Dataset Collection of Multi-Communication Technologies Monitored in Different Mobility Contexts

Jana Koteich, Nathalie Mitton

Dataset Collection of Multi-Communication Technologies Monitored in Different Mobility Contexts

Jana Koteich
Inria, France jana.koteich@inria.fr

Nathalie Mitton
Inria, France nathalie.mitton@inria.fr

***Abstract*—The ubiquitous nature of mobile devices equipped with radio communication technologies made the collection of data a commonplace especially for studying human mobility. The collection of such datasets forms the intermediate results in many scientific research projects. Therefore, with the lack of datasets, collecting and publishing data should be seriously addressed since several scientific research is based on the gathering and analysis of measurement data. In this paper, we introduce the PILOT dataset, a Privacy-preserving data collectIon of wireLess cOmmunication Technologies. The dataset is a collection of four jointly collected information in different mobility contexts. It includes three wireless communication technologies: WiFi probe-responses, BLE (Bluetooth Low Energy) beacons, and LoRa (Long Range Radio) packets, plus additional information: Acceleration, Roll, and Pitch, all collected at the same time. We provide the keys to reproduce such data collection and share the datasets already collected. The dataset is collected for approximately 90 hours, with a size of 200 MB using FiPy devices from Pycom and it is uploaded to GitHub. The dataset's utility is validated through the application of a classification machine learning model that determines the real-life situation of devices through the communication links monitored in different scenarios with an accuracy of 94%. Thus, we believe that such dataset is important for human mobility studies and applications of integrated sensing systems since it offers a new form of a classified collected data that does not exist in the already published datasets.**

***Index Terms*—Dataset generation, WiFi, BLE, LoRa, Acceleration, FiPy, IoT**

I. Introduction

These days, the rapidly evolving information technology and the development of wireless and mobile networks have promoted a new wave of information and industrial tide [1]. Several promising applications like tracking smartphones, traffic monitoring, crowd dynamics monitoring, and other scientific research are based on the gathering and analysis of measurement data. This increases the urge of creating new forms of datasets that provide a new perspective to analyze data and bring out new measurements. The collected datasets are mainly characterized by the *Model* and *Parameters* recorded. Each type of dataset can serve a new form of analysis, that is why there are always newly generated datasets with different characteristics. We identified a lack of a collective dataset that includes several traces from wireless communication technologies and sensors recorded at the same time in different mobility contexts. Such a dataset provides a new generation of collected data that would help in providing keys for studying human mobility or other applications. In this paper, we introduce a new approach to collect a dataset, characterized by mainly two novel approaches for collecting data, at the level of *Model* type and *Parameters* recorded, as follows: 1) Collecting different types of data from sensors and wireless communication technologies at a time: WiFi probe-response, BLE beacons, LoRa packets, and from the sensors: Acceleration, Roll and Pitch information. 2) The data is collected in different mobility scenarios and mainly classified into two categories: *Static* vs *Mobile*. The overall collected data till now spans about 90 hours in total in different mobility scenarios collected using a Micropython enabled microcontroller called FiPy device [2]. The dataset is released as a collection of text files and comma-separated values (CSV) files with mainly the timestamp, a unique identifier of the emitting device, RSSI (Received Signal Strength), and other information dedicated to each wireless technology. This dataset is privacy-preserving since it fully meets the GDPR specification, where the mac addresses and the device names are masked. We present in detail the description of the dataset and how the data is collected. The dataset is released with a ground truth annotation reporting the time intervals during which the scanning occurs, and a description of the place of scanning and uploaded to GitHub.

The organization of the paper is as follows: In Section II we present a brief overview of the collected datasets in the literature and some use cases that used wireless data in their studies. In Section III we present the PILOT dataset, with an illustration of the experimentation setup. In Section IV the methodology and description of the collected data is illustrated, then in Section V an overview of the dataset insights and data visualization is presented with an application based on the collected data. Section VI presents the main challenges for generating the datasets, and finally, Section VII concludes this paper and outlines some future work.

II. State of the Art

In this section, we provide two literature reviews, one dedicated to the different generated datasets for wireless communication traces in the literature and some presented methodologies

and tools for collecting them to highlight the uniqueness of our approach. The other sub-section provides different use cases that used wireless traces in their research work to address the importance and use cases of such datasets.

Datasets: During the last decade, several research studies have been conducted toward an efficient data collection from wireless networks and Internet of Things (IoT) devices for analysis [3]. In [4], Friesen et. al. present a complete data collection system developed at the University of Manitoba that uses a variety of wireless networking technologies and devices to collect inferred traffic data. They used XBee, GSM and Bluetooth modules for designing and implementing a slave probes network with the objective to collect Bluetooth device information. In [5], Vu et. al. introduce a new framework that collects location information and ad hoc contacts of humans at the University of Illinois campus. The Bluetooth MAC address is used to infer contact information and Wi-Fi MAC addresses are used to infer physical location of the phone. In [6], a data collection campaign and a dataset of BLE beacons for detecting and analysing human social interactions is is collected in a High School. CRAWDAD [7] dataset, is a repository of wireless network datasets, including datasets for WiFi, Bluetooth, and cellular networks that are collected at Dartmouth. UJIIndoorLoc [8], is a dataset that includes WiFi signal strength measurements and location information for a multi-story building, which can be used for indoor positioning research. Although useful, these approaches are limited to contact tracing applications in a single environment or for indoor positioning systems. In [9], the authors designed a system architecture to collect data from the IoT environment relying on BLE technology only and smartphones. Cecchinel et. al. [10] proposed an architecture to support the big data collected from IoT. This architecture is restricted to data generated from sensors like sonar and temperature sensors and is presented from a software perspective only. While [11] generated different types of network traffic data with the FIT IoT-LAB testbed. Their work includes a single technology (IEEE 802.15.4) and focuses on the delay and throughput of links under different message sizes and frequencies. Another technology under study is 5G. The authors of [12] produce 5G datasets that can be used to study 5G traffic malicious attacks and their characteristics. In [13], the authors proposed a new paradigm for generating 5G datasets for ISAC high precision positioning, named Multilevel-FSM. In [14] the dataset is designed for multimodal machine learning research in wireless communication. It consists of various scenarios where multimodal sensing and communication data samples are collected. Though it is a large-scale rich dataset but still we offer more diversity. To the best of our knowledge, no previous work has generated a labeled dataset from multi-communication wireless technologies and sensors at the same time in different mobility contexts as we propose with a new form of collected *parameters* in different *mobility* scenarios.

Dataset Usecases: Several approaches have leveraged from WiFi beacons or probe-response to characterize people's flow. As in [15], the author collected WiFi probe requests transmitted by people's smartphones, then used this data to characterize people's flows through a machine learning approach. While in [16], Gebru et. al. presented two possible methodologies for people counting and mobility detection based both on off-the-shelf hardware and commercial devices to scan the WiFi spectrum for probe requests in an urban environment, then by applying a ML-based scheme, they show how the data collected can be used to characterize people's flows. This approach is amenable to being deployed at the edge of the network. But as a further notice, Bluetooth has been exploited for some applications like traffic monitoring. As mentioned earlier, in [4], the authors exploited the collected data from Bluetooth devices to collect statistical representation of traffic density and flow. While in [17], Kulkarni et. al., deployed road-side Bluetooth scanners for traffic data collection and from Bluetooth information they have extracted traffic parameters for road traffic management. So, wireless communication has been pervasive in our daily lives. As such, collecting and analysing data from the wireless environment is an important approach for several use cases like traffic monitoring, characterizing people's flows which would help in studying human mobility, and for other services like positioning systems and others.

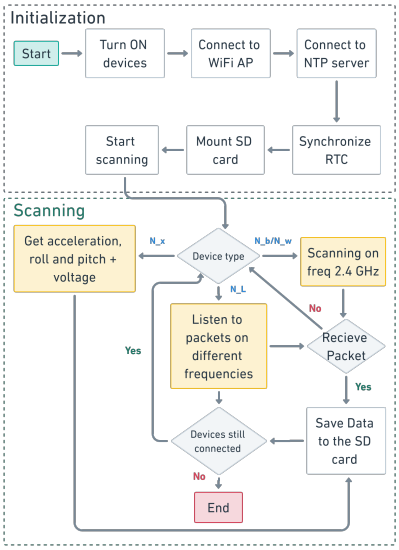


Fig. 1: Scanning process diagram

III. PILOT Dataset

PILOT dataset provides a group of collected packets from three different wireless communication technologies: WiFi, BLE and LoRa, and as additional useful information, the dataset includes the acceleration, roll, pitch and the battery voltage for collecting these data. These traces have been jointly collected all together in several mobility contexts using FiPy

microcontrollers from pycom [2]. The duration of each scanned dataset is between 10 min and 3 hours, and each recorded log file is labeled by its specific mobility scenario of scanning with a description of the conditions of scanning. The dataset is uploaded to github[1] as a collection of text files and CSV files.

A. Experimental Design

Our experimental setup (Figure 2) is the basis of the collected data and the formation of the desired dataset. The aim is to scan the different wireless communication activities in the range of the scanning device. To achieve this goal, we used Pycom FiPy devices since they support the three wireless communication technologies: WiFi, Bluetooth, and LoRa. FiPy also provides SigFox and LTE-CAT M1/ NB1, but they are not included in our scanning process since they require a subscription to an operator. The FiPy is connected to a Pytrack or Pysense expansion board from Pycom, that includes Accelerometer and an SD card module. The development boards of the FiPy include an onboard WiFi and Bluetooth antenna, and for LoRa, an external antenna is connected.

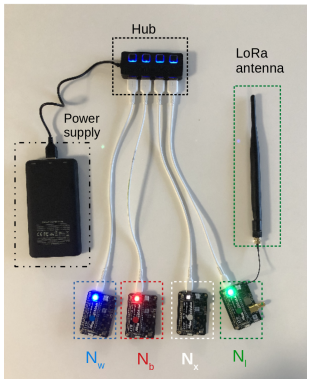


Fig. 2: Data collection setup. FiPys connected to power supply.

Since one FiPy supports several technologies at a time, first we implemented the code to collect information for the three technologies on a single microcontroller. However, we observed a significant delay in receiving BLE beacons and WiFi probe response when one device was scanning and listening to all packets. So, to avoid the delay that is caused by interruption methods, the scanning process is distributed on four devices for more precision. To get the actual time, the FiPy is connected to an access point (AP), to be able to connect to the Network Time Protocol (NTP) server that will help to synchronize the real-time clock (RTC) and get the current timestamp. The collected data by each device is saved on a secure digital (SD) card. So, the implementation runs on four FiPy devices, each one dedicated to collecting a separate type of data as follows:

- Node W (N_W): Scans for WiFi APs every 2s.
- Node B (N_B): Scans for Bluetooth devices every 1s.
- Node L (N_L): Listens for LoRa packets on different frequencies almost every 2s.
- Node X (N_X): Gets the acceleration, roll, pitch, and battery voltage/percentage.

Figure 1 shows the diagram that describes the scanning process from the start until saving the data on the SD cards.

B. Configuration

The configuration of each wireless technology is as follows:

a) WiFi Node (N_W): The WiFi node is configured to start active scanning, as the device radio transmits a probe request and listens for a probe response from an AP or active devices such as phones or laptops. The FiPy device supports 802.11b/g/n so the radio scanning is in the 2.4-GHz to 2.4835-GHz spectrum. Upon detecting a probe request, scanners log several pieces of information related to that probe, as it indicates the following named tuples: (SSID, BSSID, sec, channel, RSSI). The saved log mainly includes 1) the timestamp related to the probe request detection, 2) the service set identifier (SSID) which is the name of the device, 3) the basic service set identifier (BSSID) which is the MAC address of the radio interface the client device is currently connected to, 4) the *sec* that stands for security, 5) the channel number which is in the range of 1 to 11, and 6) the received signal strength (RSSI). The value of *sec* attribute defines the type of security where each value means the following: '0' is open, '1' is WEP, '2' is WPA-PSK, '3' is WPA2-PSK, and '4' is for WPA/WPA2-PSK.

b) Bluetooth Node (N_B): The BLE node is configured for passive scanning to receive the advertising packets (PDUs) that are retrieved every second. First, we created a Bluetooth object, then started performing a scan listening for BLE devices sending advertisements. The following named tuple with the advertisement data is received during the scanning: (mac, $addr\_type$, adv_type, $rssi$, $data$), where *data* contains the complete 31 bytes of the advertisement message. Then this *data* is parsed to get the following information: adv_flag, $scan\_tx\_pwr$, $conn\_tx\_pwr$, tx_range and adv_tx_pwr. So as with WiFi, the log is first saved with a timestamp since the advertisement is received, with the MAC address of the sender and the RSSI. Other information is retrieved like the advertising flag, where for some beacons it is unknown. Same for the $name$, in most of the received beacons the name of the device is unknown. Then we have the other information like scanning transmission power, connection transmission power, transmission range, and advertising transmission power.

c) LoRa Node (N_L): The LoRa Alliance has defined two frequency bands for the usage of LoRa technology in Europe. These bands are EU433 from 433.05 to 434.79 MHz and EU863 from 863 to 870 MHz. So in our location, only EU868 is supported. The microcontroller will be listening to different frequencies and switch between them every second. When changing frequencies, the display will show what frequency it is monitoring and the number of valid packets that have been previously seen on that frequency. In this way, we can

[1]https://github.com/Janakoteich/PILOT-Dataset-Collection-of-Multi-communication-Technologies

receive some LoRa packets that are operating on the following frequencies: 863000000, 864000000, 865000000, 866000000, 867000000, 868000000, 869000000, 864862500, 865062500, 865402500, 865602500, 865985000, 866200000, 866400000, 866600000. These frequencies are defined randomly in the configuration of the FiPy. As we increase the number of frequencies to listen on, the delay will increase for switching between selected frequencies, and as a consequence, the chance of receiving packets will decrease. So, if by chance we received data while listening on a specific frequency, the data will be saved in the file with mainly the following information: The timestamp, the spreading factor, the data itself (which will be masked for privacy issues), the frequency, RSSI, the signal-to-noise ratio (SNR), and other information.

IV. Data Collection Methodology

The dataset is collected in different scenarios, with different variations. Using the final setup (Figure 2) described in Section III, the devices are connected to the power supply to start collecting data. Each log is saved with a real timestamp in a text file on the SD card. The goal is to observe and record the variations of the wireless technologies in different mobility contexts, which are mainly categorized into two: *Static* and *Mobile* scenarios. For *Static* we defined the following cases: Home, Office, Restaurant, Bus station, University and Meetings, and for *Mobile* we have the following scenarios: Pedestrian, Car, Bus, Metro, and Trains. We took into consideration the different conditions of each scenario, as we have rural areas (auto-route), urban areas like cities, and less crowded places like villages. Some of the scenarios that fall under the category of having several conditions are illustrated below:

Train: Knowing that there are several types of trains, the data is collected in almost the different types that exist, like fast trains that travel between provinces more likely between countrysides (eg., TGV in France), or slower trains that travel between cities (like TER and Intercités in France), or those that travel between areas in a big city (eg., RER in France).

Bus: Mainly there are two kinds of buses. 1) Those that travel inside cities (urban areas) with a specific trajectory and fixed interruptions on bus stops every specified short time, and 2) buses that travel between cities (rural areas) for a long time and distance (the Autocar).

Car and Pedestrian: For Cars or pedestrians, there are mainly two conditions, being in an urban or rural place.

Home: There are several types of homes, such as ordinary apartments in rural areas, studio apartments in urban areas, collocations, platforms in a crowded city, a platform in a village, separate homes in rural areas, a Hotel, etc.

Meetings: In this category, we defined the static scenarios where people meet for professional activities such as conferences, workshops, seminars, etc.

Office: The office can be in a building surrounded by other companies and buildings. The office scenario can be categorized into a rural office and an urban office.

Figure 3, shows a map of the main train lines in France. The highlighted orange routes shows the train lines where the data was collected. The cities marked with a green circle represents the main locations of the data collected in the other scenarios (bus, home, hotel, university, etc.). Besides France, some of the datasets were collected in South Africa.

A. Privacy Preserving Techniques

According to the General Data Protection Regulation (GDPR) [18], the MAC address and the device name are considered as personal data, so to ensure the published dataset is privacy-preserving, we pseudonymize the device names which is a foundational technique to mitigate data protection risks. This is achieved by replacing the actual name of each unique MAC address with any random symbol and the MAC address is hashed using SHA-224 [19].

B. Dataset Description

One of the main characteristics of the dataset is that it is labeled. This is achieved by tracking the places and the time of scanning, then the data is retrieved from the four devices and saved in a file with a label based on the scanning conditions that were noted. In the first step we get four text files (WiFi.txt, Ble.txt, LoRa.txt, acc.txt), then using python programming language, the MAC addresses, and SSIDs' are encrypted, then each text file is transferred to a comma-separated values (CSV) file. The data is classified and uploaded on GitHub. Table I describes the organization of the dataset briefly. For each scenario, we have sub-folders named by the prefix of the name of this scenario, and each of them includes the scanned data. The table describes each scanned folder by its category, the collected information, the approximate interval of scanning with the duration, and finally the description of the scenario of scanning. At the time this paper is written, these are the collected data so far, as 120 datasets are made available with different scanning time slots and intervals for 11 mobility contexts.

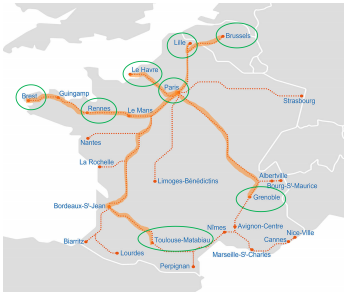


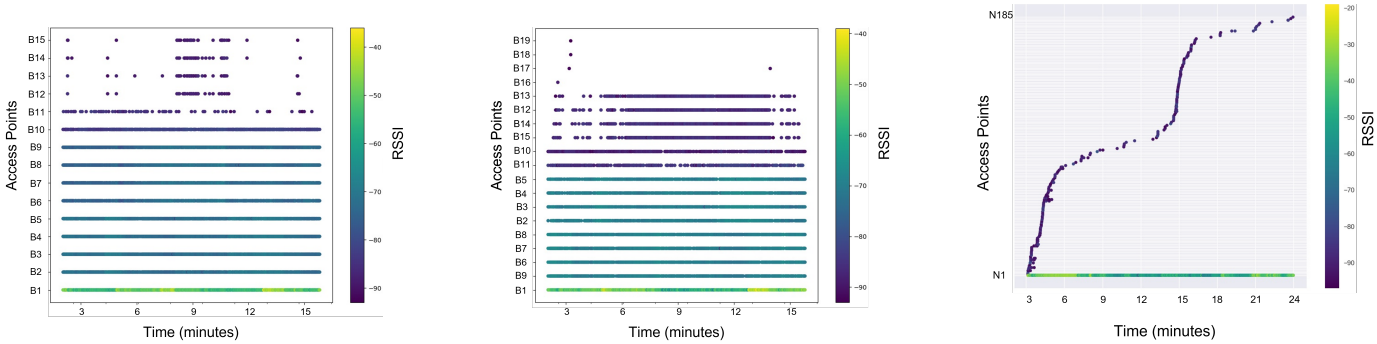
Fig. 3: Data collection locations

V. Dataset Insights Overview and Application

In this section, some primary observations of the collected data are presented. Figures 4a, 4c represent the received frames from each access point over time, with the corresponding RSSI.

TABLE I: Records from Mobile Scenarios

Mobility Scenarios	Label	WiFi	BlE	LoRa	Acceleration	From	To	Duration	Description
	Mobile								
Bus	B1	✓	✓	✓	✓	12:20:00	12:55:00	35min	Autocar between city and village - crowded
	B2	✓	✗	✓	✓	20:06:00	20:38:00	32min	Bus in a city - very crowded
Car	C1	✓	✓	✓	✓	17:26:00	18:03:00	37min	Auto-Route - rural area
	C2	✓	✓	✗	✓	13:21:00	14:04:00	43min	Auto-Route then between houses in villages
Train	T1	✓	✓	✓	✗	20:17:00	21:02:00	45min	TER between two cities
	T2	✓	✓	✓	✓	11:47:00	12:57:00	1hr, 10min	TGV
	Static								
Home	H1	✓	✓	✓	✓	08:38:00	10:42:00	2hr, 4min	Apartment in a building - village
University	U1	✓	✓	✓	✗	09:14:00	11:12:00	1hr, 58min	University campus
Office	O1	✓	✗	✓	✓	16:17:00	16:47:00	30min	Office in a rural area
	O2	✓	✓	✓	✗	17:02:00	18:13:00	1hr, 11min	Office in an urban area
Restaurant	R1	✓	✓	✓	✓	10:49:00	12:02:00	1hr, 13min	Restaurant in the city - not crowded



(a) Device S with 20 ms dwell time (b) Device L with 120 ms dwell time (c) Bus monitoring (WiFi)

Fig. 4: Dataset analysis

In Figure 4a which displays the WiFi data collected from an office, we can observe the reception of probe response frames over the whole scanning time. Knowing that the scanner is fixed, would indicate that the access points detected by the scanner are also fixed. While in Figure 4c which is related to the data collected from a bus in an urban area, the frames are appearing only for a very short duration, this is because we are losing connection with the access point because of the mobility of the bus. From these observations, we can see how each scenario has its unique pattern of received probe response frames. What is more interesting is that several datasets for the same scenario, show the same behavior of received frames. Such results could be important for building new knowledge and bringing a new perspective to analyzing human mobility.

As a further analysis, normally we should keep receiving a probe response frame from an AP as long as the AP is in the range of the device that is sending the probe request, but based on Figure 4a, we observed the absence of the frames for some time-slots even when the access point is near to the scanner and static. To investigate this issue, we had to check the configuration for scanning. Based on the implemented code for scanning, the dwell time is set to *20 ms*, as this field is used to control how long the scan dwells on each channel. So, the device is listening for 20 ms on each channel for frames then switching to the other one. To see the impact of the dwell time on the quality of scanning, we held a simple experiment as described below. First we selected two devices: **Device S** Configured with $20ms$ dwell time on each channel, and **Device L** Configured with $120ms$ dwell time on each channel. The two devices are located next to each other, and started scanning at the same time with the same conditions for 15 minutes. Figures 4a and 4b show the results of the data collected from *Device S* and *Device L* respectively. From the figures we can observe that *Device S* discovered 15 access points, while *Device L* discovered 19 access points, and *Device L* received more probe response frames from the access point **B12** than *Device S*. These observations show the impact of the dwell time on the number of received probe response frames. From these results we can deduce that to detect more access points and receive more probe response frames, it is suggested to increase the dwell time on a channel to increase the chance of discovering more access points. This observation is supported by the work done in [20] with further analysis.

Dataset Application: This dataset helps in providing new approaches for studying human mobility and understanding people's status in real-life situations. To validate this claim,

in [21] we leverage the collected traces from WiFi and BLE to determine people's real-life status. Knowing that the dataset is labeled by the scenario from which it has been collected, we built a classical machine learning model to investigate in the ability of determining people's mobility status based only on wireless information observed from their environment. Since the data is collected in 10 different scenarios and through data visualization we observed a difference between different scenarios, thus to acknowledge also this observation, it was possible through a light classical machine learning algorithm (XGBClassifier) to classify between 10 scenarios with a 94% accuracy. This is considered as a new approach for studying human mobility, thus we can see the importance of providing new forms of datasets. As a further notice, we just leveraged the WiFi and BLE data for our analysis, but we believe throughout the other parameters like the acceleration, further observations could be done.

VI. Challenges and Open Issues

From our experience, the main challenges for reproducing the dataset are at the hardware level. During scanning, it could happen that the Fipy disconnects or reboots, this causes a gap in the collected data. Also, since the Fipy does not have an internal RTC module to synchronize time, it is required to connect to an NTP server through a WiFi connection to get the real-time, but this is not applicable everywhere due to radio coverage restrictions or connection loss. Finally SD card is sometimes corrupted which causes loss of data.

Besides the hardware challenges, there is another open issue to be investigated to improve the quality of scanning which concerns LoRa. There are no beacons in LoRa, so we cannot scan to detect gateways as passive detection of a public LoRAWAN gateway is unreliable. A FiPy device can only listen on one channel and DataRates (DR) at a time, thus, we can only listen to a specific channel/data rate combination and see if by chance someone else uses the same. So, we can deduce that the radio can only monitor one frequency at a time, and we must receive an entire LoRa packet while listening to parse it, but the more time we spend on one channel, the more we might be missing from others. So, in our approach, we are still limited by the selected frequencies mentioned in Section III for listening to LoRa packets, while the goal is to capture as many as possible packets and discover the presence of LoRa base stations. Finally, knowing that the methodology for scanning and getting all four joined information at the same time was challenging, further investigations could be held to optimize the scanning setup while still getting the same quality of the output dataset.

VII. Conclusion and Future Work

In this work, we have provided a rich dataset called the PILOT dataset that includes four jointly collected information from multi-communication wireless technologies and sensors. The traces are collected in different mobility scenarios mainly categorized as *Static* and *Mobile* scenarios. The dataset is a collection of WiFi probe responses, BLE beacons, LoRa packets, and additional information from sensors like acceleration, roll, and pitch. The methodology to reproduce the dataset is illustrated, with annotations to the collected data provided on GitHub. This dataset is reproducible and will be enriched with more data with time. In the meantime, we are still collecting data for an average of 4 hours per week. In future work, we will add new scenarios for collecting data like bicycle and Motorcycle for *Mobile* scenarios, and we will add a new category for the collected dataset related to predefined mobility models that meet the expectations of more research work.

References

[1] R. Gu, H. Zhang, D. Pei, and J. Zhang, "A scalable and virtualized testbed for iot experiments," in *TRIDENTCOM*, 2018.

[2] https://docs.pycom.io/datasheets/development/fipy/ .

[3] L. Mansour and S. Moussaoui, "Cdcp: Collaborative data collection protocol in vehicular sensor networks," *Wireless Personnal Communications*, vol. 80, no. 1, 2015.

[4] M. Friesen, R. Jacob, P. Grestoni, T. Mailey, M. R. Friesen, and R. D. McLeod, "Vehicular traffic monitoring using bluetooth scanning over a wireless sensor network," *Canadian Journal of Electrical and Computer Engineering*, vol. 37, no. 3, 2014.

[5] L. Vu, K. Nahrstedt, S. Retika, and I. Gupta, "Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus," in *MSWIM*, 2010.

[6] M. Girolami, F. Mavilia, and F. Delmastro, "A bluetooth low energy dataset for the analysis of social interactions with commercial devices," *Data in Brief*, vol. 32, 2020.

[7] https://ieee-dataport.org/collections/crawdad .

[8] https://www.kaggle.com/datasets/giantuji/UjiIndoorLoc .

[9] A. E. Boualouache, O. Nouali, S. Moussaoui, and A. Derder, "A ble-based data collection system for iot," in *NTIC*, 2015.

[10] C. Cecchinel, M. Jimenez, S. Mosser, and M. Riveill, "An architecture to support the collection of big data in the internet of things," in *IEEE World congress on services*, 2014.

[11] N. Santi, R. Grünblatt, B. Foubert, A. Hameed, J. Violos, A. Leivadeas, and N. Mitton, "Automated and Reproducible Application Traces Generation for IoT Applications," in *Q2SWinet 2021*, Alicante, Spain.

[12] O. A. Fernando, H. Xiao, and J. Spring, "Developing a testbed with p4 to generate datasets for the analysis of 5g-mec security," in *IEEE WCNC*, Austin, USA, 2022.

[13] K. Gao, H. Wang, H. Lv, and W. Liu, "Toward 5g nr high-precision indoor positioning via channel frequency response: A new paradigm and dataset generation method," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 7, 2022.

[14] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "Deepsense 6g: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Comm. Magazine*, 2023.

[15] K. Gebru, "A privacy-preserving scheme for passive monitoring of people's flows through wifi beacons," in *IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, 2022.

[16] K. Gebru, M. Rapelli, R. Rusca, C. Casetti, C. F. Chiasserini, and P. Giaccone, "Edge-based passive crowd monitoring through wifi beacons," *Computer Communications*, vol. 192, pp. 163–170, 2022.

[17] A. R. Kulkarni, N. Kumar, and K. R. Rao, "Efficacy of bluetooth-based data collection for road traffic analysis and visualization using big data analytics," *Big Data Mining and Analytics*, vol. 6, no. 2, 2023.

[18] https://ec.europa.eu/info/law/law-topic/data-protection/reform/ .

[19] https://www.rfc-editor.org/rfc/rfc3874.txt .

[20] T. Choi, Y. Chon, and H. Cha, "Energy-efficient wifi scanning for localization," *Pervasive and Mobile Computing*, vol. 37, 2017.

[21] J. Koteich and N. Mitton, "Machine Learning Approach for Mobility Context Classification using Radio Beacons," in *MASCOTS2023 IEEE*, New York, United States, Oct. 2023.