Lidar data can help the VAX A/V pipeline to enhance the detection rate for low-confidence predictions for these activities. Figure 8 summarizes how support models for particular privacy-preserving sensors can help increase detection rates for different activities. Activities like Drinking/Eating and WashingDishes see a high detection rate by combining all the sensors (i.e., “AV+All”). Some activities like Exercising and HairDryer show no improvement by using any of the privacy-sensitive sensors. One reason for this is that none of the prominent clusters for these activities have instances that are marked ‘Undetected’ by the A/V pipeline.

4.3 Training Models for the Privacy Sensitive Sensors

For training models for privacy-sensitive ‘X’ sensors, the first challenge is making sure that there is training data for all activity classes.

4.3.1 Getting Labels for Undetected Activities using a Human-in-the-loop. At the end of the training phase, users can disconnect the A/V sensors and not use the A/V pipeline. At this point, the VAX pipeline will start to predict activities based on the models trained on the privacy-sensitive sensors only. However, a subset of activities might not be detected with high confidence using the A/V pipeline. These activities might also vary from one home to another. To improve the end-to-end accuracy of the VAX pipeline, we ask for limited human input for only these undetected activities in this phase. We solicit this input by asking users to perform a single instance of those undetected activity classes. As we show in Section 5, using only a single training instance for an average of 2 undetected activities (out of a total of 17 activities) for each home VAX can give an accuracy of 84%, which is better than a baseline approach where users provide one label for each of the 17 activities (accuracy of 79%), and with significantly less user burden.

4.3.2 Featurization of raw data from privacy-sensitive modalities: We train an ensemble of sensor-level models instead of a monolithic model that combines all sensing modalities. There is an opportunity cost of losing on added advantage of fusing multiple sensors; however, we prioritize reliability in long-term deployment. Such an approach would also allow for utilizing existing models (e.g., Vid2Doppler for Doppler RADAR [5]). Another benefit of using an ensemble of sensor-level models is the ease of featurization as different modalities have different kinds of featurization techniques, which are well-studied in prior literature [5, 55, 56]. For Thermal data, we use a rolling window for 5 contiguous images (8x10 pixels each) and use the maximum temperature value in °Celsius for each pixel, respectively. For the Doppler data, we use a rolling window for 20 contiguous Range-Doppler (256x32) heatmaps and break this down into two parts, (i) range-bins (i.e., distance from the sensor) of primary motion activity over time, (ii) average and variation (standard deviation) for different velocity bins over time, and concatenate the two to create final feature vector. For the Lidar data, we first remove the boundary of the room from the data, and then mark the distance of the closest object for all angles (0-360), and then take the maximum value over a rolling window of 10 contiguous samples. For the Micarray, we use a similar featurization technique as Lidar, with distance from a sound source weighted by the amplitude of the sound. As Micarray data is sampled at 1 Hz, we do not aggregate across multiple samples over time. For the other sensors (i.e., PIR, IMU, Environmental Sensors), we directly use featurized data (over 100ms) from multi-modal sensor board, mites [68].

4.3.3 Handling class imbalances in training labels: We use a minority oversampling approach called SMOTE [22] for data augmentation across activities, which removes a large number of instances. SMOTE (Synthetic Minority Oversampling Technique) works by selecting examples that are close to the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line. In general, data
augmentation leads to lower generalizable accuracy. However, we have a key advantage in our setting. Higher anomaly scores for labeling instances for a given activity lead to higher class imbalance but also suggest easier separability due to denser data distribution for normal samples. Thus, the classes that are clearly separable in a sensing modality benefit from outlier removal without suffering from errors due to sampling imbalance.

4.3.4 Reducing noise in training labels: To reduce noisy training labels, We use a confidence learning approach built on top of another approach proposed by Northcutt et al. [74], which addresses the problem of identifying noisy data in a general setting where no annotation information is available except the observed noisy labels. We start by calculating out-of-sample probabilities $P^{(x)}_{ik}$ for all data by training naive classifiers $\theta^{(x)}$ over featurized data across all 'X' sensors, respectively. For each sensor $x$, we calculate a confidence joint matrix, $C^{(x)}_{ij}$, to partition and count labels across multiple classes. Diagonal entries of $C^{(x)}_{ij}$ count correct labels, and non-diagonals capture asymmetric label error counts (As an example, $C^{(x)}_{i3,j1} = 10$ is read, as per sensor $x$, “Ten examples are labeled 3 but should be labeled 1”). This is further normalized to $Q^{(x)}_{ij}$ to make sure each sample row for $C^{(x)}_{ij}$ sums to the marginal probability of activity from $\theta^{(x)}$, and probability across all activities sums to 1. Finally, we classify samples with the lowest confidence calculate activity level threshold $\delta^{(x)}_i$, i.e., any sample predicted as activity label $i$ would be classified as noise if $P^{(x)}_{ii} < \delta^{(x)}_i$ for label noise characterization, and remove samples which are classified as noise across the majority of sensors. One critical hyperparameter for this approach is the selection of naive classifier $\theta^{(x)}$. We selected an ensemble of KNN and SVM with Gaussian kernel, which are shown to be the most robust classification methods for training with imperfect labels in prior literature [17], and are empirically shown to work best on our training data. Finally, we train the same classifier on clean data to provide the final trained model for each sensor, and for final activity predictions, we return the activity label for the highest prediction probability aggregated across all sensors.

4.4 Putting it together into an end-to-end VAX pipeline

Our entire pipeline for VAX is written in Python with over 10k lines of code and consists of three components, (i) training A/V model from reference homes, (ii) generating activity labels using trained A/V model and support models from ‘X’ sensors and (iii) training ‘X’ sensors with labels from A/V model. For training across all components, we use module implementation for a variety of machine learning methods (i.e., KNN, SVM, OPTICS, and SMOTE) from scikit-learn and imblearn libraries. Our noise reduction algorithm is built on top of Cleanlab [74], a generic package for identifying and reducing noise in datasets. We open-sourced our entire VAX system [80], which can be used out of the box with label predictions from any off-the-shelf A/V and can be extended to add more variety of ‘X’ sensors with no implementation changes.

5 EVALUATION AND RESULTS

We evaluated VAX to answer the following questions.

(1) How well does our A/V pipeline detect activities with a model trained on different sets of reference homes and predicts activity labels along with support models in a new home?

(2) How well does the VAX pipeline do in two scenarios?

(a) Using activity labels detected from the A/V pipeline with no user input.

(b) Using activity labels detected from the A/V pipeline along with one sample from user input on activities that are not detected by the A/V pipeline.

(3) How does VAX compare with a baseline approach in terms of accuracy and user burden?
5.1 Performance of A/V pipeline

Figure 9 shows the performance of our A/V pipeline (trained in leave one home out fashion) across participants for activities detected with our A/V pipeline (i.e., confidence score higher than cutoff values, 60% for audio-only ensemble and 40% for audio-video ensemble). We observed that 75% of the activities follow these criteria and thus are labeled using the A/V pipeline. Across these detected activities, our A/V pipeline shows high accuracy, precision, and recall. The information captured from the A/V modality is also consistent across multiple homes, i.e., an A/V model trained on a set of reference homes can capture activity patterns in new homes with support from privacy-sensitive sensors. Figure 12 shows the confusion matrix across activities predicted from the A/V pipeline and ground truth. We observe that A/V is good at identifying most activities with an observable sound signature, such as WashingDishes, Blender, etc. Our pipeline also does well for activities like HairBrush, or Exercising, which have unique motion signatures compared to other activities in the set. Our A/V pipeline has the highest confusion among activities like Baking and FridgeOpen. While both these activities involve large motion, they involve opening and closing doors, and the video-based ML models often confuse such activities. Apart from these two activities, any activity involving very little motion and insufficient sound, such as CookingOnStove, was hard to detect using audio or video-based models.

In Figure 14, we evaluate the impact of the number of reference homes on the accuracy and detection rate of activities from the A/V Model without any support from ‘X’ sensors. We observe that as the number of reference homes increases, the count of detected activities increases monotonically. We also observe that even with a low detection rate, the A/V pipeline has good accuracy, which shows that the A/V pipeline is reliable in detecting activities even with a smaller count of reference homes. This result means that when the labels generated by the A/V models are used to train other sensors, not all user actions will be labeled, and training might be slow. More importantly, given the precision of the A/V models is high, the labels will be accurate.

Figure 11(a) shows the F1-Score for our A/V pipeline across activities performed by all ten participants. We observe that our A/V pipeline can classify some activities (like Blender, WashingDishes, Exercising, etc.) across almost all homes. We also observe that across all homes, we are missing true detection for at least one activity (e.g., P3) and, at most, five activities (e.g., P7), and around two activities on average across all homes.
5.2 Performance of VAX pipeline

Once we get labels from the A/V pipeline, we train our privacy-sensitive sensors with the ‘X’ pipeline in leave-one-instance-out cross-validation for a given participant. Figure 11(b) shows the performance of VAX when trained on labels provided by the A/V pipeline and no user input. We see no detection for activities that the A/V pipeline could not detect accurately with VAX. Thus, we anticipate that when deployed in people’s homes, VAX will identify the undetected activities by looking at activities not present in A/V inferences and asking the user to provide one labeled instance for those undetected activities. Figure 11(c) shows F1-Score across all activities for all participants once we have one labeled instance (either from the A/V pipeline or a simulated human input). We observe that the VAX pipeline can reliably classify activities with only one sample for undetected classes. We also see that F1-score for activities included in the A/V pipeline improves. This happens due to the reduction in noise in overall training data with more ground truth labels.

Figure 10 shows an accuracy of 85%, precision of 85%, and recall of 85% for VAX with one label for each class (either automatically received from A/V models or simulated). This result highlights VAX’s capability of fusion of ‘X’ sensors to classify various activities successfully. Figure 13 shows the confusion matrix for activity detection from privacy-sensitive sensors. We observe that detection across almost all activities is good, showing a clean differentiating signature in one or more ‘X’ sensors. We observe that ‘X’ sensors do not perform well for some activities such as HairBrush, HairDryer, and Shaver In Use. These are all bathroom activities performed at roughly the same location. Given Doppler RADAR is sensitive to small motions, with more training samples, the models should be able to detect these activities. However, confirmation of this hypothesis remains a future work. Figure 15 shows the F1-score for activity detection across different sensors. We observe that for most of
the activities, Thermal, Doppler, and Lidar sensors show a good F1-score. This points out that in the future, even one of these sensors might give sufficient accuracy across all activities and be more practical in terms of cost. Other sensors like Micarray show good F1-scores across a subset of activities like HairBrush, Coughing, HairDryer, etc., where human movement is limited, and their positional difference is captured by Thermal, Lidar, or Micarray sensors, thus relying completely upon small variations in hand/torso movement from Doppler sensor.

5.3 Comparing VAX with other Baseline Approaches

Figure 16 shows a head-to-head comparison of activity detection for the VAX pipeline with a few baseline approaches. We consider two different baseline approaches: (i) pre-trained models on the privacy-sensitive ‘X’ sensors using data from reference homes and run prediction on a new home in a leave-one-home-out fashion, i.e., X-Only (pre-trained) and (ii) trained a model on the in-home using data from privacy-sensitive ‘X’ sensors by asking users to perform the activities and collect labels, i.e., X-Only (no pre-training). We compare these approaches with VAX in two different settings.

a) No user input is provided (Figure 16(a)): There are no labels to train X-Only (no pre-training) model when user input is not provided. In comparison to X-Only (pre-trained), VAX performs considerably better across all activities (74% vs. 38%). Furthermore, when we compare activities that are detected with our A/V pipeline (15 out of 17 on average), the performance of VAX increases from 74% to 84%, whereas there is a
b) User input is provided (Figure 16(b)): For the X-Only (no pre-training) case, we provided one label per activity, i.e., 17 labels/home to train models. For the X-Only (pre-trained), we start with the pre-trained models using data from the reference homes and then fine-tune the models with one label per activity in the test home, i.e., 17 labels/home. For VAX, we only provided one sample for the activities undetected.
by the A/V pipeline (2 labels/home on average). Even with considerably less labeling effort for VAX (2 labels/home vs. 17 labels/home), VAX provides better performance (accuracy of 84%) than both baseline approaches (accuracy of 79% and 78%). We also observe that VAX’s accuracy on all activities with user input on activities undetected with the A/V pipeline is similar to VAX’s accuracy without any user input on activities detected with the A/V pipeline, thus showing that VAX’s performance is limited only by the ability of off-the-shelf, pre-trained A/V models to separate chosen activities.

6 DISCUSSION
This section discusses the limitations of VAX and opportunities for future enhancements.

6.1 VAX hardware: modular design to disaggregate sensors
Our results indicate a limitation of our current VAX prototype. To simplify our deployment task, we integrated all the sensors, including the USB camera/microphone and privacy-sensitive sensors, and the Intel NUC compute node as a single monolithic rig. This rig could then be mounted on a tripod and carried around into different locations in the same home, across homes, and placed in a single location. However, that also meant a single vantage point for all the sensors in each room. Hence, we could not get any labels for some activities in some locations due to occlusion (e.g., the person’s back was to the A/V sensor, and furniture was in the way) or just poor angles. This is important since the A/V models are otherwise mostly agnostic to the vantage point, and if they are unable to provide labels, we are unable to train models for the privacy-sensitive sensors at all. In the future, making VAX modular so that at least the Audio/Video sensors can be placed at different locations could be beneficial to detecting all activities. Similarly, some privacy-sensitive sensors are directional (e.g., the Thermal sensor, the PIR movement sensor) with a field of view, while others are more sensitive depending on the location (e.g., the IMU to measure vibrations). Thus potentially having more than one of the privacy-sensitive sensors in a room at different locations could help boost accuracy. Furthermore, we can opportunistically place multiple sensors of the same sensing modality in different environments and train them together with a portable A/V sensing solution (i.e., using your phone to capture A/V labels) to extend activity recognition beyond a single location to the entire house with a single VAX system.

6.2 Real world in-the-wild study
Our current results, which show an average accuracy of 84% without any user input on a smaller set of activities and with around two user labels per home (for harder-to-detect activities) across all activities, showcase the challenge with accurate HAR. Notably, while we were able to achieve this level of accuracy by evaluating VAX in a diverse set of 10 homes, our data collection itself lasted for a few hours in each home where we asked the participant to do these activities while we collected Audio and Video data and the ground truth for each activity instance start/stop times. The data collection itself was a significant endeavor for our team. However, this data collection setup is still not what an in-the-wild deployment of a system such as VAX would look like. We imagine a scenario where the VAX rig is deployed in a person’s home for a particular duration (a day or two) while the participant goes about doing activities totally unconstrained. During this time, the models for the privacy-sensitive sensors would be trained and refined, and on the third day, the system would start to predict activities, and the Audio/Video sensors could be unplugged. We leave the exploration of this totally unconstrained in-the-wild deployment to future work.

6.3 Exploring privacy sensitive modalities
Our evaluation (see Figure 15) shows that Doppler, Thermal, and Lidar sensors can capture a wide range of activities with a high accuracy. However, these sensors need a direct field of view for detecting activities,
and installing several of these sensors across all home environments might incur a heavy cost to users. One approach to reducing cost is to opportunistically install sensors based on the type of activities detected in a given environment, i.e., for a kitchen setting, installing a Thermal and a Doppler sensor might be sufficient as most activities have a distinctive thermal signature (FridgeOpen, Baking, etc.), or movement patterns (CookingOnStove, Chopping+Grating, etc.). Similarly, installing a Doppler sensor might be sufficient for a bathroom setting as almost all activities have different movement patterns (HairDryer, ShaverInUse, and WashingHands), limited location variability due to short spaces, and low variability in the thermal signature. Another approach is to explore new privacy-sensitive modalities beyond what we used in our paper. For example, IMUs in mobile and wearable devices (smart watches, smart garments, etc.) can capture rich signal variability for various activities involving body movement patterns. Some recent works also used down-sampled audio signals (16KHz to 1KHz) to suppress privacy-sensitive information (i.e., speech) and show that it can still reliably detect a set of audio-based activities [71], whereas other recent works have built custom hardware platforms (i.e., person detection sensor [100]) which uses a camera under the hood but hides the complexity of the ML implementation inside the hardware module, and only exposing privacy-sensitive information. Incorporating these privacy-sensitive modalities with VAX could provide more opportunities for reducing costs and detecting a richer set of activities.

6.4 Applications of VAX beyond activity recognition

The ideas presented in the paper to utilize A/V models to bootstrap privacy-sensitive sensing modalities can be extended beyond human activity recognition. One interesting application area could be robotics, where A/V sensors can train other sensing modalities to provide additional support in noisy conditions when A/V data is unreliable, i.e., navigating self-driving cars safely through bad weather conditions. Another interesting area could be attributing activities to individuals in a multi-person setting (i.e., more than one person using the kitchen) by correlating information from different users’ personal devices (i.e., mobile phones and smart devices) with variability in data captured by VAX sensors.

7 CONCLUSION

In this paper, we present VAX, and end to end system that utilizes a set of off-the-shelf audio and video ML models to provide activity labels, for in-situ training of various privacy-sensitive sensors. VAX’s A/V pipeline combines the output of these models and provides a non-linear mapping to a consistent set of activities. VAX also proposes a method to train a variety of privacy-sensitive sensors with noisy labels from the A/V pipeline. We deployed and evaluated VAX across participants in 10 homes, performing 17 different activities. Our evaluations show that our A/V pipeline can detect 15 out of 17 activities with no human supervision. Further, we show that with just one user input provided for the undetected activities (on average 2 out of 17) our privacy-preserving sensors can detect all 17 activities with a 85% accuracy. Ultimately, VAX’s hybrid approach provides a compelling starting point for bootstrapping even more accurate HAR models in the future.

8 ACKNOWLEDGMENTS

This work was partially supported by NSF Award (SaTC-1801472, CSR-1526237) and the CMU’s CyLab Security and Privacy Institute. We gratefully acknowledge a gift by JP Morgan Chase to support research on smart buildings at Carnegie Mellon University. We would like to thank Sudershan Bhoovaraghavan, Han Zhang, and Thomas Pope for initial prototype development and feedback provided through multiple revisions of the VAX hardware. We would like to acknowledge Sudershan Bhoovaraghavan, Han Zhang, Haozhe Zhou, Vimal Mollyn, and Aigerim Massabayeva for their feedback on the paper and system design throughout. We also thank our anonymous reviewers for their constructive feedback on our paper.
REFERENCES


Heju Li, Xin He, Xukai Chen, Yinyin Fang, and Qun Fang. 2019. Wi-motion: A robust human activity recognition using WiFi signals.


117:24  •  Patidar et al.


