frontiers

Dataset for Cyber-Physical Anomaly Detection in Smart Homes

Yasar Majib [1,*], **Mohammed Alosaimi** [1], **Andre Asaturyan** [2], **and Charith Perera** [1]

[1] *School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom*

[2] *The Building Research Establishment, BRE Group, Bucknalls Lane, Watford, Hertfordshire, United Kingdom*

Correspondence*:
Corresponding Author
majiby@cardiff.ac.uk

ABSTRACT

Smart homes contain programmable electronic devices (mostly IoT) that enable home automation. People who live in smart homes benefit from interconnected devices by controlling them either remotely or manually/autonomously. However, high interconnectivity comes with an increased attack surface, making the smart home an attractive target for adversaries. NCC Group and the Global Cyber Alliance recorded over 12,000 attacks to log into smart home devices maliciously. Recent statistics show that over 200 million smart homes can be subjected to these attacks. Conventional security systems are either focused on network traffic (e.g., firewalls) or physical environment (e.g., CCTV or basic motion sensors), but not both. A key challenge in developing cyber-physical security systems is the lack of datasets and test beds. For cyber-physical datasets to be meaningful, they need to be collected in real smart home environments. Due to the inherited difficulties and challenges (e.g. effort, costs, test-bed availability), such cyber-physical smart home datasets are quite rare. This paper aims to fill this gap by contributing a dataset we collected in a real smart home with annotated labels. This paper explains the process we followed to collect the data and how we organised them to facilitate wider use within research communities. Dataset can be accessed from (Link 1) and (Link 2).

Keywords: Dataset, Smart Building, Anomaly Detection, Internet of Things, Cyber-Physical, Human Activities

1 INTRODUCTION

The smart home concept is becoming increasingly popular, day by day, due to the abilities it offers e.g., improving quality of life of the occupants by increasing convenience, comfort, privacy, and security. The emergence of the Internet of Things (IoT) has made it possible to join smart devices, connected to physical world with the Internet. This allows remote monitoring, automation and control by either user, other devices, or other systems over the Internet. As a result of this massive impact of IoT devices on our lives, smart homes became dynamic and more complex, with various devices and sensors generating massive amounts of data every second. This data has various applications such as understanding users' behaviour, improving efficiency (e.g. energy), detecting anomalies and many more. In this report, we present a dataset that contains data from a variety of sources such as cyber (network traffic), smart devices, and environmental sensors. Along with individual RAW datasets, we have also provided a dataset based on a singular timeline by merging the separate datasets into one, making it easier to analyse user behaviour and detect anomalies

based on behaviour. Furthermore, we have inflated smart devices and environmental sensors' datasets from a lower frequency e.g. max one second to a higher frequency according to the cyber dataset (because of its very high frequency). The total time period of the dataset is four weeks out of which three and half weeks were acted by the main actor (actor for normal activities) and three days by the second actor (actor for anomalous activities). The main motivation for this study is to create a dataset that can be used by researchers to train and evaluate machine learning (ML) models for smart homes, behaviour analysis of users and to detect anomalies in the behaviour of users. By combining data from multiple sources, a complete picture of user behaviour can be visualised which can also help in detecting anomalous activities that may not be possible from a single data source. The resulting dataset can be utilised to improve the performance of existing ML models or to develop new ML models to capture the complexity of smart home environments. To capture the data, we used a combination of tools and techniques.
We captured network traffic using TShark (2023) and HomeAssistant (2023) (HA) for activities of smart devices. We created a custom tool to capture environmental data from sensors including temperature, humidity, motion, illumination, air quality, proximity, pressure, and noise. The devices were installed on two different floors of the house, connected via WiFi, Ethernet or ZigBee (ZB). After capturing the data, we processed, cleaned, normalised and repaired it to remove any capturing errors or static/irrelevant information. We then merged the separate datasets into a single timeline, creating a unified dataset from multiple sources. The final form of the dataset is a CSV file with 564 features from a large number of heterogeneous devices. In conclusion, we present a comprehensive, processed, cleaned, normalised, and ready-to-use dataset consisting of cyber-physical sources. It can be used to train machine learning models for smart home applications e.g. activity detection, users' behaviour recognition, and context-aware anomaly detection.
This dataset will help researchers to find answers to research questions i) if both types, cyber and physical, on singular timescale can enhance results activity recognition or anomaly detection in smart homes, ii) what are the different techniques to fuse data from multiple sources in smart homes, and iii) comparison between dataset for both activity recognition and anomaly detection in smart homes. One of the use cases of the dataset can be detecting behaviour of a new actor in a smart home, the new actor can be an intruder with some level of high-level understanding of the legitimate resident of the smart home. The dataset contains activities of two different researchers/actors, the first actor is consider as primary resident of the smart home and the second actor (introduced for a short-time) can be consider as intruder/anomalous actor. This dataset, thus, can be used to detect behavioural anomalies (drifts) of users in a smart home environment.

1.1 Motivation

The rise of smart devices and IoTs for domestic use has resulted in exponential growth in the data being generated by smart homes. The data provides valuable insights into user's behaviour and tendencies and thus can be used to develop context-aware applications for smart home. However, analysing smart home data is overwhelming because of its diversity and heterogeneity of smart devices, environmental sensors, and network traffic on top of the complexity of users' behaviour. Furthermore, smart home dataset is typically siloed into separate data streams of information on different time scales and frequencies, such as environmental sensors or data from smart devices or network traffic packets, which makes it difficult to gain a holistic understanding of users' behaviour. For example, Figure 1(v) reflects a generic idea of cyber-physical data from human activity in a smart home, text in blue represents continuous activity whereas all other activities are binary. In order to better understand the activity from data, it is vital to have a holistic view of cyber and physical data from the activity. It is also important to have multiple data sources to develop an anomaly detection algorithm with better accuracy with lesser false positives. To address

these challenges, we present a novel smart home dataset that combines network traffic, smart device, and environmental sensors data into a single timeline, enabling an ample analysis of users' behaviour to detect anomalies in it.
The main motivation for this dataset is to enable researchers to develop novel ML models or enhance/optimise existing ML models that can predict and detect anomalies in users' behaviour. The existing, publicly available datasets, are either contains data from physical sensing or network traffic whereas our dataset contains both cyber and physical data streams. Our aim, by publishing this dataset, is to facilitate research into smart home applications that are tailored and context-aware.We believe that this dataset will be valuable to researchers in the fields of smart homes, IoT, and machine learning, as well as to developers of applications. We hope that this dataset will enable the development of useful and innovative applications that promote energy efficiency, enhance the user's experience (UX), and most importantly enhance privacy and security of users in future smart homes.

1.2 Contribution of the Work

Our aim is to explain the rationale behind the need to have a cyber-physical smart home dataset and also to explain in detail how we collected the dataset, which will help researchers better use the dataset for their own analysis. This work presents a new dataset that combines multiple (cyber-physical) sources, captured in a smart home environment over a period of four weeks. The dataset includes data from network traffic, smart devices such as smart TV, voice assistants, smart power plugs, smartphone, smart light bulbs, motion sensors, and security cameras, as well as environmental sensors such as temperature, humidity, illumination, proximity, air quality, and noise sensors. The dataset contains activities of two different researchers who acted independently, over different time frames, knowing least about each other's activity patterns, this can help in detecting anomalous behaviour in Activities of Daily Living (ADLs). The dataset was converted into a single timeline, keeping timestamp of network packets due to the rapid frequency offered. By doing that we inflated the dataset according to the frequency of the cyber dataset, and we filled the data of each stream with latest known values. The contribution of this work is twofold; i) it provides an ample view of the smart home environment by capturing the interactions between cyber, smart devices, and environmental sensors data; ii) it includes a timeline that merges all data sources as a single source, making it easier to analyse and develop ML models.

2 RELATED WORK

In this section, we will introduce a number of smart home datasets that are already publicly available to researchers in smart home domain. Some of these datasets provide both network traffic and sensors' reading data, similar to our work. These datasets are useful for researchers to perform experiments, evaluate and test smart home applications along with privacy and security. There were two datasets published by MIT Media Lab's "House_n" repository used by various researchers for smart home domain. Tapia et al. (2004) includes single occupant data from environmental sensors, energy consumption and wearable sensors for a period of three months. Another dataset by Intille et al. (2005), also known as MIT PlaceLab, provides a multi-occupant dataset for three weeks that includes audio-visual recordings on top of environmental sensors, energy consumption, and wearable sensors. Georgia Tech Aware Home (GTAH) Research Initiative offered an on-request dataset that provides features from sensors and smart devices with four participants, This dataset Kientz et al. (2008), includes motion sensors, door sensors as well as smart devices like thermostats. DOMUS is a publicly available smart home dataset that contains data from temperature,

humidity, light, and motion sensors provided by Gallissot et al. (2011). It was collected from a real smart home for a period of almost a year. The sensors were installed in different areas of the house to monitor activities of its occupants as well as environmental conditions. It has a high sampling rate, which provides fine-grained temporal information about environmental conditions while enabling researchers to analyse the data from different time scales and identify patterns and trends might not be apparent at coarser time scales. DOMUS has been adopted in many studies to develop ML models for activity recognition, occupancy detection, and energy consumption prediction in smart homes. Another publicly available dataset published by UCI Smart Home (UCIH) is similar sensors and devices for ADLs and behaviour in smart homes with a single participant by Anguita et al. (2012). Center for Advanced Studies in Adaptive Systems (CASAS) dataset provides data from various sensors and devices for ADLs and user's behaviour as well as energy consumption in multiple smart homes and participants Cook et al. (2013). The Ambient Assisted Living (AAL) Research and Application System (ARAS) dataset is also published in the same year by Alerndar et al. (2013), it includes multi-user data from motion, contact and environmental sensors including information about user's activities for a couple of months. A non-intrusive load management (NILM) dataset named UK-DALE (Domestic Appliance-Level Electricity) was shared by Kelly et al. (2015) for six homes in the UK. This dataset contains both aggregated and appliance-level power consumption. UK-Dale contains 5000 hours of power readings from lighting, kitchen, audiovisual, laundry and heating category appliances with a frequency of one second. A cyber-only dataset was collected by Miettinen et al. (2017) which provides network traffic and smart devices data for public use, IoT-Sentinel focused on cyber threats and attacks detection by fingerprinting IoT devices. Lastly, Luca Arrotta et al. (2022) recently published a sixteen hours multi-occupant dataset collected using environmental sensors, and wearable and smart devices from a single house. Table 1 provides a holistic comparison between other publicly accessible smart-home datasets with our dataset on the basis of Activities of Daily Living, User Behaviour, Energy Consumption, Network Traffic, Environmental Sensors, and Smart Devices. All datasets are publicly available except GTAH is available on request. In contrast with all previous other datasets, our dataset provides cyber (network traffic), physical (environmental sensors), smart devices, energy consumption of individual devices, user behaviour as well as activities of daily living on a synchronised single shared timeline.

3 DEPLOYMENT SCENARIO AND SETUP

3.1 House Description

The data collection setup was based in a house named Zero-Bills, built on the Building Research Establishment (BRE) Innovation Park in Watford. This house was designed for the future concept of a zero-carbon emission plan. It is based of timber and steel construction, the house is highly insulated, with a solar loft and an overheating mitigating mechanism. The house has three floors, we used the ground and the first floor (for a few activities) for this dataset collection project. The second floor has some sensors installed by BRE, data from these sensors is available in the dataset. Dimensions of the house are `9400mm x 8200mm` (depth x width) approximately. Visuals of both the front side (in red boundary) and rear side of the house are shown in Figure 1(i) and Figure 1(ii) respectively.

Table 1. Publicly Available Datasets' Comparison

Dataset	Year	Paper	H	R	Sensors and Devices	Type of Data	ADLs	UB	EC	NT	Env	SD
MIT	2004	Tapia et al. (2004)	2	S	Environmental and contact sensors	ADL, user behaviour	✓	✓			✓	
PlaceLab	2005	Intille et al. (2005)	1	S	Environment, contact sensors and energy monitoring sensors	ADL, user behaviour, energy consumption	✓	✓	✓		✓	
GTAH	2008	Kientz et al. (2008)	1	M	various sensors and smart devices	ADLs, user behaviour, energy consumption	✓	✓	✓		✓	✓
DOMUS	2011	Gallissot et al. (2011)	1	M	Environmental, air quality, and motion sensors, smart devices, and power plugs	ADL, user behaviour, energy consumption	✓	✓	✓		✓	✓
UCISH	2012	Anguita et al. (2012)	1	S	Accelerometer and gyroscope	ADL	✓					
CASAS	2013	Cook et al. (2013)	32	B	Motion, door, contact sensors	ADL, user behaviour, energy consumption	✓	✓	✓		✓	
ARAS	2013	Alerndar et al. (2013)	2	M	Environmental, contact, motion and other sensors	ADL, user behaviour	✓	✓			✓	
UK-Dale	2015	Kelly et al. (2015)	5	M	Power plugs	Energy consumption			✓			
IoT-Sentinel	2017	Miettinen et al. (2017)	NA	NA	Network traffic, communication between devices	IoT device identification, cyber threats and attacks				✓		✓
MARBLE	2022	Luca Arrotta et al. (2022)	1	M	Environmental sensors and wearable devices	ADL, user behaviour	✓	✓			✓	✓
CUBRE	2023	This paper	1	S	Environmental sensors, smart devices, network traffic, energy consumption	ADL, user's behaviour, network traffic, energy consumption, environmental conditions	✓	✓	✓	✓	✓	✓

Index: H = Houses (NA = Not Applicable), R = Residents (S=Single, M=Multi, and B=Both), ADLs = Activities of Daily Living, UB = User's Behaviour, EC = Energy Consumption, NT = Network Traffic, Env = Environmental Sensors, and SD = Smart Devices.

3.2 Floor Plans

3.2.1 Ground Floor

The entrance door is almost middle of the house facing stairs to first floor, the left section of the house on this floor is being used for storage/garage. Most of the activities take place on this floor, it has a kitchen, dining area, and lobby as usable areas as shown in Figure 1(iv). The garage area is accessible from outside using a different door, it was being used as a storage at the time of data collection.

3.2.2 First Floor

Unlike ground floor, area of first floor of the house is fully utilised. This floor consists of a bedroom, toilet, study room, and kids room as shown in Figure 1(iii). There were fewer activities taking place on this floor when compared to ground floor.

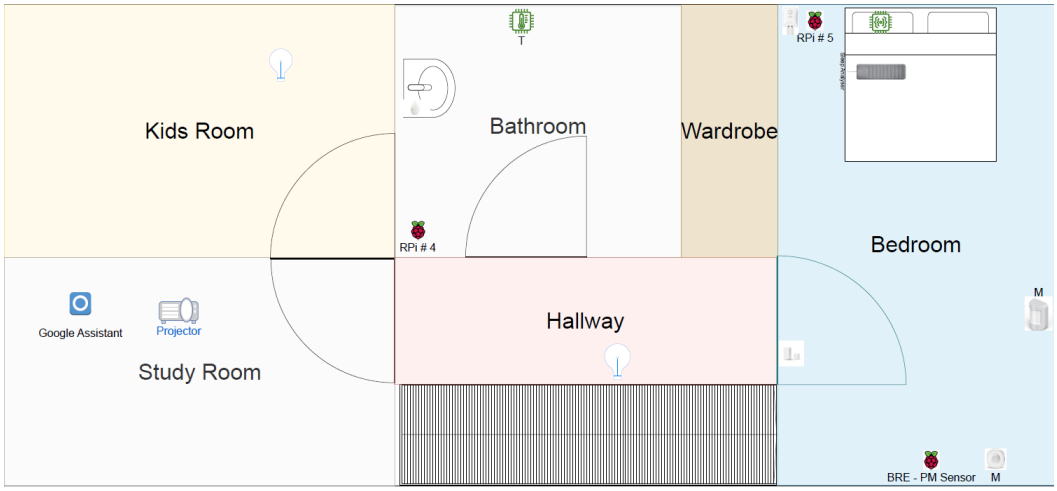
3.2.3 Second Floor

Similar to ground floor, the second floor is not fully used, there is a solar loft at the rear side of this floor. However, there are some sensors installed on this floor which are reflected in datasets. There was no activity that took place on the second floor.

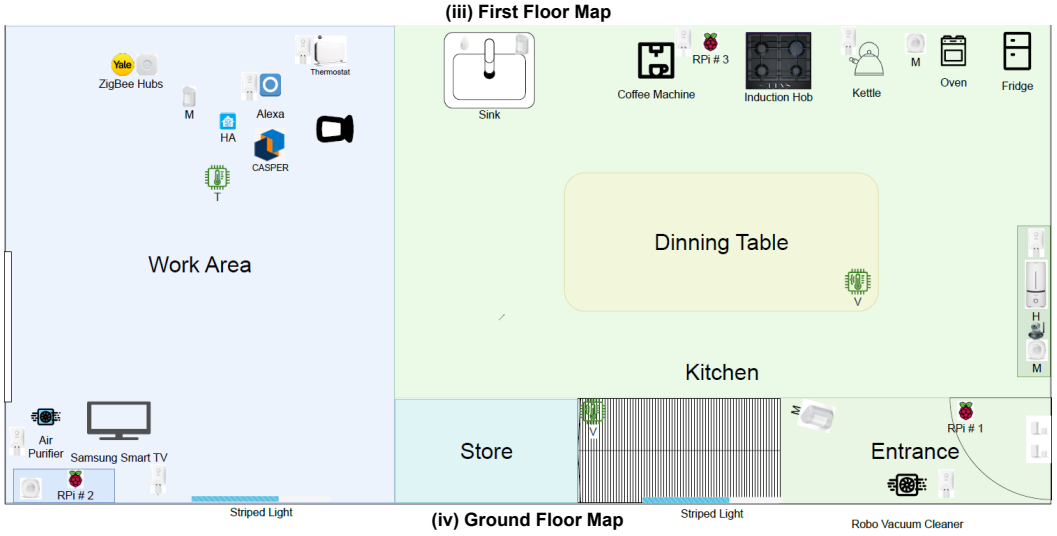
3.3 Physical Architecture

In this section, we discuss the physical elements e.g. components, connectivity, and arrangement. We have a heterogeneous physical setup that generates data through various states and in formats. A holistic view of our physical architecture is presented in Figure 2(i). We discuss our physical architecture in a

(i) Front view of the House (ii) Rear view of the House



(iii) First Floor Map



(iv) Ground Floor Map

Human Activity	Enters Home	Passes from Kitchen	Enters Living room	Sits on Sofa	Turns on Smart TV	Check messages on phone
Physical Activity	Contact sensor detects door event	Motion sensor detects occupancy	Noise sensor detects change in environment	Smart watch records sitting movement	Illumination sensor detects change in luminous intensity	
Cyber Activity	Sensor data to Smart Hub	Phone connects to wifi network	Noise levels sent to server	Sends data to network (via mobile or direct)	Light sensor sends status update to Smart Hub and TV sends/receives network packets	Mobile sends/receives network packets

(v) A generic view of cyber-physical data from human activity in smart home

Figure 1. Zero-Bills House in BRE and a generic example of cyber-physical activity

structured format to make sure it is understood clearly as it is the key component of the data collection setup. The physical architecture consists of four components:

3.3.1 Core Devices

These are mission-critical devices that are mandatory for operations in data collection process, failure of any of these devices could result in major faults in the whole operation. There are three core devices in the architecture:

- **Gateway (GW)** as discussed earlier that the GW is the most important device in this physical architecture, it is responsible for various tasks e.g., data storage, data ingestion, network packet capture, and services e.g., Dynamic Host Configuration Protocol (DHCP), Hyper-Text Transfer Protocol (HTTP) and Internet Sharing using Network Address Translation (NAT) Protocol, an RPi based industrial IoT gateway, OnLogic FR201 hardware was used for this purpose.
- **HomeAssistant (HA)** is based on the HomeAssistant Operating System (HAOS) installed on an 8GB RAM variant of RPi4 device with a 256GB Solid-State Drive (SSD) storage. One other option was to use a docker container on GW to install HA, this could have added extra load on GW hardware which may have resulted in low performance.
- **Wireless Access Point (WAP)**: Our initial plan was to assign this task to GW but there was a limitation on how many devices can be controlled using a built-in wireless card on RPi. Since the number of devices was too many so we introduced TP-Link TL-WR940N in the network to be used for creating wireless network as a bridge to GW (no DHCP was configured on this device).

3.3.2 Intermediary Devices (M)

These are the devices in the middle, connected physically or over the network with end-point devices, we had two types of intermediary devices RaspberryPi (RPi) v4 was used to extract data from environmental sensors and transfer to GW and Smart Hubs i.e. Philips Smart Hub, connected to HA via Ethernet, creating separate ZB networks for some EPDs.

3.3.3 End-Point Devices (EPDs)

These are the sensors or actuators which are the actual source of data. There are three types of EPDs in the dataset:

- Smart devices directly connected to HA over ZB
- Smart sensors connected to HA through smart hubs (Internal or External WiFi)
- Environmental Sensors (EnvS) attached to RPi over Internal WiFi

3.3.4 Physical Networks

There were six different Layer-1 physical networks in the architecture:

i. Ethernet Core: this network physically connects GW, HA, WAP, and Smart Hubs using Ethernet cables. The network ID of this network is `10.11.12.0` and the broadcast ID is `10.11.12.255`.

ii. Ethernet Management: This interface was used to transfer captured packets to the configuration laptop, this network was created to ignore the transfer over the core network which could have doubled the data storage for packet captures.

iii. Internal WiFi: The internal WiFi was created using WAP with Wi-Fi Protected Access Pre-Shared Key version 2 (WPA2-PSK), the SSID of the network was `shtb`.

iv. External WiFi: Any external WiFi network can be considered for this physical network e.g., a hot-spot sharing network using a mobile phone or an open-to-public network `BRE-Visitors`.

v. GSM Network: An data SIM was used as a backup in case of unavailability of external WiFi networks. A D-Link DWM-222, LTE (4G) USB Modem, used for GSM connection. This switch-over was manually performed.

vi. ZB LAN: A USB-based ZB dongle connected to HA was used for this network.

The most busy device in the whole architecture is the Gateway (GW) which performs a number of critical operations e.g., ingestion and transformation, forwarding or storage of data in final format.

3.4 Logical Architecture and Data Pipeline

In this section, we discuss the logical architecture of our data collection setup. As seen in Figure 3, we have an architecture based on six sections, separated horizontally. The most critical section in the logical architecture is ”Core Interfaces” which interacts with almost all parts of data, the two core components HA and GW hosts most of the core interfaces with the exception of one data receiving point i.e. a weekly comma-separated values (CSV) file from BRE which holds data from BRE owned sensors and smart devices.

3.4.1 Core Interfaces

The HA works on both IP and ZB protocols to gather data from smart devices (connected directly using ZB, Internal WiFi, or Internet) and perform automation steps e.g. controlling temperature using a smart radiator after sensing using temperature sensors. It then stores the data in MariaDB database as states for each entry. The MariaDB server database can be accessed using the `phpMyAdmin` package using a web browser on GW to download a CSV file containing the log of all entities connected to HA. The GW hosts an `HTTP server` that receives data from EnvS via RPi, it adds some information as reflected in ”a T process in a dual circle” in the processes section and stores the sensor log into a CSV file. GW also enables, in addition to NAT service, DHCP, and Domain Name Service (DNS) services to allow local (IoT) devices to get dynamic IP addresses and perform DNS queries. All traffic passing through the NAT service was captured using `tshark` package reflected by ”a C process in a dual circle” in the processes section and stored in pcap format, a new pcap file is created every hour to prevent load on RAM of the GW hardware. The routing interface also allows some IoT devices connected via Internal WiFi to access internet-based cloud services. Email client runs on the configuration laptop, and saves the CSV formatted BRE data dump, received weekly, to local storage.

3.4.2 Cloud Services

There were a few cloud services utilised, comparatively less frequently, as most of the data was being stored in local storage. Cloud Services were used to access cloud applications by some IoT devices, audio/video streaming services (Smart TV and Alexa), or email services to receive external data via email.

3.4.3 Local Data Points

The main data generating (source) are the IoT devices connected either through ZB via a ZB wireless network or IP, both IoT and EnvS, connected through Internal WiFi. IoT devices send data to HA directly or via Cloud Services whereas EnvS are directly ingesting data to GW’s HTTP service. BRE Network section was only known through email of sensor and device logs in CSV format. The local Storage section has three different types of data formats i) CSV files, ii) pcap files, and iii) MariaDB database (eventually converted into CSV). Internal Processes as discussed earlier are T for transforming the data coming from EnvS and C for capturing the data flowing through routing services running on GW.

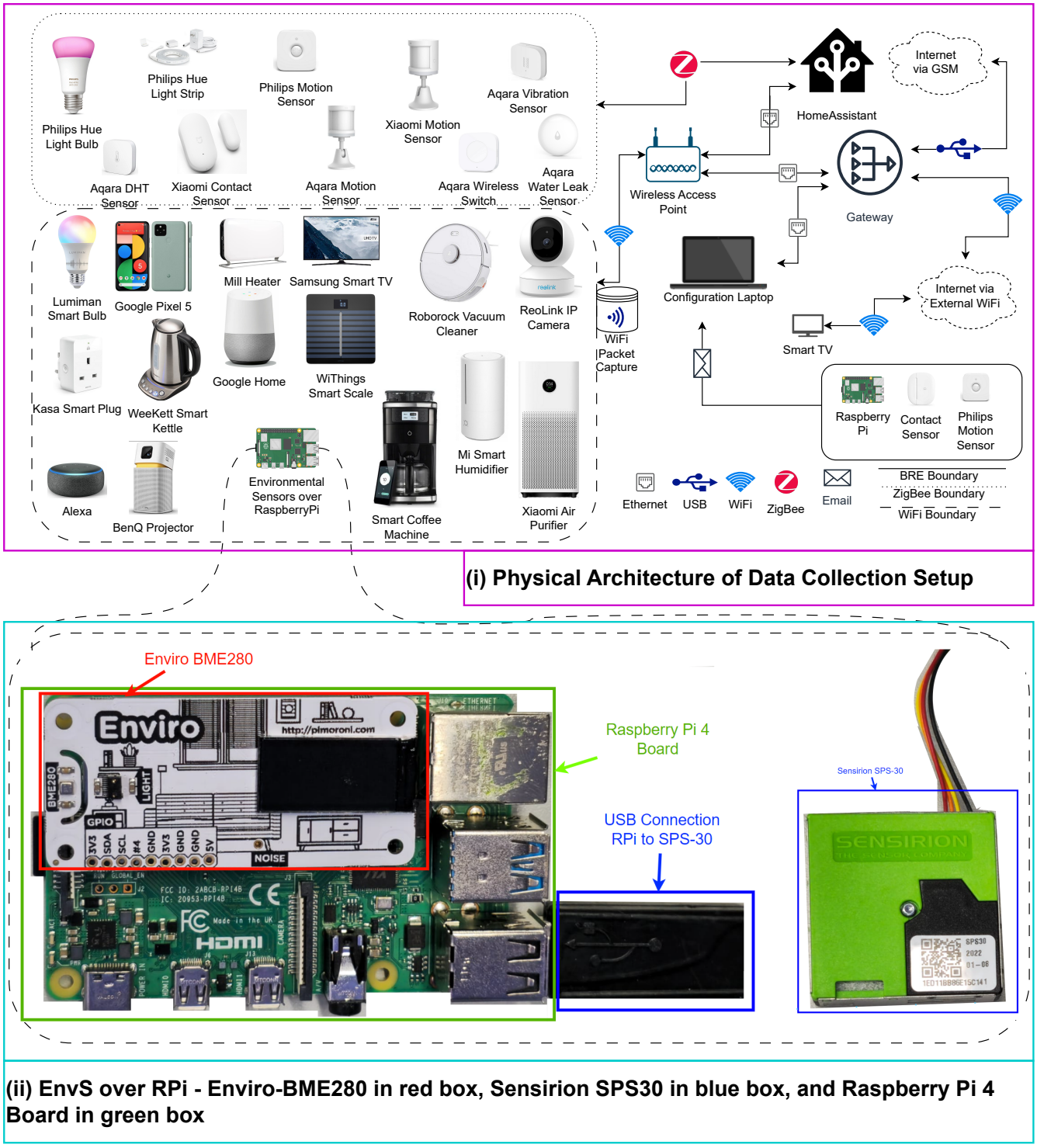


Figure 2. Physical Architecture and EnvS over RPi for Data Collection Setup

3.4.4 Environment Sensors (EnvS) Node Specification

We developed the sensor nodes using three hardware components i) Enviro BME280 Environmental Sensor, Sensirion SPS30, and iii) RPi4. Figure 2(ii) displays all three components integrated together. BME280 contains Temperature, Humidity, Noise, Pressure, Illumination, and Proximity; it was installed as a HAT on top of RPi. SPS30 is an AirQuality (AQ) sensor connected to RPi via a USB port. It provides

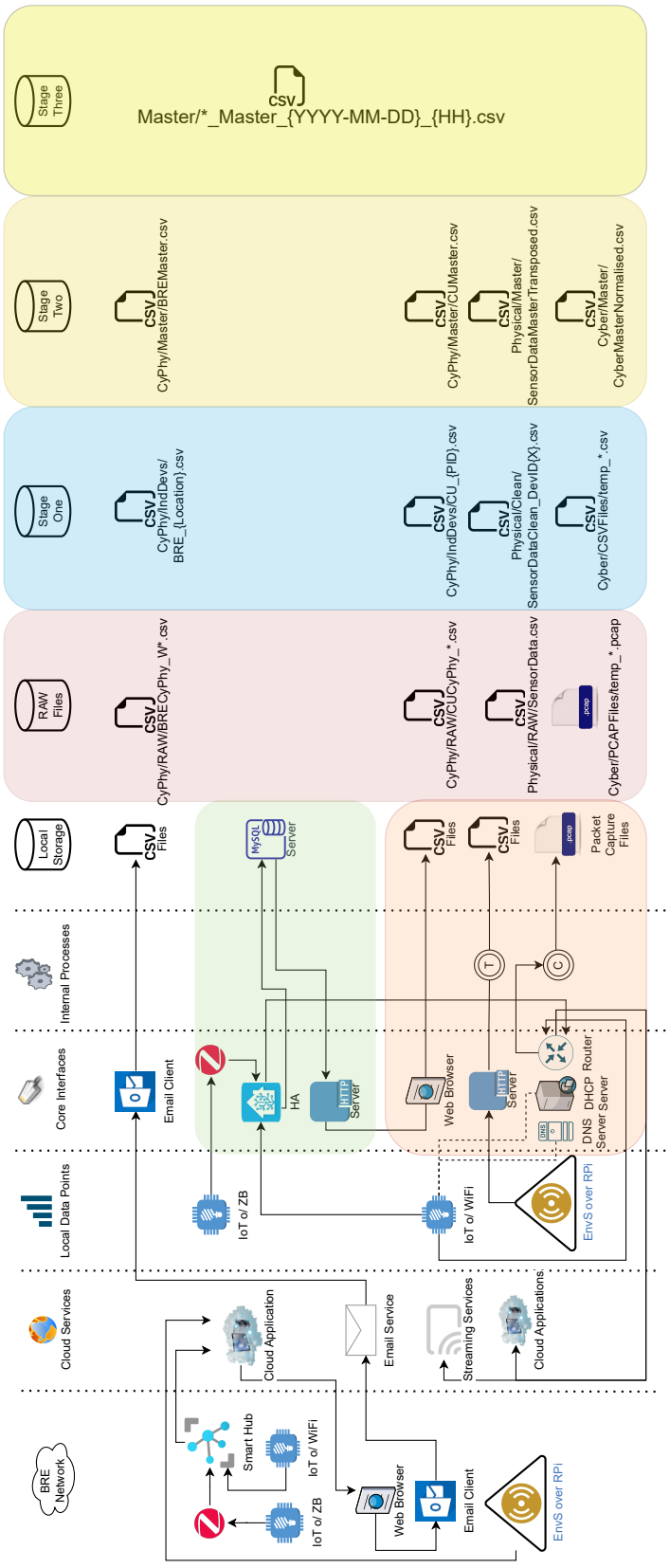


Figure 3. Logical Architecture of Data Collection Setup: Orange background area represents Gateway, and green background area shows HA.

various types of AQ measurements. The frequency of EnvS data was one second, details of the data specifications are discussed in Section 4.1.1.

3.5 Smart Devices' Specifications

In this section, we discuss the smart devices we used in the data collection setup. Table 2 in lists the devices' communication, location, data, and frequency information followed by individual device's specifications. Product Identification Number (PID) can be used to extract data for a particular device from the dataset, only one item i.e. Samsung Smart TV has a PID or -1. Details on individual device specification and entities can be found in Supplementary Content 7. Section 1 contains specifications of each device in its sub sections and table of entities of each device can be found in Section 3. Sub section numbers in Supplementary Content Document (SCD) 7 are synchronized with this document e.g. Section 3.5.9 contains specifications brief details on Kasa Smart Plug and Section Section 2.9 provides specifications of the same smart device and same case for Xiaomi Motion Detector details are in Section 3.5.2 and Section 2.2.

- binary_sensor.smartthings_contact_sensor_pid_{X}_moving
- binary_sensor.smartthings_contact_sensor_pid_{X}_tamper
- binary_sensor.smartthings_contact_sensor_pid_{X}_battery_low
- binary_sensor.smartthings_contact_sensor_pid_{X}_contact
- sensor.smartthings_contact_sensor_pid_{X}_z_axis
- sensor.smartthings_contact_sensor_pid_{X}_y_axis
- sensor.smartthings_contact_sensor_pid_{X}_x_axis
- sensor.smartthings_contact_sensor_pid_{X}_battery
- sensor.smartthings_contact_sensor_pid_{X}_temperature
- binary_sensor.contact_sensor_pid_{X}_moving
- binary_sensor.contact_sensor_pid_{X}_tamper
- binary_sensor.contact_sensor_pid_{X}_battery_low
- binary_sensor.contact_sensor_pid_{X}_contact
- sensor.contact_sensor_pid_{X}_z_axis
- sensor.contact_sensor_pid_{X}_y_axis
- sensor.contact_sensor_pid_{X}_x_axis
- sensor.contact_sensor_pid_{X}_battery
- sensor.contact_sensor_pid_{X}_temperature

In-depth details of individual devices are presented in sub-sections. Each individual device has a unique internal PID, in case of multiple items for the same device, the {X} can be replaced with a 3-digit (zero-filled) PID from the Table 2 to get the entity_id for a particular device, for example, there were three Samsung SmartThings Contact Sensors used in the dataset, each of them has different PID i.e. 3, 4, and 6. The resulting entity ID for `binary_sensor.smartthings_contact_sensor_pid_{X}_moving` from PID 3 will become `binary_sensor.smartthings_contact_sensor_pid_003_moving`.

3.5.1 Samsung SmartThings Contact Sensor

Samsung SmartThings Contact Sensor, also known as Samsung SmartThings Multipurpose sensor is a battery-powered device, connected via ZB. It provides open/close event data as well as environmental readings to SmartThings hub as well as integrate-able with third-party ecosystems. We installed three of these sensors on Oven, Fridge, and the main Entrance doors. Table 2 in SCD 7 shows the entity_ids and format of the data provided by Samsung SmartThings Contact Sensor and Section 2.1 provides detailed specifications of this sensor. An image of Samsung SmartThings Contact Sensor is provided with Supplementary Content 7 as Figure 1.

3.5.2 Xiaomi Motion Detector

The Xiaomi Motion Detector detects body heat and movement in a close area up to a 7-meter range using an infrared sensor and communicates with HA via USB ZB Dongle. It is a battery-powered sensor and can

Table 2. Smart Devices used in Data Collection

PID	Specs	Device Name	Location	Layer 1 Network	First Connection
4	3.5.1	Samsung Smart Things Contact Sensor	Oven	ZB	HA - ZB
3	3.5.1	Samsung Smart Things Contact Sensor	Fridge	ZB	HA - ZB
6	3.5.1	Samsung Smart Things Contact Sensor	Entrance Door	ZB	HA - ZB
138	3.5.2	Xiaomi Motion Sensor	Kitchen	ZB	HA - ZB
139	3.5.2	Xiaomi Motion Sensor	Workarea	ZB	HA - ZB
136	3.5.3	Xiaomi Contact Sensor	Bedroom	ZB	HA - ZB
81	3.5.4	Aqara Wireless Switch	NA	ZB	HA - ZB
82	3.5.4	Aqara Wireless Switch	NA	ZB	HA - ZB
77	3.5.6	Aqara Vibration Sensor	Kitchen	ZB	HA - ZB
72	3.5.5	Aqara Temperature and Humidity	Bedroom	ZB	HA - ZB
74	3.5.5	Aqara Temperature and Humidity	Workarea	ZB	HA - ZB
71	3.5.5	Aqara Temperature and Humidity	Kitchen	ZB	HA - ZB
73	3.5.5	Aqara Temperature and Humidity	Washroom	ZB	HA - ZB
79	3.5.6	Aqara Vibration Sensor	Bedroom	ZB	HA - ZB
78	3.5.6	Aqara Vibration Sensor	Stairs	ZB	HA - ZB
75	3.5.7	Aqara Motion Sensor	Entrance	ZB	HA - ZB
40	3.5.10	Philips Hue Bulb	Kids Room	ZB	HA - ZB
145	3.5.11	Philips Hue LightStrip	Workarea	ZB	HA - ZB
62	3.5.12	Philips Hue Motion Sensor	Bedroom	ZB & IP	HA - IP - Philips Hue Hub
146	3.5.11	Philips Hue LightStrip	Stairs	ZB	HA - ZB
83	3.5.8	Aqara Water Leak Sensor	Kitchen	ZB	HA - ZB
84	3.5.8	Aqara Water Leak Sensor	Washroom	ZB	HA - ZB
114	3.5.26	Smart Coffee Machine	Kitchen	IW	Google Pixel 5
109	3.5.24	WeeKett Smart Kettle	Kitchen	IW	HA - IP
107	3.5.25	Mi Smart Antibacterial Humidifier	Kitchen	IW	HA - IP
113	3.5.22	Mill Smart Electric Radiator	Workarea	IW	HA - IP
-1	3.5.23	Samsung Smart TV	Workarea	IW OR EW	HA - IP
108	3.5.19	Xiaomi Air Purifier	Workarea	IW	HA - IP
143	3.5.21	IP Camera - Reo Link	Kitchen	IW	HA - IP
115	3.5.20	Benq Projector	Study Room	IW	HA - IP
125	3.5.13	Google Pixel	Mobile	IW OR EW	HA - IP
94	3.5.14	WiThings Smart Scale	Washroom	IW	Google Pixel 5
20	3.5.9	Kasa Smart Plug	Workarea	IW	HA - IP
21	3.5.9	Kasa Smart Plug	Workarea	IW	HA - IP
22	3.5.9	Kasa Smart Plug	Workarea	IW	HA - IP
91	3.5.17	Roborock Vacuum Cleaner	Entrance	IW	HA - IP
24	3.5.9	Kasa Smart Plug	Entrance	IW	HA - IP
53	3.5.18	Lumiman Smart Bulb	Hallway	IW	HA - IP
13	3.5.15	Amazon Echo Dot	Kitchen	IW	HA - IP
23	3.5.9	Kasa Smart Plug	Kitchen	IW	HA - IP
18	3.5.9	Kasa Smart Plug	Kitchen	IW	HA - IP
19	3.5.9	Kasa Smart Plug	Kitchen	IW	HA - IP
17	3.5.9	Kasa Smart Plug	Bedroom	IW	HA - IP
63	3.5.16	Google Speaker	Study Room	IW	HA - IP

Index: ZB = ZigBee, IW = Internal WiFi, HA = HomeAssistant, EW = External WiFi, IP = Internet Protocol.

be installed easily on any type of door or window. Can also be connected to other smart home ecosystems. Other than the primary sensor i.e. motion, it provides battery information and power outage count as well. Detailed specifications are available in Section 2.2 of SCD 7 whereas Table 7 show the details of data and frequency of each entity_ids. For both Xiaomi Motion Detectors, PIDs are not configured in the entity_id. An image of Xiaomi Motion Detector is provided with Supplementary Content 7 as Figure 2.

3.5.3 Xiaomi Contact Sensor

The Xiaomi Contact Sensor is a battery-powered, ZB-based device primarily used with the manufacturer's hub but can also be used with other third-party smart home ecosystems. Only one device of this model is installed in the bedroom on first floor. Table 8 provides a list of entities in the Smart Devices' dataset

and detailed specifications can be found in Section 2.3 of SCD 7. An image of Xiaomi Contact Sensor is provided with Supplementary Content 7 as Figure 3.

3.5.4 Aqara Wireless Switch

The Aqara Wireless Switch comes with a push button, battery-powered, which can communicate with the HA via ZB without needing any physical installation. Various operations can be triggered upon button press event using any supported smart home hub. Apart from providing button press events, it also provides device temperature. We had two of these buttons, both of these devices were added in HA but not used during the data collection process. Technical specifications are listed in Section 2.4 of SCD 7 and all data entities provided by this switch are presented in Table 9. An image of Aqara Wireless Switch is provided with Supplementary Content 7 as Figure 4.

3.5.5 Aqara Temperature and Humidity Sensor

The Aqara Temperature and Humidity Sensor (AqaraDHT) is a compact, battery-powered device with a high-precision sensor for temperature and humidity measurements, which can communicate with any support hub via ZB. We have installed four of these sensors in the following locations; GF: Kitchen, and Workarea, FF: Bedroom, and Washroom. Details of data provided by this sensor are available in Table 1 with technical specifications in Section 2.5 of SCD 7. An image of Aqara Temperature and Humidity Sensor is provided with Supplementary Content 7 as Figure 5.

3.5.6 Aqara Vibration Sensor

Aqara Vibration sensor is also based on ZB protocol with a built-in battery, it is compatible with Aqara hub as well as other ecosystems. We installed three vibration sensors of this model in Kitchen, Stairs, and Bedroom. It primarily sends vibration events but also provides device temperature and 3-axis, the detailed specification can be found in Section 2.6 with a complete list of entities from Table 7 of SCD 7. An image of Aqara Vibration Sensor is provided with Supplementary Content 7 as Figure 6.

3.5.7 Aqara Motion Sensor

Aqara Motion Sensor, similar to other Aqara devices, is also battery-operated and can be connected to Aqara hub or other smart home ecosystems via ZB. There was only one motion sensor of this make was installed at main entrance door. Other than occupancy events, it also provides illuminance and device temperature information. A detailed specification is provided in Section 2.7 and all entity_ids are listed in Table 10 of SCD 7. An image of Aqara Motion Sensor is provided with Supplementary Content 7 as Figure 7.

3.5.8 Aqara Water Leak Sensor

Similar to the same manufacturer items, Aqara Water Leak Sensor provides connectivity over ZB, and has a built-in battery. In addition to water leakage information, it provides device temperature information as well along with other entities. We installed two of these sensors in Kitchen and Washroom. Detailed specifications are shown in Section 2.8 and all entities are listed in Table 13 in SCD 7. An image of Aqara Water Leak Sensor is provided with Supplementary Content 7 as Figure 8.

3.5.9 Kasa Mini Smart Plug by TP-Link

Kasa Mini Smart Plug is WiFi based device, which can be connected with Kasa Smart App as well as various smart home ecosystems. It provides current and voltage information as well as other key properties,

all details are available in Table 4 with detailed specifications in Section 2.9 of SCD 7. We installed eight of these smart plugs with various devices at Entrance, Kitchen, Workarea, and Bedroom. An image of Kasa Mini Smart Plug by TP-Link is provided with Supplementary Content 7 as Figure 9.

3.5.10 Philips Hue Light Bulb

Philips Hue Light Bulb is a ZB device that connects to HA through Philips Hue Hub or a ZB repeater. Its electrical connection is B22 standard. It is a multi-colour smart bulb that provides bulb status to the smart hub. Detailed specifications of this light bulb are listed in Section 2.10 and dataset entities are provided in Table 3 in SCD 7. There was only one of these bulbs installed in the kids' room on first floor. An image of Philips Hue Light Bulb is provided with Supplementary Content 7 as Figure 10.

3.5.11 Philips Hue LightStrip

Philips Hue LightStrip is a form of rope light with smart features and various colours. It is also integrateable with HA and other ecosystems over ZB via Philips Hue Smart Hub or a repeater. We installed two of these lights on the Stairs and Workarea. Table 11 lists all data entities related to this light strip with detailed specifications in Section 2.11 of SCD 7. An image of Philips Hue LightStrip is provided with Supplementary Content 7 as Figure 11.

3.5.12 Philips Hue Motion Sensor

Philips Hue Motion Sensor is also a ZB device with a built-in battery, it is easy to install and integrate sensors. In addition to motion events, it provides temperature and illumination readings as well. All entities with their details are shown in Table 12 and detailed specifications can be found in Section 2.12 of SCD 7. Only one motion sensor by Philips was installed in the Workarea. An image of Philips Hue Motion Sensor is provided with Supplementary Content 7 as Figure 12.

3.5.13 Google Pixel 5

Google Pixel 5 comes with 8 GB RAM and 128GB storage with an Octa-core 2400 MHz processor, a detailed specification can be found in Section 2.13. It provides a lot of data points like other Android phones to HA, Table 5 in SCD 7 lists all entities provided by Google Pixel 5 in the dataset. We had only one mobile phone, used to access various applications and was kept with actors most of the time. An image of Google Pixel 5 is provided with Supplementary Content 7 as Figure 13.

3.5.14 WiThings Smart Scale

WiThings Smart Scale provides body weight along with additional features like body composition. All features of these devices are mentioned in Section 2.14 of SCD 7. This device does not connect to HA but rather provides a connection with its own WiThings Health Mate mobile application (which was installed in Google Pixel 5) hence data was not available in HA. An image of WiThings Smart Scale is provided with Supplementary Content 7 as Figure 14.

3.5.15 Alexa Dot 3rd Generation

Amazon Alexa is a well-known voice assistant, connected to HA using WiFi. It provides media information and alarm-related details to HA, all entities listed in Table 14. The specification can be found in Section 15 of SCD 7. We frequently used Alexa Dot for daily activities, it was placed in the Workarea. An image of Alexa Dot 3rd Generation is provided with Supplementary Content 7 as Figure 15.

3.5.16 Google Home

Google Home or Google Speaker is also a well-known voice assistant native to the Google ecosystem. It can also be integrated with HA via WiFi. Detailed specifications of Google Home are listed in Section 2.16 of SCD 7. There was only one Google Home placed on first floor, but there is no data available in HA. An image of Google Home is provided with Supplementary Content 7 as Figure 16.

3.5.17 Roborock Vacuum Cleaner S5 Max

Roborock Vacuum Cleaner was connected to HA via WiFi, it provides status and stats to HA. There were very few activities performed with this device. The specification can be found in Section 17 and a list of entities in Table 15 of SCD 7. We had one robot cleaner, placed at entrance area near the door. An image of Roborock Vacuum Cleaner S5 Max is provided with Supplementary Content 7 as Figure 17.

3.5.18 LUMIMAN Smart Bulb

Lumiman Smart Bulb is multi-colour bulb connected via WiFi with HA, providing status only. It has E27 electrical connection, detailed specifications are provided in Section 2.18 and Table 16 of SCD 7. We have one of these smart bulbs from Lumiman which was installed in the hallway of first floor. An image of LUMIMAN Smart Bulb is provided with Supplementary Content 7 as Figure 18.

3.5.19 Xiaomi Air Purifier 4 Pro

Xiaomi Air Purifier 4 Pro was connected to HA via WiFi, it purifies air as well as provides environmental information e.g. Air-Quality, Temperature, and Humidity. The detailed specification of this Air Purifier is available in Section 2.19 and a list of dataset entities is available in Table 17 in SCD 7. It was installed in Workarea. An image of Xiaomi Air Purifier 4 Pro is provided with Supplementary Content 7 as Figure 19.

3.5.20 BenQ GV1 Projector

BenQ GV1 is a smart project with multiple input options, it has a built-in battery and USB-C connection for laptop/mobile display connectivity, detailed specifications are available in Section 2.20 of SCD 7. there are no data entities for this device. We had the smart projector installed in Study Room. An image of BenQ GV1 Projector is provided with Supplementary Content 7 as Figure 20.

3.5.21 ReoLink E1 Pro Camera

ReoLink E1 Pro was used to capture video of activities being performed by actors, it was not installed in HA. It was connected to the Internet for uploading recordings to ReoLink Cloud. Since it was not installed in HA only cyber activity is available in cyber dataset. Specifications of the ReoLink E1 Pro camera are provided in Section 2.21 of SCD 7. An image of ReoLink E1 Pro Camera is provided with Supplementary Content 7 as Figure 21.

3.5.22 Mill WiFi Portable Heater 1200W

Mill WiFi Portable Heater was connected to HA via WiFi. Apart from managing temperature in the room, it also provides environmental information along with daily and yearly electricity usage analysis. The heater was installed in Workarea. Detailed specifications of this heater are available in Section 2.22 and all entities are listed in Table 18 of SCD 7. An image of Mill WiFi Portable Heater 1200W is provided with Supplementary Content 7 as Figure 22.

3.5.23 Samsung Smart TV

Samsung Smart TV was connected to HA via Internal WiFi in the first place but later switched to external WiFi due to network issues. This TV was used to watch news most of the time to create activities. The TV is owned by BRE and was installed in Workarea. Specifications of the TV is provided in Section 2.23 and detail of entities are listed in Table 19 of SCD 7. An image of Samsung Smart TV is provided with Supplementary Content 7 as Figure 23.

3.5.24 WeeKett Smart WiFi Kettle

Even though WeeKett Smart Kettle is a WiFi device but not integrate-able with HA, we connected this kettle with Google Pixel 5 to get data, one of the smart features of this device was to keep water hot to a specific temperature. It was installed in the Kitchen and was used to make tea most of the time. A list of all important specifications is provided in Section 2.24 and entity_ids in Table 20 of SCD 7. An image of WeeKett Smart WiFi Kettle is provided with Supplementary Content 7 as Figure 24.

3.5.25 Mi Smart Antibacterial Humidifier

Mi Smart Antibacterial Humidifier keeps humidity levels as configured to make the environment pleasant. It was installed in kitchen area, effects can be seen in the Physical dataset in comparison to data provided by this Humidifier, details of entities are provided in Table 21, specifications are also provided in Section 2.25 of SCD 7. An image of Mi Smart Antibacterial Humidifier is provided with Supplementary Content 7 as Figure 25.

3.5.26 Smart Coffee 2nd Generation

Smart Coffee machine was used frequently for multiple types of activities. There were two same-model coffee machine used in the data collection project, the first one had hardware issues so replaced by a new one. Detailed specifications of the coffee machine are available in Section 2.26 of SCD 7. The coffee machine was installed in the Kitchen, there is no data available in HA, perhaps cyber dataset contains network packets to/from it. An image of Smart Coffee 2nd Generation is provided with Supplementary Content 7 as Figure 26.

4 DATA MODELLING AND DATA DESCRIPTION

The dataset consists of physical, cyber, and cyber-physical four sub-datasets, captured simultaneously, by inter-connected devices at the same location but each dataset contains different data types and formats. We captured the following type of subsets for this dataset:

- Physical (In-situ sensors): Environmental reading of temperature, humidity, light, noise, and air quality.
- Cyber (Network Router): All internal network traffic passes via NAT router (Gateway).
- Smart Devices (Home Assistant): Smart devices connected to the home assistant directly or via intermediate devices or via the internet.
- Video Recordings (while performing the activities): An IP camera was installed in the kitchen area to record video of activities being performed by actors.

4.1 Data Collection

In this section, we briefly discussed each individual dataset. Section 4 of SCD 7 contains in-depth technical details on each dataset.

4.1.1 Physical

The physical dataset was captured using two different environment-sensors kits i) Enviro BME280 (Environment sensors e.g. Temperature, Humidity), and ii) Sensirion (Air Quality), both add-on devices were physically installed on an RPi using HAT and USB interface respectively. The Enviro kit was installed from the start on all RPis, only one RPi (RPi-3) had Sensirion. Refer to section 4.1 of for technical details about Physical dataset, including list of all entities and RPis. Later after November 2, 2022, all five RPis were equipped with both kits. We developed two services that run on RPi Client (RPi-C) at startup to collect and transfer data to a data storage device via HTTP. We developed a service for a data storage device that listens to HTTP requests to receive and store the data from all RPis to a single CSV formatted file locally. Table 22 of SCD 7 show the Device ID, IP Address, MAC Address, Name, and location of each RPi-C in the house. Following is the header of the CSV file: In the above list PIR (x), AirQuality (x), Accelero (x), Gyro (x), Pressure (x), and Motion (x) values are static in the dataset. The timestamp is generated at the time of sensor readings, DevID is unique and hard-coded in each RPi-C, and DevIP is generated by data-storage service using the ”remote_addr” function in Python requests library. The frequency of HTTP requests is one second which is also similar to Enviro sensors, whereas Sensirion air-quality (SRS-30) values were read every two seconds. Air quality data was read by another service and stored the current values in a .txt file which is read by main service. In case of a network-related failure, the data is stored locally on RPi-C with a timestamp and sent to HTTP service once the service is reachable. This dataset is stored in SensorData.csv in ”Physical” directory, size of this file is 2.7 gigabytes.

4.1.2 Cyber (Network Router)

The cyber dataset captured at the router contains network traffic to/from all internal devices connecting to the internet using ”shtb” wireless network. The Gateway plays key role in this dataset because it provides all services required to access to the internet, to internal devices, as well as capture the network traffic. The services offered by Gateway are DHCP, DNS, and NAT. Refer to section 4.2 in SCD 7 for technical details about Cyber dataset including devices' hostnames, MAC and IP addresses. Gateway has two Ethernet ports eth1 and eth0, the first is connected to an internal WAP which bridges wireless network to eth1. The second port, eth0, is directly connected to the maintenance laptop, this network is not captured. Because the dataset was transferred to a laptop, over a network, at the end of every day while the eth1 network is being captured would have created another footprint of this transfer that would significantly increase the storage requirement of the cyber dataset. The idea was to capture the network traffic without adding any exception so an additional maintenance network was introduced. Gateway's Internet Protocol (IP) address was `10.11.12.1` with Network ID `10.11.12.0`, Broadcast ID was `10.11.12.255` and netmask was 255.255.255.0 of eth1 network. Tshark TShark (2023), a network packet capture tool, was used to dump the network captures to a pcap file. The pcap files were created at one-hour frequency i.e. there is a new file created every hour with start timestamp in filename. Table 23 in SCD 7 provides a list of IP addresses and mac addresses for all internal devices, the DHCP lease time was configured to be unlimited to save any conflict in device identification using IP address. The dataset consists of network packet capture format (.pcap) files stored in ”Cyber” directory, file name format is `temp_{HourID}_{YYYYMMDDhhmmss}.pcap` for example the file name of the first hour is: `temp_00001_20221019140448.pcap`.

4.1.3 Smart Devices (Home Assistant)

This dataset contains data from smart devices connected to HA, most of the smart devices were connected to HA, either directly using IP, ZB, via an intermediate device, or internet. This makes HA the most critical

component of this dataset. HA was connected to Gateway first via wireless network and later Ethernet, in both scenarios, via WAP, this is the reason for two DHCP lease entries for HA in Table 23 SCD 7 as IP `10.11.12.22` and `10.11.12.54` respectively. The devices were manually added one by one in the respective areas in HA. This dataset was exported using the phpMyAdmin add-on in HA, the output format is CSV which contains the followings headers:

4.1.4 Smart Devices (BRE)

This dataset is captured by BRE, it includes data from motion, contact, air quality (various), temperature, tamper, humidity, and illumination sensors. The average frequency of this dataset is 5 seconds, but there was a power cut for the central hub due to a human error the dataset has one week of missing data. , There are a total number of 21 devices in the dataset, and each device contains one or more sensors, Table 25 of SCD 7 enlists all locations whereas the sensor codes can be matched in Table 24. The location column has a pre-fix `UK-WAT-ZB` which reflects the house "Zero-Bills" followed by MTS ID (internal network number). The later part of the location holds, separated by "-", type code of device e.g. OC for contact sensor, floor number (starting from 00), and area information e.g. K1 for kitchen and B1, B2 for Bedroom 1 and 2 respectively.

4.1.5 Video Recordings (while performing the activities)

Video recordings of the activities were captured using ReoLink IP Camera. It starts video recording after detected motion in the field of view, and the recorded video is sent to the cloud service provided by the manufacturer. The video files were later downloaded from the cloud service and stored in "VideoRecordings" directory.

4.2 Data Cleaning

All datasets except Video Recordings were processed in some way. Both RAW and processed datasets are available in the repository. For a comprehensive look at all the datasets' locations and properties, please refer to Table 26 in SCD 7.

4.2.1 Physical Dataset

We had control over physical data format at the time of capturing and ingestion which makes it easier to handle, clean, and normalise. However, there was an error that affected values of a few sensors while capturing stage. There are seven columns that have single (for each device) static values, for stage one we removed static columns (PIR, AirQuality, Accelero, Gyro, Pressure, and Motion) as well as DevIP, and DevID from the RAW dataset. There were three sensors that had errors at capturing stage i) (Temp) temperature values were dipping, ii) (Lux) illumination and iii) (Humidity) humidity values rising abruptly. This issue started after November 02, 2022 post-installation of AirQuality Sensors (Sensirion SRS-20) in EnvS (1,2,4, and 5), EnvS 3 had this issue since the start because SRS-30 was installed since the beginning. This behaviour for all Humidity and Temperature/Lux sensors was for two iterations and one iteration respectively every time. To resolve this issue in data, we calculated average of previous row and next row to update the dip in temperature and rise in humidity and lux. Figure 4(i) and Figure 4(ii).

4.2.2 Cyber (Network Router)

We converted cyber dataset in a four-step operation i) convert hourly pcap files to csv using TShark (2023) utility, ii) added start time (taken from each file name) of each file to timestamp column (_ws.col.Time) and convert into timestamp format, iii) join all csv files as one data-frame, and iv) Normalise the resulting

dataset by mapping protocols and converting strings columns into integers. The output CSV file contains the following columns:

1. frame.number: Frame number of packet
2. _ws.col.Time: Epoch formatted timestamp
3. _ws.col.Source: Source IP Address
4. _ws.col.Destination: Destination IP Address
5. _ws.col.Protocol: Protocol
6. _ws.col.Length: Length of packet (in bytes)
7. _ws.col.Info: Remaining information of the packet (structure depends on the type of protocol)

The last column "_ws.col.Info" contains information based on the previous columns e.g. column structure varies based on protocol e.g. in the case of TCP the Info column structure will be different than in case of UDP, ICMP, or ARP protocols.

4.2.3 Smart Devices Dataset (CU)

This dataset was occasionally exported from MariaDB of HA using phpMyAdmin add-on in CSV format. These CSV files have duplicate entries which were dropped after reading the CSV files and creating a single CSV file containing data from start to end time. After creating a single CSV file, we extracted data for entities and save a separate file for each PID. Later, a one-second frequency timestamp was added manually from a specific time frame and then synced with each column with the new timeline. Missing values in the resulting dataset were filled up using forward fill and then backward fill operations, the outcome was saved as CUMaster dataset. A sample for this inflating process for entity binary_sensor.smartthings_contact_sensor_pid_004_moving of PID 4 is demonstrated in Figures 4(iii) and 4(iv). It is pertinent to notice that the tilt in the red lines (red lines are from the RAW dataset) shows missing values whereas purple lines (purple lines are from an inflated dataset) show continuous values in parallel to the tilted red-line.

4.2.4 Smart Devices Dataset (BRE)

The original form of this dataset is also based on multiple files (one file per week). We, at first, merge all files and converted location column along with associated sensor columns to transpose into columns for each location-sensor. Secondly, we repeated the same operation of inflating this dataset similar to CU dataset with a one-second frequency to make it ready to be synchronised with all other datasets.

4.3 Data Fusing and Integration

In this section, we discuss the data fusion and integration of multiple datasets. We compiled most information about the single and multi-dimensional datasets in Table 26 of SCD 7. It is important to note the time scales of each stage of all datasets because of the variations during data processing. Datasets' locations are mentioned in Type and Stage where Type is a directory located in the root directory and Stage is a directory located in each type directory.

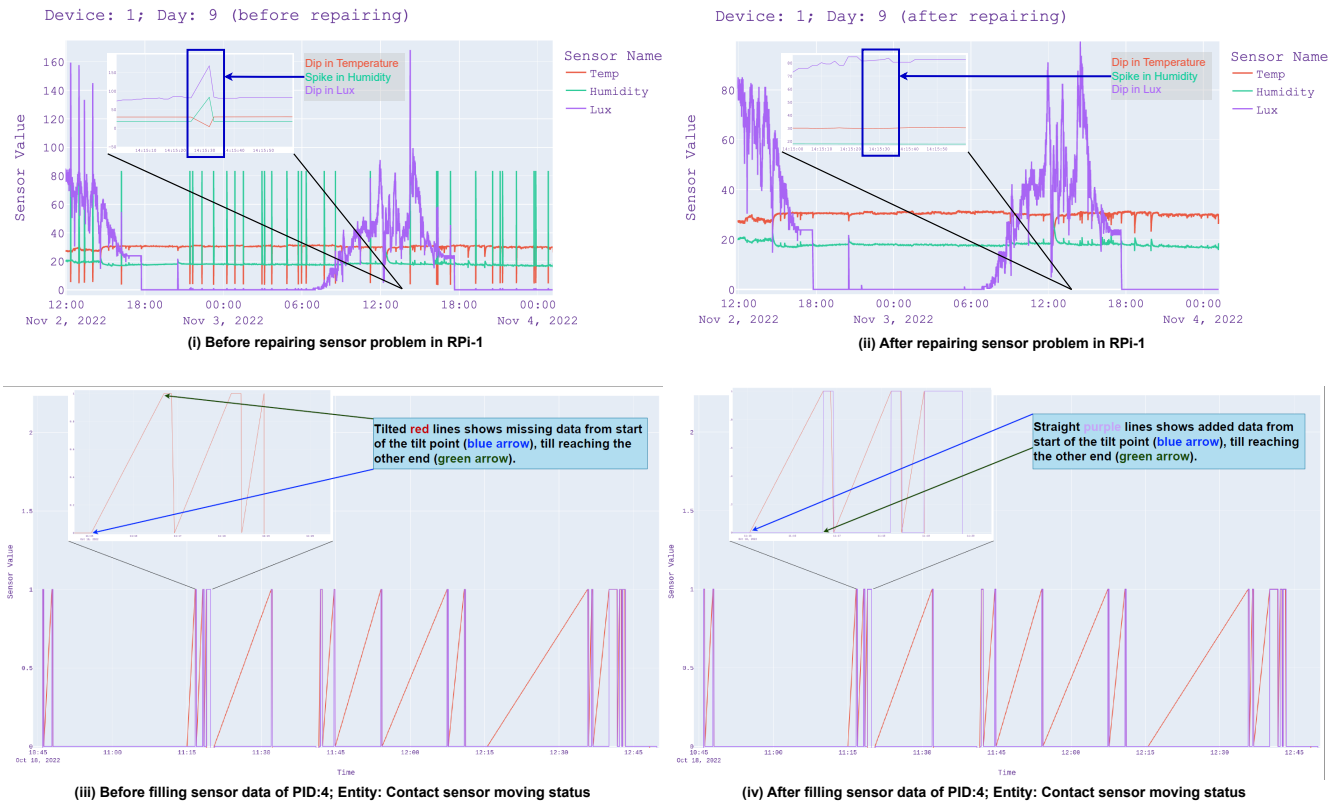


Figure 4. Data Cleaning and Filling (Missing Data)

5 ANOMALY CREATION AND ANNOTATION

5.1 Anomaly Creation Setup

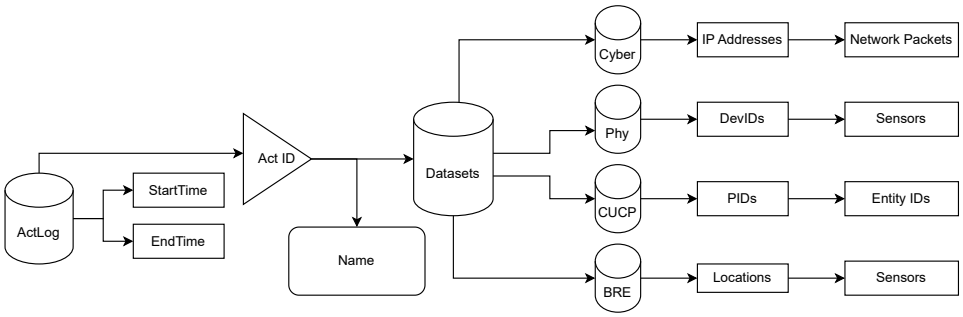
In this section, we discuss the anomaly creation setup for our dataset. The first actor performed most of the activities in the house from the start time till November 7, 2022, and after November 11, 2022, till the end of the dataset timeline. The second actor performed activities on 8-10 November 2022. Both actors didn't discuss the way they would perform the activities, thus the activity signatures should be different from each other. This labelling will help develop supervised ML algorithms or to evaluate unsupervised algorithms on the dataset. Both actors logged start time, end time, and details about the activities they performed, these files are discussed and listed in Table 27 of SCD 7.

5.2 Annotation and Labelling Process

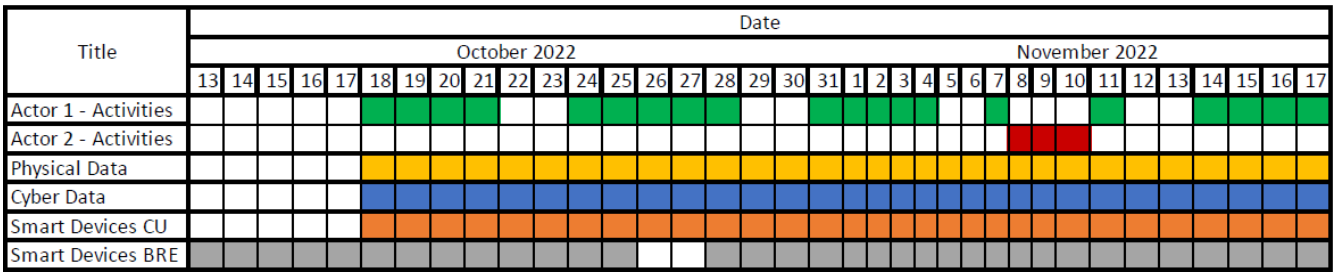
As discussed in the previous section both actors logged their activities with start time and end time. On top of that we have mapped a list of databases and associated devices which are associated with each activity, Figure 5(i) demonstrates a holistic view of these mappings. Furthermore, ActivityLabels.json contains these relations in JSON format. Figure 5(ii) shows a holistic view of the timelines of all datasets and activities of both actors.

5.2.1 Miscellaneous Files

Apart from the datasets, there are some pertinent files to be noticed. These files are listed and explained in Table 27 of SCD 7. Each file contains different information and structure, e.g. CSV files contain a list of activities performed by both actors as well as headers of Physical dataset where as JSON files contain list



(i) Activity Labelling Structure



(ii) Datasets and Actors Timelines

Figure 5. Activities Label Structure and Datasets & Actors Timelines

mapping for normalisation as well as information about devices. There is only one XLSX file that contains information about PID and entities of each device model (in case of multiple devices for one make/model).

6 DISCUSSION ON DATA QUALITY AND LIMITATIONS

In this section, we discuss quality of each dataset and then the limitations we faced during the data collection process. There were some issues faced when collecting physical dataset, the temperature, humidity, and lux readings were fluctuating randomly, which apparently looks like a hardware issue. The issue in the physical dataset was reversed but the source of the problem can not be determined. Physical dataset is based on one-second frequency due to the hardware limitations of sensors, we tried to reduce the time frame but this resulted in the script crashing. On the other hand, the SRS-30 reading also has a similar issue so we set up a five-second delay between each iteration of sensor readings. For our cyber dataset, we planned to capture all traffic on the network, whether communicating between device to device directly or via Gateway. But due to limitations in the built-in wireless adapter, which can only handle 9 maximum wireless clients simultaneously, we installed an extra WAP in the system. With this trade-off, we were only able to capture traffic from internal devices to other networks via Gateway's NAT service. In HA dataset, due to limitations in the system, the data had to be manually downloaded/exported once in a while. Due to this issue there is a duplication in the resulting dataset, which we resolve by dropping all duplicated entries in the collective dataset from all files. The BRE dataset was not captured for one week time, due to some accident that resulted in turning off the central hub. This incident was not reported next week of incident, resulting in one week of missing data.

Some of the limitations in the collection process were i) limited access to the BRE Innovation Park e.g. only on weekdays 7 AM till 7 PM, and ii) the BRE sensing network is internal and restricted so it was not possible to create a dataset pipeline to collect/ingest all data in single timestamp format from the beginning rather we have to do it in post-processing. Thus it was not possible to collect a dataset for nycthemeral

cycle of activities e.g. night sleep activities. Another limitation faced due to non-functional lavatory, one of the necessity of living, which made it out of question to record all activities inherently.

7 CONCLUSIONS AND FUTURE WORK

In this work, we presented a novel and real-world dataset from cyber (network traffic), smart devices, and physical sources in smart home settings. The dataset includes activities of two actors, both actors performed and recorded their activities at different times and in fashion. The main actor performed activities for three weeks and the secondary actor performed activities for a few days. These settings can be looked at as the training dataset for actor 1 and the testing dataset for actor 2 (considering actor 2 as an anomalous actor). We have also inflated the dataset based on the frequency of the network traffic. We created a master dataset that holds all features of various sources in a single table, which can facilitate actors for the development of ML models for users' behaviour analysis and anomaly detection in a cyber-physical environment.
For future, a nycthemeral cycle based activity dataset of similar nature (cyber-physical) may be captured and made publicly available so that the research can be done on wider ranger of activities. Another participation may be done by capturing and sharing similar dataset[s] based on longer period of time e.g., at least year, will be helpful for better understanding and bench-marking. Similar to this dataset, which was captured in a smart home scenario can be captured in buildings with data from Building Management Systems (BMS) along with its related network traffic for advancement in research on Built Environments.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

FUNDING

This work has been supported by the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which has been funded by the UK EPSRC under grant number EP/S035362/1.

SUPPLEMENTAL DATA

The Supplementary Material for this article can be found online at: [Supplementary Content] Dataset for Cyber-Physical Anomaly Detection in Smart Homes.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the IEEE Data Port `https://dx.doi.org/10.21227/sez1-2928` and Cardiff University Research Portal `http://doi.org/10.17035/d.2023.0259651425`.

REFERENCES

Alerndar, H., Ertan, H., Incel, O. D., and Ersoy, C. (2013). ARAS human activity datasets in multiple homes with multiple residents , 232–235

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. (2012). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. *International Workshop on Ambient Assisted Living and Home Care* , 216–223

Cook, D. J., Crandall, A. S., Thomas, B. L., and Krishnan, N. C. (2013). CASAS: A Smart Home in a Box. *IEEE Computer* 46, 62–69

Gallissot, M., Caelen, J., Bonnefond, N., Meillon, B., and Pons, S. (2011). Using the Multicom Domus Dataset

HomeAssistant (2023). Home Assistant. *Home Assistant* Https://www.home-assistant.io/, Accessed: 2023-05-03

Intille, S. S., Larson, K., Beaudin, J. S., Nawyn, J., Tapia, E. M., and Kaushik, P. (2005). A living laboratory for the design and evaluation of ubiquitous computing technologies. *CHI Extended Abstracts* , 1941–1944

Kelly, J. L., Kelly, J., and Knottenbelt, W. J. (2015). The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data* 2, 150007–150007

Kientz, J. A., Patel, S. N., Jones, B., Price, E., Mynatt, E. D., and Abowd, G. D. (2008). The Georgia Tech aware home. *CHI Extended Abstracts* , 3675–3680

Luca Arrotta, Claudio Bettini, and Gabriele Civitarese (2022). The MARBLE Dataset: Multi-inhabitant Activities of Daily Living Combining Wearable and Environmental Sensors Data. *Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* , 451–468

Miettinen, M., Marchal, S., Hafeez, I., Frassetto, T., Asokan, N., Sadeghi, A.-R., et al. (2017). IoT Sentinel Demo: Automated Device-Type Identification for Security Enforcement in IoT. *IEEE International Conference on Distributed Computing Systems* , 2177–2184

Tapia, E. M., Intille, S. S., and Larson, K. (2004). Activity recognition in the home setting using simple and ubiquitous sensors

TShark (2023). tshark. *TShark* Https://www.wireshark.org/docs/man-pages/tshark.html Accessed: 2023-05-03