On the Use and Construction of Wi-Fi Fingerprint Databases for Large-Scale Multi-Building and Multi-Floor Indoor Localization: A Case Study of the UJIIndoorLoc Database

Sihao Li, *Student Member, IEEE*, Zhe Tang, *Student Member, IEEE*, Kyeong Soo Kim, *Senior Member, IEEE*, and Jeremy S. Smith, *Member, IEEE*

***Abstract*—Large-scale multi-building and multi-floor indoor localization has recently been the focus of intense research in indoor localization based on Wi-Fi fingerprinting. Although significant progress has been made in developing indoor localization algorithms, few studies are dedicated to the critical issues of using existing and constructing new Wi-Fi fingerprint databases, especially for large-scale multi-building and multi-floor indoor localization. In this paper, we first identify the challenges in using and constructing Wi-Fi fingerprint databases for large-scale multi-building and multi-floor indoor localization and then provide our recommendations for those challenges based on a case study of the UJIIndoorLoc database, which is the most popular, publicly-available Wi-Fi fingerprint multi-building and multi-floor database. Through the case study, we investigate its statistical characteristics with a focus on the three aspects of (1) the properties of detected wireless access points, (2) the number, distribution, and quality of labels, and (3) the composition of the database records, and then identify potential issues and ways to address them in using the UJIIndoorLoc database. Based on the results from the case study, we not only provide valuable insights on the use of existing databases but also give important directions for the design and construction of new databases for large-scale multi-building and multi-floor indoor localization in the future.**

***Index Terms*—Indoor localization, Wi-Fi fingerprint database, UJIIndoorLoc.**

I. Introduction

THE rapid development of technologies for wireless communication and mobile devices has brought about a host of new applications and services, a notable example of which is location-based services (LBS) [1]. Global navigation satellite systems (GNSS) and cellular networks can accurately localize mobile users and devices in an outdoor environment, which, however, cannot be a viable option for indoor localization due to the lack of line-of-sight (LOS) signal paths [2]. Since the location of mobile users and devices is essential to LBS, the indoor localization techniques not based on GNSS and cellular networks have been increasingly attracting attention from both researchers and practitioners.

Indoor localization can be done with and without ranging [1]. Ranging-based approaches rely on the distances between a point of interest and multiple known locations—i.e., *anchor nodes* like wireless access points (WAPs)—or differences of them in determining the unknown position of the point of interest through trilateration/multilateration. Time of arrival (ToA) technique uses the travel time between the unknown location of a user or a device and anchor nodes [3], [4], while time difference of arrival (TDoA) technique uses the time differences between the arrivals of the user's signals at anchor nodes [5], [6]. Instead of arrival times or their differences, angle of arrival (AoA) technique uses angles of signal arrivals, which can be estimated by measuring time differences of arrivals between individual elements of an antenna array [7], [8]. Due to the reflections and multi-path interference introduced by indoor structures and obstructions, however, it is challenging to accurately estimate the arrival times or their differences of received radio signals through non-line-of-sight (NLOS) signal paths in time-based approaches. Likewise, the angle measurement in AOA techniques could be affected by indoor obstacles as well as the user's body posture and way of carrying devices. In large-scale multi-building and multi-floor indoor localization, the performance of ranging-based techniques cannot be comparable to that of single-floor indoor localization. Given the requirement of deploying numerous anchor nodes throughout buildings and floors, we can conclude that ranging-based techniques are unsuitable for large-scale multi-building and multi-floor indoor localization.

In ranging-free indoor localization techniques, the location of a user or a device is not estimated based on distance-related information with trilateration/multilateration. In the case of *location fingerprinting*, which is by far the most popular indoor localization technique, the information measured at a point of interest—e.g., received signal strength indicator (RSSI), channel state information (CSI), and geomagnetic field intensity—is used for localization, which is supposed to be unique to each location and thereby serves as a location fingerprint. Specifically, in Wi-Fi fingerprinting using RSSIs as its location fingerprints, there are offline and online phases in its operation [9]: During the offline phase, the RSSIs at known locations—called reference points (RPs) in the literature—are collected and stored in a fingerprint database; during the online phase, the current location of a user or a device

This work was supported in part by the Postgraduate Research Scholarships (under Grant PGRS1912001), the Key Program Special Fund (under Grant KSF-E-25), and the Research Enhancement Fund (under Grant REF-19-01-03) of Xi'an Jiaotong-Liverpool University.

Sihao Li and Zhe Tang are with the School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, P. R. China, and also with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, U.K. (e-mail: {Sihao.Li19, Zhe.Tang15}@student.xjtlu.edu.cn).

Kyeong Soo Kim is with the Department of Communications and Networking, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, P. R. China (e-mail: Kyeongsoo.Kim@xjtlu.edu.cn).

Jeremy S. Smith is with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3GJ, U.K. (e-mail: J.S.Smith@liverpool.ac.uk).

is estimated based on the RPs whose RSSIs most closely match the newly-measured RSSIs at the location. One of the most significant advantages of Wi-Fi fingerprinting is that it does not require additional hardware or infrastructure (e.g., costly sensors and dedicated base stations serving as anchor nodes) and, therefore, can be used in any environment equipped with Wi-Fi networks, including offices, hospitals, campuses, and shopping malls, making it a cost-effective solution. For Wi-Fi fingerprinting, CSI also can be used as location fingerprints, which, unlike coarse-grained RSSIs, can provide fine-grained indicators consisting of both amplitude and phase information during signal propagation and enables even single-WAP localization. A significant drawback of CSI-based Wi-Fi fingerprinting, however, is the requirement of unique network interface cards (NICs) and device drivers for the acquirement of CSI (e.g., Intel 5300 NICs) [10].

Due to their recent evolution and penetration into many areas as new enabling technologies, artificial intelligence (AI)/machine learning (ML) algorithms are frequently used to improve the performance of location fingerprinting techniques. The K-nearest neighbors (KNN) algorithm is one of the most employed ML algorithms for indoor localization due to its simplicity and efficiency [11], [12]. However, its localization performance would be degraded by the complexity of indoor environments, especially in large-scale multi-building and multi-floor indoor localization, due to the increased spatial variability and dynamics of Wi-Fi signals. Moreover, KNN relies on stationary signal information and does not handle temporal dynamics well, further limiting its effectiveness. In contrast, deep neural networks (DNNs) can address the issues in large-scale indoor localization [13], [14], which can model complex relationships between the input features and output labels and thereby efficiently handle the spatial variability and dynamic signals encountered in large-scale multi-building and multi-floor indoor localization. In addition to classical feedforward neural networks (FNNs) used in earlier works (e.g., [13]), more advanced DNNs like convolutional neural networks (CNNs) [15], [16] and recurrent neural networks (RNNs) [17] are employed as well due to their improved robustness and generalization capability.

Small-scale indoor localization covers only a single floor of multi-floor buildings or enclosed space (e.g., a classroom) and utilizes location fingerprints based on a limited number of WAPs. In such an environment, constructing and managing a fingerprint database is relatively straightforward, and most indoor localization algorithms can accurately estimate the location of a user or a device. However, it is not that straightforward to extend not only the way of constructing and managing fingerprint databases but also localization algorithms to large-scale multi-building and multi-floor indoor localization, where the characteristics and the dimension of input signals are considerably complicated and more extensive. Large-scale multi-building and multi-floor indoor localization have to address the following unique issues in comparison to their small-scale counterpart:

- *Scalability*: The higher dimension (i.e., the number of detected WAPs) and the large number (i.e., the number of RPs) of location fingerprints.
- *Irregularity*: Differences in location coverages and internal structures (e.g., floor plans) among buildings and floors and uneven spatial distribution of RPs.

Note that the localization performance heavily depends on the underlying Wi-Fi fingerprint database. There are bodies of research reporting outstanding performance of their proposed indoor localization algorithms, which are based on their custom-built fingerprint databases covering limited areas with simple internal structures like corridors and Lab spaces, mainly because constructing large fingerprint databases is labor-intensive and time-consuming; even worse, most of those fingerprint databases are not publicly available. However, it is desirable to compare a newly-proposed indoor localization algorithm with the existing ones on an equal basis, preferably based on publicly-available, well-established fingerprint databases. For this purpose, Torres-Sospedra et al. provided the *UJIIndoorLoc database* [18], a large-scale publicly-available multi-building and multi-floor Wi-Fi fingerprint database, which has been the most widely-used fingerprint database for benchmarking multi-building and multi-floor indoor localization algorithms in the literature. Though there have been numerous studies on indoor localization based on the UJIIndoorLoc database, a systematic case study dedicated to large-scale multi-building and multi-floor Wi-Fi fingerprint databases has yet to be seen.

In this paper, therefore, we take the UJIIndoorLoc database as a representative example of large-scale multi-building and multi-floor Wi-Fi fingerprint databases and investigate its statistical characteristics based on comprehensive analyses to identify potential issues and ways to address them in constructing and using a Wi-Fi fingerprint database for large-scale multi-building and multi-floor indoor localization. The results of our work in this paper provide valuable insights into the use of existing databases and give important directions for the design and construction of new databases in the future.

The rest of the paper is organized as follows: Section II reviews publicly-available fingerprint databases that are well-known in the literature. Section III presents the results of the case study of the UJIIndoorLoc database. Section IV discusses the challenges and provides recommendations in constructing and using large-scale multi-building and multi-floor Wi-Fi fingerprint databases. Section V concludes our work in this paper.

II. Review of Publicly-Available Fingerprint Databases

We review publicly-available fingerprint databases well known in the literature and provide their taxonomy in this section.

A. UJIIndoorLoc

UJIIndoorLoc is the first publicly-available multi-building and multi-floor Wi-Fi fingerprint database, which covers a total surface of over 108,000 m^2 of three four- or five-floor buildings on the University Jaume I (UJI) campus in Castelló de la Plana, Spain [18].

The UJIIndoorLoc database provides 21,048[1] records measured at pre-established RPs (933 in total) and random locations. To guarantee statistical independence between datasets, the validation dataset of the UJIIndoorLoc database was measured three months after the training dataset. The UJIIndoorLoc database is quite flexible in that the localization with it can be based on *classification* of building, floor, and location identifiers (IDs), *regression* of three-dimensional (3D) location coordinates, or *their hybrid* given its large-scale multi-building and multi-floor nature.

Due to these advantages and flexibility, the UJIIndoorLoc database becomes the most widely-used reference for benchmarking multi-building and multi-floor indoor localization algorithms in the literature (e.g., selected as the official database of the IPIN 2015 competition [19]).

B. WicLoc

WicLoc fingerprint database covers the tenth floor of the new main building at the Beihang University in Beijing, China [20]. The floor occupies an area of about 1,600 m^2 and has 28 rooms with a size of 3.75 m×8 m each and a circular corridor. The WicLoc database consists of the users' daily location information, corresponding Wi-Fi RSSIs, and other characteristics like users' step counts and turns based on *crowdsourcing*, which could significantly reduce the labor cost for data collection. More than one hundred WAPs detected in the floor are segmented into pre-defined 2 m×2 m grids.

C. TUT 2017 and TUT 2018

TUT 2017 and TUT 2018 are single-building and multi-floor Wi-Fi fingerprint databases from the Tampere University of Technology in Tampere, Finland, the details of which are described in [21] and [22], respectively. Both databases provide training and test datasets, whose record consists of RSSIs, floor, longitude, and latitude. The TUT 2017 database is based on the records collected at random RPs (i.e., no grid-based or pre-established mapping) by volunteers with 21 devices in a five-floor building with a footprint of about 22,570 m^2 (i.e., a size of about 208 m×108 m). The TUT 2018 database, on the other hand, is based on the records collected at grid-based RPs in a three-floor building, where grid spacing of 5 m×1 m is used for training and test datasets, respectively.

D. BLE RSSI Database based on Apple iBeacon

This Bluetooth low energy (BLE) RSSI database was created for a real-world evaluation of the semi-supervised deep reinforcement learning (DRL) model proposed for indoor localization in the context of a smart city [23]. The database is based on the records by iPhone 6S smartphones of the RSSIs from Apple iBeacon [24] devices mounted on the ceiling of the first floor of the Waldo Library at Western Michigan University in Michigan, USA. A grid of 13 iBeacon devices was deployed to cover an area of 200 ft.×180 ft., which contains many pillars blocking the propagation of iBeacon signals. The database provides 820 and 600 labeled data points—i.e., labels composed of location, timestamp, and RSSIs—for training and testing, respectively, and 5,200 unlabeled data points for semi-supervised learning.

[1]The total number of records mentioned in [18] is 21,049, but the actual number of records contained in the two CSV files (i.e., "trainingData.csv" and "validationData.cvs") publicly released by the same group is 21,048.

E. Taxonomy of Fingerprint Databases

Table I provides a taxonomy of the fingerprint databases discussed in this section.

First, fingerprint databases can be categorized into *single-floor*, *single-building and multi-floor*, and *multi-building and multi-floor* databases based on their coverage.

Second, as for the way of collecting RSSIs, there are two categories of *crowdsourcing*, which is based on volunteers, and *insourcing*, which is purposefully designed and systematically carried out by the participants of the project. The former often leads to crowded RSSI records at random RPs but requires less effort since only volunteers are required. The latter may result in a more structured and tidier RSSI database, but it requires considerable labor during the construction.

Third, based on the capabilities of RSSI measurement devices, the databases are also categorized into the two groups of *Wi-Fi RSSI* and *multi-network RSSI*, the latter of which includes RSSIs for other types of networks (e.g., cellular networks and BLE) as well.

III. A Case Study of the UJIIndoorLoc Database

In this section, we present the results of our statistical investigation of the UJIIndoorLoc database. The UJIIndoorLoc database [18] provides 21,048 records (i.e., measurements) in total—i.e., 19,937 records in the training dataset and 1,111 records in the validation dataset—covering the three buildings on the UJI campus, each of which consists of RSSIs from 520 WAPs and 9 fields of location and measurement information (i.e., 529 fields per record). The histograms in Fig. 1 show the overall distributions of the RSSIs in the training and validation datasets.

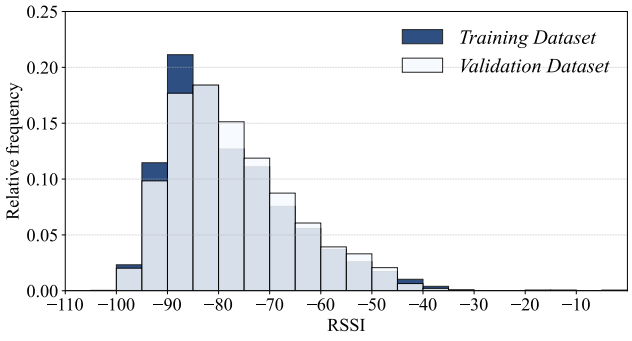


Fig. 1: RSSI histograms.

A. Label-Based Record Analyses

The quality of indoor localization depends on the location and measurement information provided by a fingerprint label. Table II lists the fields of the UJIIndoorLoc fingerprint label, where the first six and the rest three are for location and measurement information, respectively.

TABLE I: Taxonomy of publicly-available fingerprint databases.

Category	Implementation	UJIIndoorLoc [18]	WicLoc [20]	TUT 2017 [21]	TUT 2018 [22]	iBeacon [23]
Coverage	Single-floor.		✓			✓
	Single-building and multi-floor			✓	✓	
	Multi-building and multi-floor	✓				
Collection	Crowdsourcing		✓	✓		
	Insourcing	✓			✓	✓
Fingerprint	Wi-Fi RSSI	✓	✓	✓	✓	
	Multi-network RSSI					✓

TABLE II: Fields of the UJIIndoorLoc fingerprint label.

Field	Description	Range of Values
BUILDINGID	Building identification	[0, 2]
FLOOR	Floor identification	[0, 4]
LONGITUDE	Longitude	
LATITUDE	Latitude	
SPACEID	Room identification	[1, 254]
RELATIVEPOSITION	Room/corridor marker	[1, 2]
USERID	User identification	[0, 18]
PHONEID	Phone identification	[0, 24]
TIMESTAMP	Capture time	

1) Location Information: Fig. 2 shows building-level record distributions of the UJIIndoorLoc database, where there is a noticeable difference in the record distributions of the training and validation datasets. Fig. 2 (a) shows that,

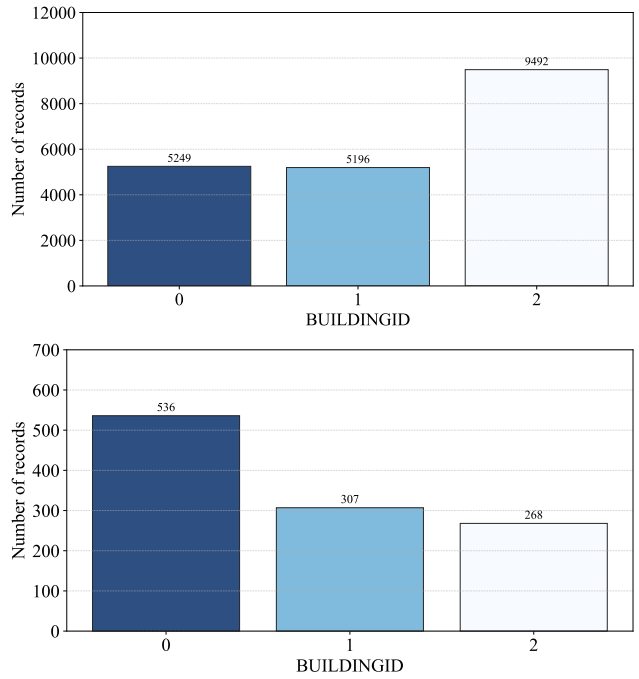


Fig. 2: Building-level record distributions: (a) Training and (b) validation datasets.

though Building 2 have just one more floor than Building 0 and 1, its number of records in the training dataset is almost the same as the sum of those of Building 0 and 1. Regarding the validation dataset, on the other hand, it is Building 0 that has the most number of records as shown in Fig. 2 (b). Fig. 3 also shows uneven record distributions over floors of each building, which results from the accessibility of the rooms (e.g., the rooms on the third floor of each building are more easily accessible than the others in the training dataset).

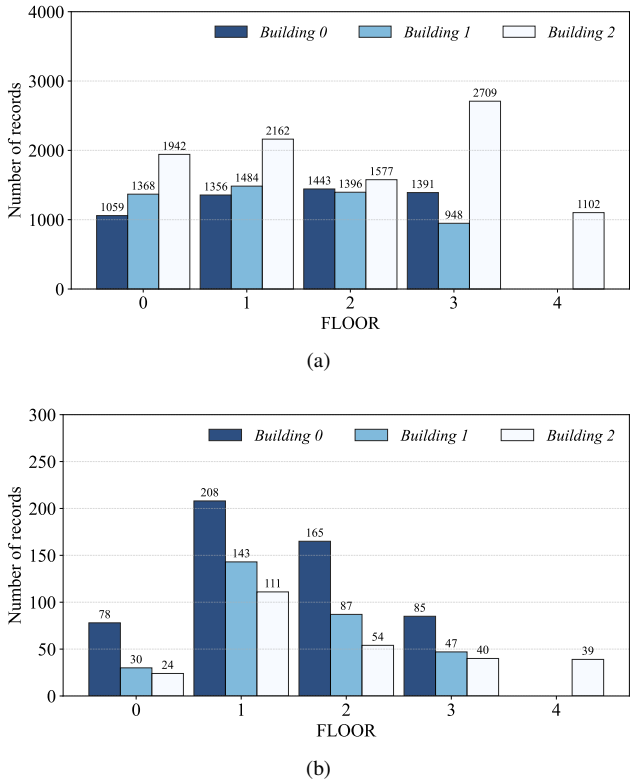


Fig. 3: Floor-level record distributions: (a) Training and (b) validation datasets.

The uneven record distribution of the training dataset could be explained by its building-level RP distribution and the histogram for the number of records per RP in Figs. 4 and 5. Note that RPs were established only for the training dataset in the UJIIndoorLoc database [18, Table X]; the records of the validation dataset do not provide SPACEID and RELATIVEPOSITION on which RPs are based. With the location coordinates of LONGITUDE and LATITUDE alone, therefore, the number of records per RP cannot be calculated.

As shown in Fig. 4, the number of RPs in Building 2 is significantly larger than those in Building 0 and 1—i.e., like the numbers of records shown in Fig. 2 (a)—due to the allocation of RPs for the training dataset of the UJIIndoorLoc database [18]. Considering the RP-level record distribution of the training dataset shown in Fig. 5 together, we can conclude that the allocation of RPs and repeated RSSI measurements at those RPs are critical to the record distributions over buildings and floors.

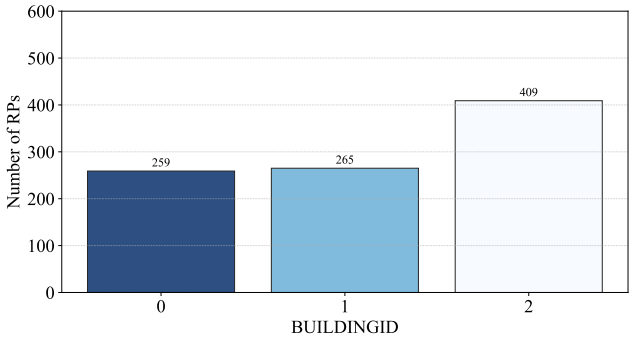


Fig. 4: Building-level RP distribution of the training dataset.

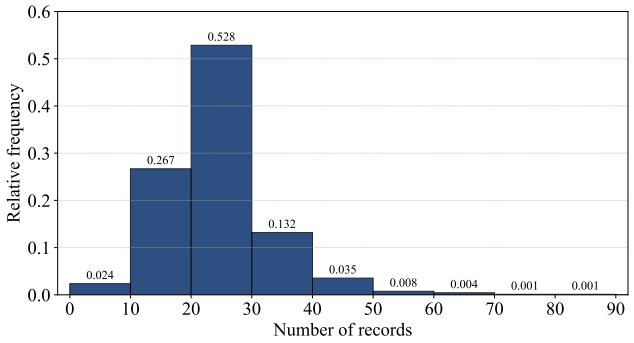
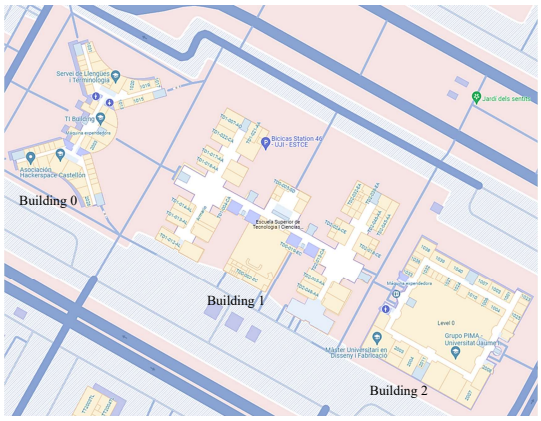


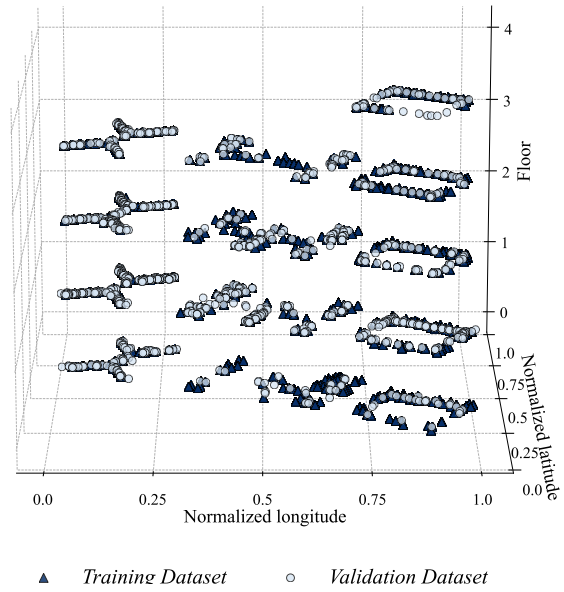
Fig. 5: Histogram for the number of records per RP of the training dataset.

Fig. 6 (a) and (b) show the floor plans of the three buildings on the UJI campus map with their corresponding BUILDINGIDs and the 3D visualization of record coordinates of LONGITUDE, LATITUDE, and FLOOR, respectively. As expected from the floor plans, the spatial distribution of the records of Building 1 is much more complicated than those of Building 0 and 2. One noticeable feature of the spatial distributions is that the coverage of the records of the training dataset is quite incomplete on the top floor of Building 2, where there are virtually no training records for the lower right part of the floor, while the validation records are more or less evenly distributed along the corridors. Such clear coverage gaps between the datasets can negatively affect the model's localization performance on the validation dataset. Data augmentation techniques can be used in this regard to fill the spatial gaps in the training dataset and thereby improve the generalization capabilities and localization performance of a model [25].

2) Users and Phones: The UJIIndoorLoc database was constructed based on the measurements by 18 users for the training dataset (i.e., USER 1 to 18), but the USERID field was not recorded and set to 0 for the validation dataset. Fig. 7 shows a user-level record distribution of the training dataset, where we can observe that there are a few dominant contributors like USER 11 with 4,516 records and USER 1 with 2,737 records. Likewise, we can also identify the dominance of a couple of phones in Fig. 8, which shows phone-level record distributions.



(a)



(b)

Fig. 6: Record spatial distributions: (a) Buildings on the UJI campus map from Google Earth and (b) 3D visualization of record coordinates.

Note that the position of a user's phone (e.g., height and direction) affects RSSI measurements as does the model and software/firmware version of the phone; at the exact the same location, a combination of different user and phone could result in different RSSI values. Given the dominance of a few users and phones in constructing the database, we can also apply techniques like data augmentation [25] and data resampling [26] to reduce the risk of bias and overfitting and thereby improve a model's generalization as discussed in Section III-A1.

The data analysis in Fig. 8 reveals that PHONE 13 and 14 hold the majority of the training dataset records. Additionally, PHONE 13 is the most significant contributor to the validation dataset. Notably, there are a few instances where the same phones appear in both datasets, which challenges the model's generalization to new, unseen phones. This observation is critical as it accurately reflects the real-world scenario, where the phones used may not be limited to those in the training

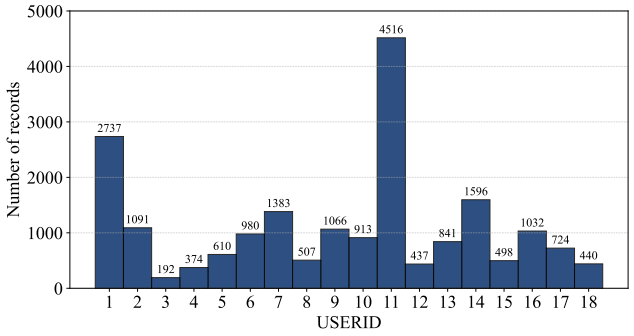
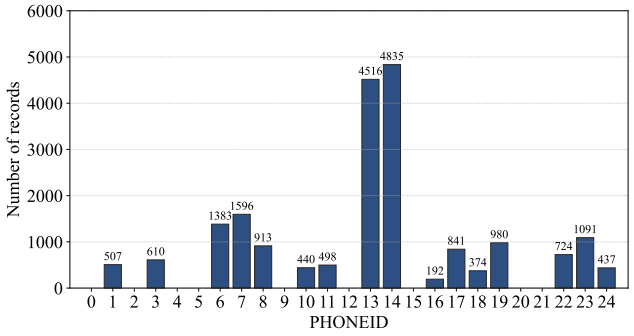
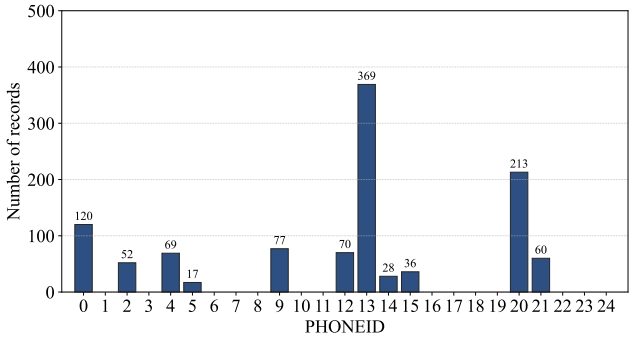


Fig. 7: User-level record distribution of the training dataset.



(a)



(b)

Fig. 8: Phone-level record distributions of the (a) training and (b) validation datasets.

dataset. Note that, without the records by PHONE 13 and 14, the remaining records of the training dataset shown in Fig. 9 quite poorly cover the entire space in comparison to the whole records shown in Fig. 6 (b).

3) Timestamp: The time of record capture is provided by TIMESTAMP field in Unix time format, which was set by a centralized server to avoid the issue of devices' different timing settings [18]. As shown in Fig. 10, RSSIs measured with the same phone at the same location but at different times (i.e., just 13 seconds' difference in the example) could be different due to several factors such as time-varying wireless channels and the movements of a user measuring RSSIs, which affects the accuracy and reliability of indoor localization systems. This could be a major issue specifically for highly dynamic indoor environments such as airports, shopping malls,

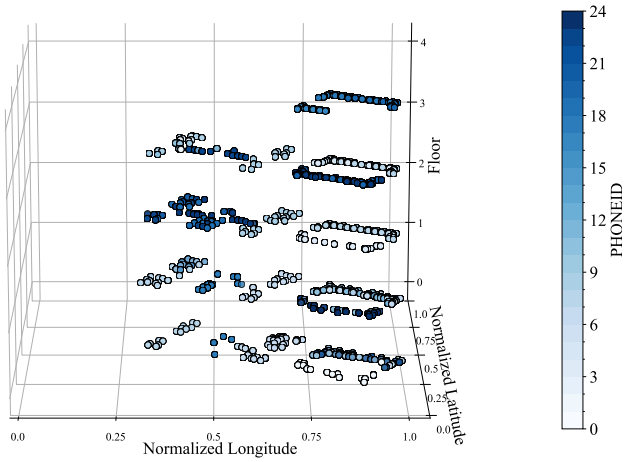


Fig. 9: Phone spatial distribution of the training dataset (excluding PHONE 13 and 14).

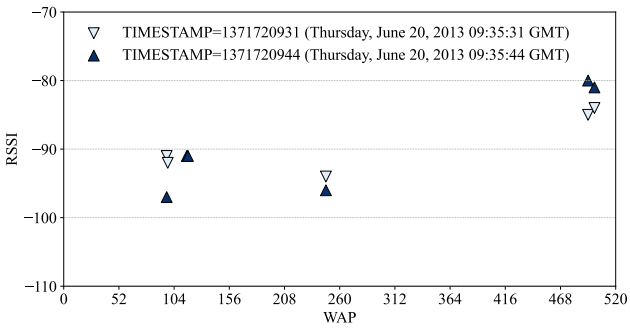


Fig. 10: RSSIs in the training dataset measured at different times with PHONE 19 at the same location of LONGITUDE=-7300.81899 and LATITUDE=4864817.599.

and hospitals.

B. Field-Level RSSI Analyses

For the analyses of RSSIs presented in this section, we exclude the RSSI values of undetected WAPs in each record by setting them to *NaN*, which stands for "Not a Number".

1) Statistical Characteristics: Fig. 11 shows the scatter plots of the field-level RSSI mean, median, and first and third quartiles of both training and validation datasets, while Fig. 12 shows the histograms of their means, medians, and standard deviations. By *field-level*, we mean each of the 520 RSSI columns (i.e., fields) of the UJIIndoorLoc database, the RSSIs of which are from one specific WAP.

From the field-level RSSI statistics shown in Fig. 11, we can observe that most of the means and medians of both training and validation datasets fall between -95 and -70, though the spread of the training dataset is slightly greater than that of the validation dataset. As most of the RSSI values are weak (i.e., <-70) and volatile [27], they are susceptible to environmental changes such as people and furniture movement. As shown in Fig. 12 (c), it is also worth to note that the standard deviations of the RSSIs are less than 2 for around 15% of the WAPs. This would indicate a situation where these WAPs yield similar

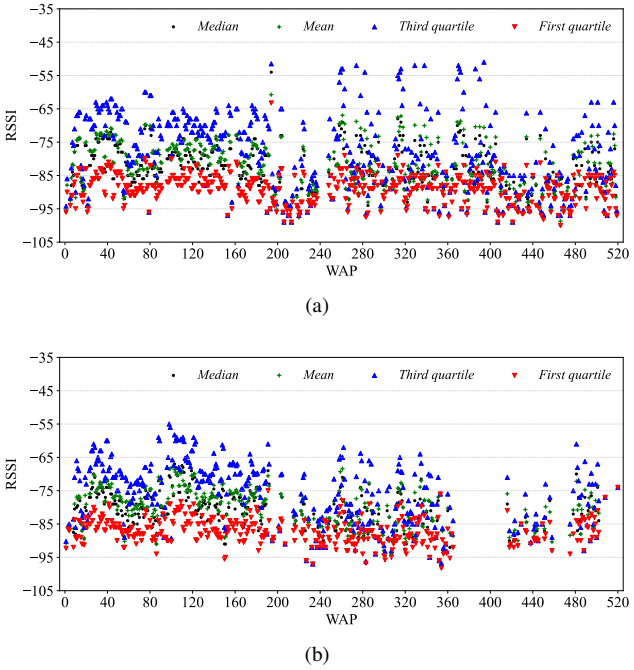


Fig. 11: Scatter plots of the field-level RSSI mean, median, and first and third quartiles of (a) the training and (b) validation datasets.

RSSIs (e.g., they are barely detected at only a few RPs), which are of little help during the localization.

From Fig. 11, we also observe that some WAPs are present in either of the datasets but not both: For example, Fig. 11 (a) shows a gab of WAPs around 240 in the training dataset, while Fig. 11 (b) shows the absence of WAPs in the range of [370, 410]. We can think of the following reasons for the inconsistencies of detected WAPs between the datasets:

- There are changes in the environments and the WAPs such as the addition of new furniture or equipment blocking the signals from WAPs and the replacement of failed WAPs with new ones. Note that a validation dataset was constructed three months after the training dataset.
- Some WAPs are not detected by certain phones due to hardware and software issues. Again, note that different sets of phones, with a few dominant ones, were used for the training and validation datasets as discussed in Section III-A2. The diversity of phones used in collecting fingerprints is the key to address this.
- There are significant differences in the measurement locations between the datasets as discussed with Fig. 6 (b) in Section III-A1.

2) Unique Values: Unique values of the RSSI from each WAP can provide an insight into the distinct characteristics of the corresponding WAP in indoor localization. Specifically, through the analyses of field-level unique values, we can identify the RSSI values that are constant over multiple locations and/or measurements (e.g., weak signals that are barely detectable) and thereby hardly contribute to location fingerprinting due to their lack of uniqueness. Those values can be removed to reduce the cost of maintaining large-scale multi-building and multi-floor Wi-Fi fingerprint databases, which not only speeds up training but also lower the risk of overfitting.

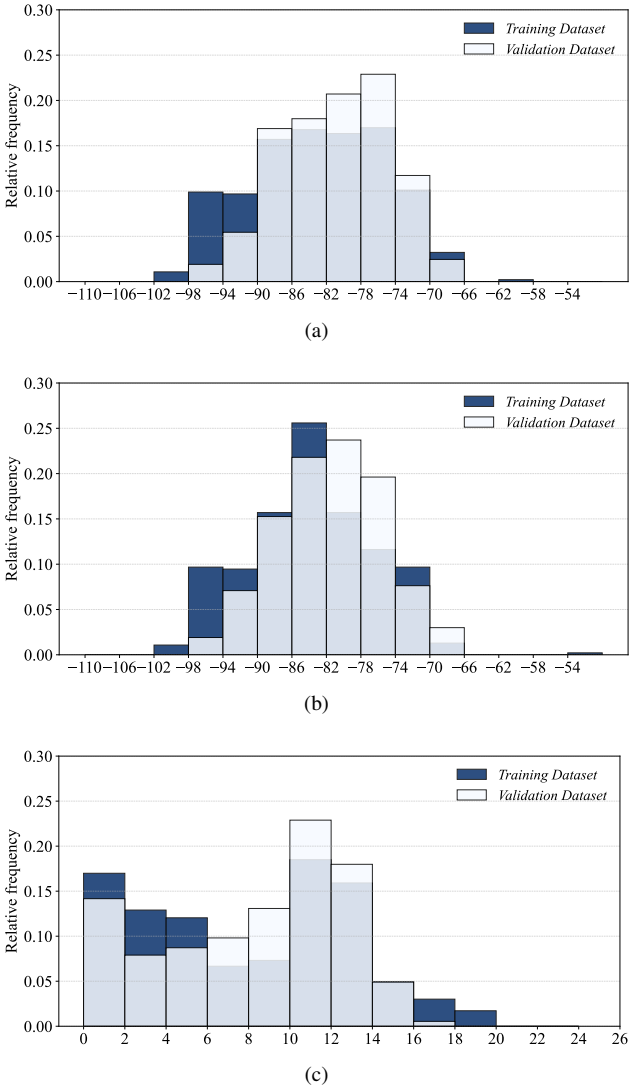


Fig. 12: Histograms of the (a) mean, (b) median, and (c) standard deviation of field-level RSSIs.

Fig. 13 shows the histograms of the number of unique RSSI values per WAP. The histograms reveal that a significant number of WAPs have smaller numbers of unique values; about one-third of the WAPs in the training dataset and half of the WAPs in the validation dataset have unique values of less than five. This implies that the signal strengths from those WAPs are so weak that they are detected only at a few RPs or their RSSIs are identical across many RPs. The scatter plots of building-level unique RSSI values shown in Fig. 14 visualizes the significance of each WAP in a clear way.

3) Spatial Distribution: Fig. 15 shows an example of field-level RSSI spatial distributions, which is based on the RSSIs from WAP486 of the training dataset. From Fig. 15, we can observe that, though WAP486 is detected at various places, its RSSIs are much stronger on the second floor of Building 2, which would indicate its deployment location.

Estimating the locations of all WAPs in a large-scale multi-

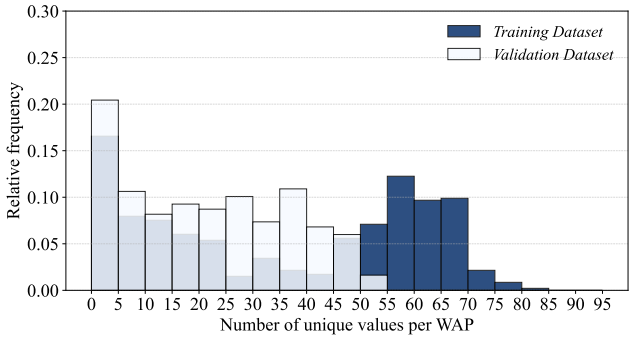
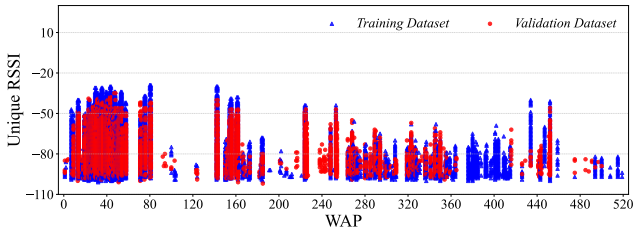
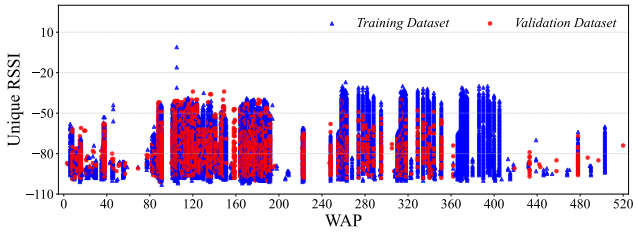


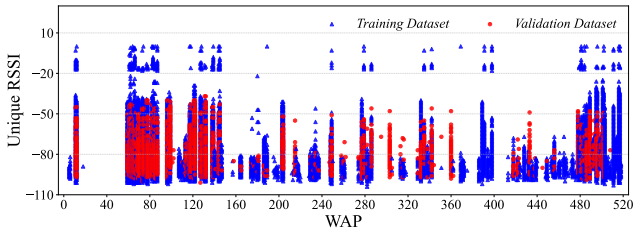
Fig. 13: Histograms of the number of unique RSSI values per WAP.



(a)



(b)



(c)

Fig. 14: Scatter plots of unique RSSI values per WAP: (a) Building 0, (b) Building 1, and (c) Building 2.

building and multi-floor environment is not practical, but our discussions based on Fig. 15 provides a practical alternative: If few critical WAPs can be identified or a target area can be hierarchically divided into smaller sub-regions, we can focus on a small number of WAPs with a reduced input dimension and approximately estimate their locations through the analysis of their RSSI spatial distributions.

C. Record-Level Fingerprint Analyses

Due to the lack of information for SPACEID, RELATIVEPOSITION, and USERID in the validation dataset, we base the record-level fingerprint analyses mainly on the training dataset.

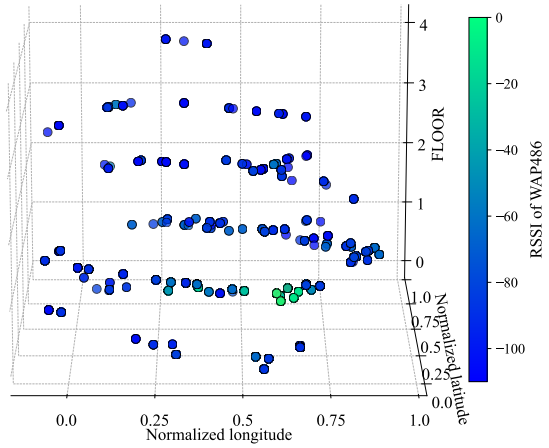


Fig. 15: Spatial distribution of the RSSIs from WAP486 of the training dataset in Building 2.

Fig. 16 shows the distributions of the number of detected WAPs per record for both training and validation datasets. The bins with the highest relative frequency for the training and validation datasets are [16, 20] and [12, 16], respectively, which is understandable given the much smaller numbers of records, users, and phones and the shorter measurement period of the validation dataset.

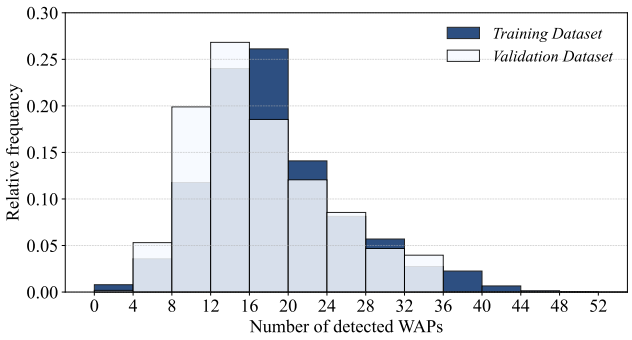
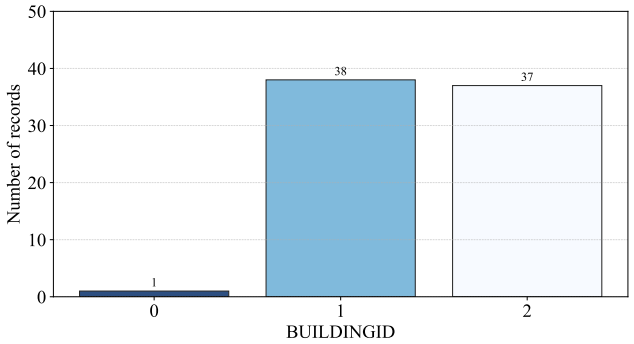


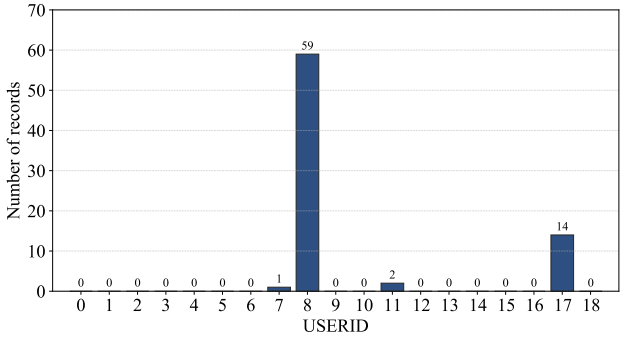
Fig. 16: Distributions of the number of detected WAPs per record.

Note that the left-most bar of the histogram of Fig. 16 indicates that there are few records with few or no detected WAPs. As the records with no detected WAPs, namely *WAP-absent records*, can impair the performance of indoor localization by mapping the same fingerprint of RSSIs of non-detected WAPs[2] to multiple locations, it is imperative to investigate WAP-absent records. Fig. 17 shows the distribution of WAP-absent records of the training dataset per building, user, and phone. From Fig. 17 (a), we observe that Building 0 has significantly fewer WAP-absent records than Building 1 and

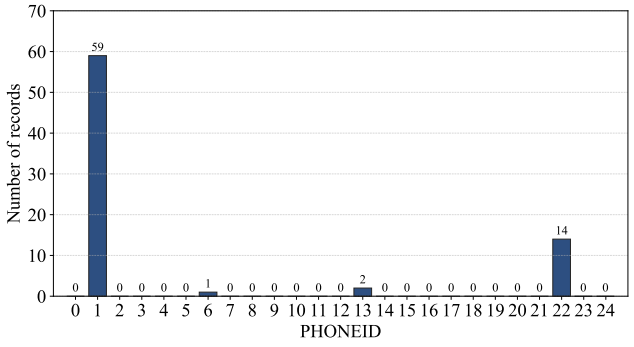
[2]The RSSIs of non-detected WAPs of the UJIIndoorLoc database are typically set to -110 before training a localization model [13].



(a)



(b)



(c)

Fig. 17: Distribution of WAP-absent records of the training dataset per (a) building, (b) user, and (c) phone.

2, while Fig. 17 (b) and (c) show that USER 8 and 17 and PHONE 1 and 22 are mostly associated with WAP-absent records.

Note that the validation dataset does not contain WAP-absent records, which could be explained as follows:

- PHONE 1 and 22 were not used for the validation dataset. The WAP-absent records, therefore, could be related with the sensitivity of specific phones as shown in Fig. 17.
- The number of records of the validation dataset is significantly lower than that of the training dataset, so there were fewer chances of collecting WAP-absent records during the construction of the validation dataset.
- As the records of the validation dataset were collected three months later than those of the training dataset, the environment had changed and/or some WAPs had been removed or gone down.

We also investigate how many WAPs are detected at more than one building, which could impair the performance of indoor localization as well. The results are summarized in Table III, where L_A denotes the number of WAPs detected at Building i for $i \in A$. We can see that the number of WAPs

TABLE III: The number of WAPs detected at building(s).

Dataset	Building	Number of detected WAPs		
Training	0	L_0=200	$L_{0,1}$=59	
	1	L_1=207	$L_{1,2}$=82	$L_{0,1,2}$=3
	2	L_2=203	$L_{0,2}$=7	
Validation	0	L_0=183	$L_{0,1}$=46	
	1	L_1=170	$L_{1,2}$=65	$L_{0,1,2}$=2
	2	L_2=125	$L_{0,2}$=3	

detected at Building 1 and 2 is higher than that at Building 0 and 1, which is understandable given the shorter distance between Building 1 and 2 as shown in Fig. 6 (a). The existence of WAPs detected at Building 0 and 2 and all three buildings, however, is unexpected, and their effects on localization and related pre-processing would be interesting topics for further investigation.

In addition to the building-level analyses, we carry out an RP-level analysis of the RSSI samples measured at the same RPs based on the sample Pearson correlation coefficient (PCC), which measures the linear correlation between two samples [28], [29]. For each pair of the RSSI samples of $X=[x_1, \ldots, x_{520}]$ and $Y=[y_1, \ldots, y_{520}]$ measured at an RP, their sample PCC is given by

$$r_{X,Y} = \frac{\sum_{i=1}^{520}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{520}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{520}(y_i - \bar{y})^2}}, \quad (1)$$

where

$$\bar{x} = \sum_{i=1}^{520} x_i \quad \text{and} \quad \bar{y} = \sum_{i=1}^{520} y_i. \quad (2)$$

$r_{X,Y}$ ranges from -1 to 1, where 1 indicates a positive correlation, 0 indicates no correlation, and -1 indicates a negative correlation. Fig. 18 shows the histogram of the sample PCCs for all possible pairs of the RSSI samples measured at the same RPs in the training dataset, which indicates strong positive correlations for most pairs. There are still few pairs with very low correlations; samples common to those pairs could impair the localization performance of a model and, therefore, be filtered out as outliers.

D. Preliminary Experimental Results

To demonstrate potential benefits of the analyses presented in Sections III-A to III-C, we carry out preliminary experiments based on them. All the experiments were run on a workstation with an Intel Core i9-9900X CPU, 128 GB RAM, and two Nvidia GeForce RTX 2080Ti GPUs running Ubuntu 20.04.2 LTS, and all the models are implemented using Python 3.8.5. As the metric of the indoor localization performance, we use the EVAAL 3D error [19].

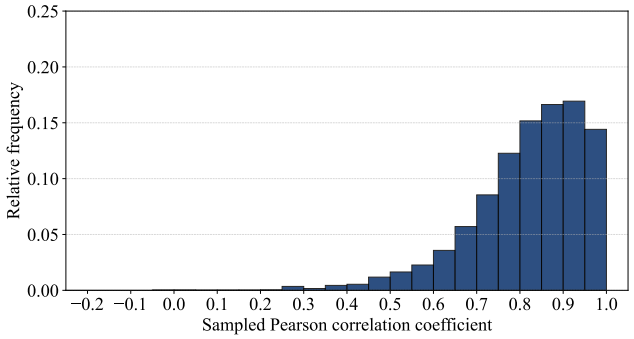


Fig. 18: Histogram of the sample PCCs for all possible pairs of the RSSI samples measured at the same RPs in the training dataset.

First, we assess the effects of WAP-absent records on indoor localization performance using the hierarchical RNN model of [17], whose results are shown in Fig. 19. As discussed in

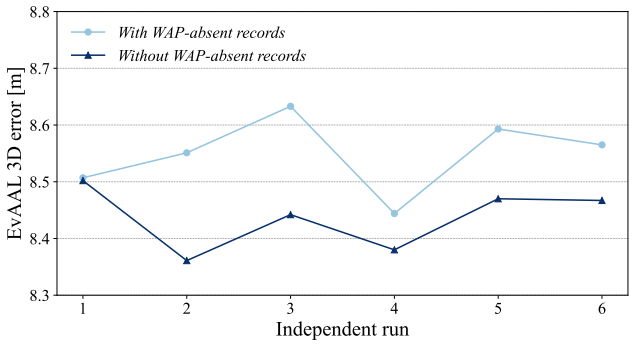


Fig. 19: Mean 3D error of the hierarchical RNN [17].

Section III-C, the results clearly indicate that we can improve indoor localization performance by filtering out the WAP-absent records from the database.

Second, we consider more complicated pre-filtering to exclude WAPs not providing unique information. Specifically, we filter out the following WAPs:

- WAPs with less than 3 unique RSSI values.
- WAPs with RSSI standard deviations less than 1.

After we apply the proposed pre-filtering to the training dataset, we also exclude in the validation dataset the WAPs that are already pre-filtered from the training dataset. As a result, the total number of WAPs of both training and validation datasets is reduced from 520 to 416, which also significantly reduces the training dataset size from 51,518 KB to 33,640 KB.

To investigate the effects of the pre-filtering process, we evaluate the performance of floor classification using the conventional ML algorithms of KNN [30], C-support vector classification (SVC) [31], and C5.0 decision tree algorithm [32] as in [33]. Fig. 20 shows the confusion matrices with and without pre-filtering, where we can observe that the filtering improves the classification performance slightly over all the algorithms. Note that, however, the major benefit of the filtering is the significant reduction of training time from 5 min to 3 min, i.e., by 40%. We also evaluate the performance of indoor localization using the hierarchical RNN [17] and summarize its results in Table IV, which demonstrate that the pre-filtering reduces 3D error but decreases floor hit rate.

TABLE IV: Localization performance of the hierarchical RNN [17] with and without pre-filtering.

	EvAAL 3D error [m]	Floor hit rate [%]
Without pre-filtering	8.62	95.2
With pre-filtering	8.51	92.0

Finally, we investigate the effects of the representation of missing RSSIs (i.e., NaN) in the database by evaluating the localization performance with different numerical values. Table V summarizes the average floor hit rate of the floor classification based on conventional ML algorithms, and Table VI provides the 2D error and success rate of the hierarchical RNN [17], where the success rate is the percentage of successful classification of both building and floor.

TABLE V: Average floor hit rate of the floor classification based on SVC, KNN, and C5.0 with different numerical values for missing RSSIs.

NaN Representation	Average floor hit rate [%]		
	SVC	KNN	C5.0
100	82.0	79.5	78.9
-105	86.3	**85.3**	81.0
-110	**86.8**	85.0	**81.4**

TABLE VI: Localization performance of the hierarchical RNN [17] with different numerical values for missing RSSIs.

Representation	2D error [m]	Success rate [%]
100	10.949	79.6
-105	**8.274**	92.4
-110	8.312	**93.1**

The three numerical values for missing RSSIs are selected based on the following rationales:

- 100 is the original representation used in the UJIIndoorLoc database.
- -105 is consistent with the minimal value of RSSIs, i.e., -104.
- -110 is the most frequently used in the literature for research based on the UJIIndoorLoc database.

The results from Tables V and VI indicate that there is room for improvement in numerical representation of missing RSSIs.

IV. Challenges and Recommendations

Sections III reveals the issues of the fingerprints in the UJIIndoorLoc database in their spatial coverage, measurement practices, and lack of diversity in users and phones, which could have negative effects on the performance of a localization model trained with the database. Here we provide our recommendations on the challenges in using existing fingerprint databases and constructing new ones for large-scale multi-building and multi-floor indoor localization.

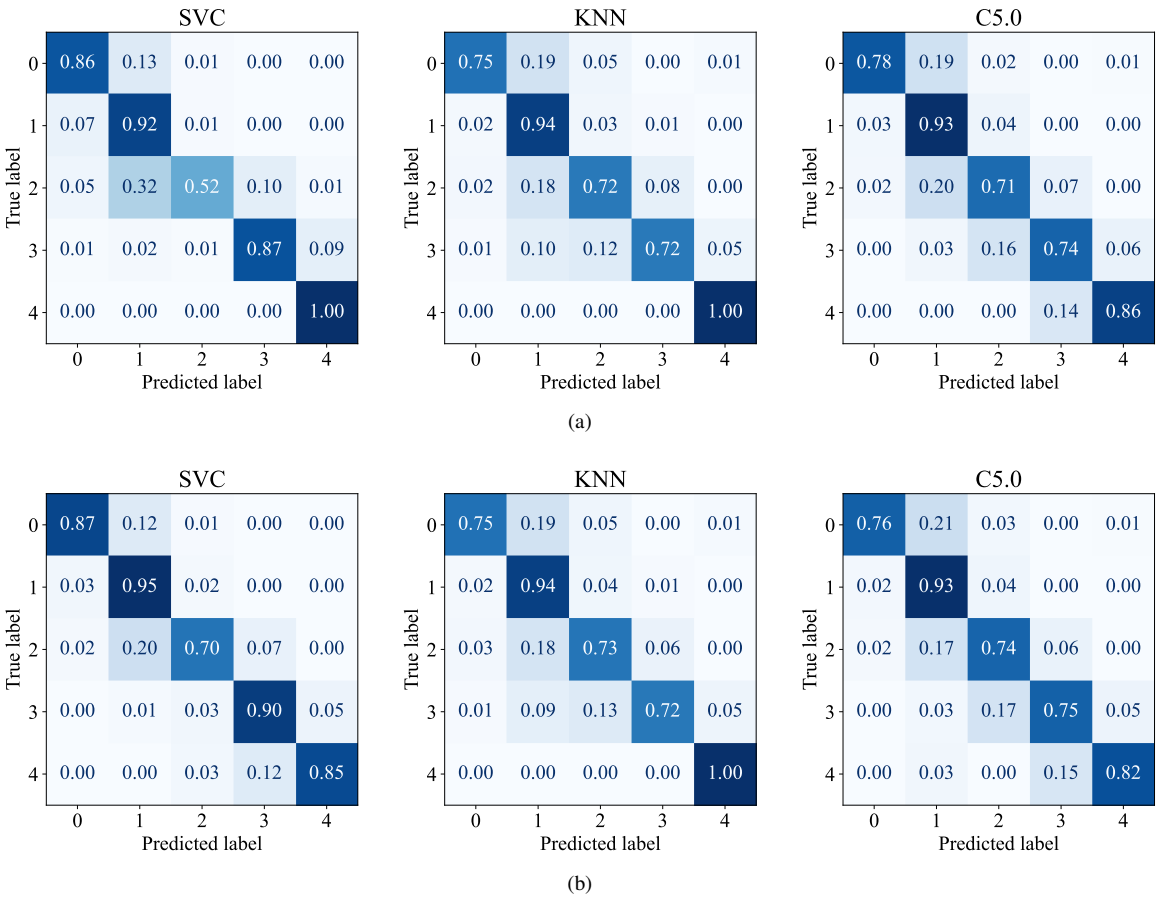


Fig. 20: Confusion matrices of the floor classification based on SVC, KNN, and C5.0: (a) Without WAP-wise feature filtering and (b) with WAP-wise feature filtering.

A. *On the Use of Existing Fingerprint Databases*

1) Localization Algorithms and Models: The accuracy of localization has been a top priority in studying localization algorithms and models, often at the expense of their space and time complexity, because the underlying scenario is that both training of a model during the offline phase and using a trained model for location estimation during the online phase are done at a centralized server with plenty of computing and power resources; under this scenario, a user device just reports the RSSIs measured at an unknown location to and receives estimated location information from the server.

Achieving good localization accuracy is relatively straightforward with a fingerprint database constructed for a controlled small-scale environment over a short period of time, which, however, is challenging with a fingerprint database constructed for a large-scale multi-building and multi-floor environment over a long period of time. In the case of the UJIIndoorLoc database, the validation dataset was collected three months after the training dataset with fewer users and devices, resulting in fewer detected WAPs. To address the issues resulting from the differences between training and validation/testing datasets in large-scale multi-building and multi-floor indoor localization, therefore, techniques like transfer learning [34] and domain adaptation [35] could be applied for the improvement of the generalization and robustness of a model.

In addition to the accuracy and robustness of localization, user privacy is also an important factor. The conventional scenario based on the client-server model raises privacy concerns due to the collection of user information (i.e., location fingerprints) by a centralized server, especially during the online phase. As the computational power of user devices increase, it is possible to perform localization tasks directly on user devices (e.g., based on pretrained models downloaded from a server) without submitting user information to a server, which can protect users' privacy.

As for response time, smaller models have an advantage over larger ones, which can also provide reasonable localization performance in small-scale indoor localization, but their direct application to large-scale indoor localization cannot guarantee the same level of performance. One promising solution in this regard is knowledge distillation [36], which can compress large models without much sacrificing their performance. This process has garnered much success in many research areas but not in indoor localization yet.

Therefore, it is essential to strike a right balance among accuracy, robustness, response time, and privacy in studying localization algorithms and models by taking into account the following important aspects:

- The number of records required for model training.
- The model size and computational power required for

localization.

- The amount of data collected from users and stored on the device.
- The structure and architecture of an indoor localization system (e.g., cloud-based and on-device).

2) Data Balancing: The spatial complexity of building structures and the accessibility of RPs pose challenges to large-scale indoor localization, which lead to uneven spatial distribution of records. As discussed in Section III-A, the UJIIndoorLoc database shows significant imbalance in the number of records (e.g., between west and east corridors of Building 2). The UJIIndoorLoc database also shows the dominance of a couple of users and phones during the construction of both datasets. Such data imbalance in space, user, and device distributions could result in poor and biased training results, so data balancing during the construction of Wi-Fi fingerprint databases is essential to achieving unbiased results as well as good localization performance from trained models.

To handle data imbalance, we can apply data augmentation and/or data resampling techniques to existing fingerprint databases: As for data augmentation, straightforward application of conventional techniques (e.g., [25], [37]) could provide satisfactory results when original records already have a good space coverage like those of Building 0 shown in Fig. 6 (b). When a building structure is complicated and original records poorly cover the space like the top floor of Building 2 or the bottom two floors of Building 1 of the UJIIndoorLoc database, however, we need a more sophisticated data augmentation schemes like generative adversarial network (GAN)-based ones [38].

As for data resampling, we can apply advanced data resampling techniques like stratified sampling [39] and weighted random sampling [40] as well as conventional up- and down-sampling to obtain more evenly-distributed datasets. For example, we can apply weighted random sampling to Floor 1 and 2 of Building 2 of the UJIIndoorLoc database.

3) Data Preprocessing: Based on the comprehensive analyses in Section III-B, it has been observed that certain WAPs are not playing a critical role in accurately identifying a specific building or floor. The significance of each WAP can vary depending on the particular building and floor. For instance, during data augmentation tasks, certain WAPs within the range of [20, 40] in the UJIIndoorLoc training dataset are detected in Building 0 but have negligible relevance for Building 2 as shown in Fig. 14. Therefore, when selecting inputs for indoor localization models and complex data augmentation algorithms, it is crucial to consider the importance of each WAP at different levels.

For instance, leveraging the hierarchical nature of multi-building and multi-floor indoor localization, we can apply the following strategies to limit the number of input dimensions by a step in manner:

As illustrated in Fig. 21, once the building-level classification model is established, for floor classification, we can use only the WAPs that are important for the given building as inputs. Similarly, for coordinate-level localization, we can use only the WAPs that are important for the given floor as inputs. This hierarchical data evolution training framework can

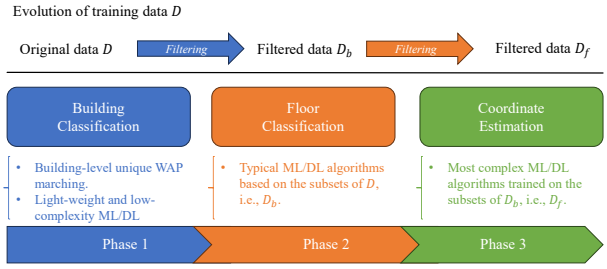


Fig. 21: Hierarchical and hybrid/fusion data evolution training framework. Note that the filtering operations in data evolution are performed based on both feature-wise and record-wise analyses.

significantly reduce the input dimension of indoor localization models and the cost of retraining models when the database is updated. However, the downside of this framework, illustrated in 22, is that it requires the construction of a series of separate models for each level of localization, which could be a challenge in terms of computational resources and time.

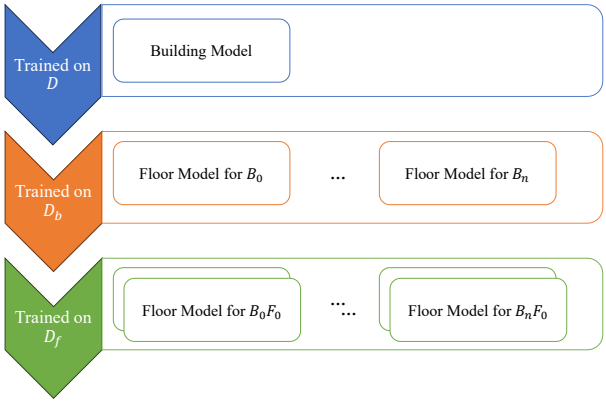


Fig. 22: Downside of the hierarchical data evolution training framework, multiple models are required for each level of localization.

It is also worth to investigate pre-filtering of WAPs with all-NaN values in the database. In the case of the UJIIndoorLoc database, as discussed in Section III-B, Out of 520 WAPs, 55 and 153 WAPs were never detected in the training and validation datasets, respectively. Excluding these WAPs could substantially reduce the input dimension of indoor localization models and storage requirements for the database, particularly for large-scale scenarios. There have been proposed several WAP selection schemes [41]–[44], but their application to large-scale multi-building and multi-floor indoor localization databases are to be carefully investigated.

WAP-absent records in the database—i.e., resulting from the lack of WAPs around the user or the device's inability to detect WAPs at the time of RSSI measurement—can deteriorate indoor localization performance and cause issues like the cold start problem [45], a decrease in robustness, and poor generalization. In the case of the UJIIndoorLoc database, users with PHONE 1 or 22 are more likely to generate WAP-absent records on the top floor of Building 1 as discussed

in Section III-C. To avoid such issues, we can simply filter out WAP-absent records before training a model. We can also apply more advanced techniques like data imputation to handle such missing data [46].

4) Missing Value Representation: It is important to understand that different databases may represent NaN with different numerical values. For example, the UJIIndoorLoc database labels NaN values as 100, which is typically converted to -110 during the preprocessing. While a larger variation in RSSI values for different locations could help room or floor classifications, it is unsuitable for coordinate regression as demonstrated in Tables V and VI of Section III-D. Therefore, it is important to numerically represent missing RSSIs based on the statistics of RSSI values in the database, which can ensure accurate indoor localization for both classification and regression.

B. On the Construction of New Fingerprint Databases

1) RP Selection: It would be ideal if we could tackle the issue of uneven record distributions during the construction of a database through a more systematic way of RP selection, because imbalanced data may cause issues like bias and overfitting in classification [47]. Proper selection of RPs during the construction of a database, however, is challenging in that the following, seemingly-contradicting requirements should be met simultaneously:

- RPs should cover entire monitoring areas in a balanced way.
- RPs should be easily accessible for data collection.

As discussed in Section III-A1, the UJIIndoorLoc database shows both building- and floor-level uneven record distributions due to the accessibility of the restricted spaces like chemical laboratories and private offices [18].

We can apply *post-construction* techniques as practical alternatives in this regard: For example, we can utilize spatial data augmentation [25] for Building 0 and 1 and stratified sampling [48] for Building 2. As a smaller number of RPs with balanced space coverage is better than a larger number of RPs with poor space coverage even for the application of the post-construction techniques, the selection of RPs was and remains a top priority for the construction of fingerprint databases.

2) Measurement Practices: How to measure fingerprints based on the selected RPs is another important factor for not only the construction of a fingerprint database but also its maintenance and update.

The frequency of RSSI measurements (e.g., daily or weekly), together with the total period of database construction (e.g., over a month or a year), needs to be carefully determined. With a low measurement frequency, some WAPs may not be detected during the measurements due to their being in a standby or sleeping operation mode, while a high measurement frequency incurs a higher labor cost. Note that the effects of a low frequency of RSSI measurements could be compensated for using a longer period of database construction. As environment changes like people's presence and the use of various electronic devices significantly affect the measured RSSIs, using different times of the day for measurements (e.g., during or after work hours) is also required to increase the temporal diversity of a fingerprint database.

Together with the measurement frequency and time and the database construction period, the way of visiting RPs is important, too. For example, we can visit and measure RSSIs at the same RPs repeatedly over a construction period for small-scale indoor localization, while we can divide RPs and visit a part of them during each measurement for large-scale multi-building and multi-floor indoor localization.

Though we carefully plan RSSI measurements at properly selected RPs for the construction of a fingerprint database, it will be useless if we do not have enough human resource to carry out the plan. As discussed in Section II-E, a hybrid data collection strategy combining in/outsourcing with volunteering can be adopted to reduce the high labor cost. Even with such a hybrid data collection strategy, a core group of participants is still a key to the successful construction of a good fingerprint database, who is to provide high-quality RSSI measurements by strictly following the measurement plan.

3) User, Device, and Network Diversity: To increase the diversity of collected fingerprint data for the robustness of a trained model, the measurements should be done with multiple users of different physical characteristics (e.g., height) with different models of devices. For example, depending on a couple of devices could result in lots of WAP-absent records due to the special characteristics or even errors of certain devices as discussed in Section III-C.

To further increase the diversity of location fingerprints and thereby provide more robust localization service in various indoor environments, we can collect and provide other types of location fingerprints like RSSIs of BLE [24] and cellular networks [49] and geomagnetic field intensity [50] as well as Wi-Fi RSSIs.

4) Database Maintenance and Update: Once an indoor localization system is deployed in the field with a constructed fingerprint database, there will be increasing requirements for the maintenance and update of the fingerprint databases, the activities of which include addition of fingerprints from new WAPs, replacement of fingerprints from existing WAPs, and removal of fingerprints no longer relevant.

In addition to manual collection of fingerprints by human participants, automatic collection of fingerprints using lightweight, battery-powered anchor devices deployed at selected RPs could be considered to further reduce the labor cost [51]. Also, exploiting unlabeled RSSIs submitted by users during the online phase of an indoor localization system deployed in the field is another interesting option in this regard [52].

V. Conclusions

Wi-Fi fingerprinting has become a dominant technology for indoor localization due to its major advantage of usability in any environment equipped with Wi-Fi networks without requiring additional hardware or infrastructure. As the localization performance of Wi-Fi fingerprinting heavily depends on the quality of the underlying fingerprint database used to

train an ML model for location estimation, a study on the use and construction of fingerprint databases becomes as important as that on localization algorithms and models, whose major focus, however, has been limited to databases covering a single floor or building.

The UJIIndoorLoc Wi-Fi fingerprint database represents a significant advancement in the field of multi-building and multi-floor indoor localization. While many researchers have used this database as a benchmark to evaluate the performance of their proposed algorithms and models, there have been few studies dedicated to multi-building and multi-floor fingerprint databases. This paper aims to fill this gap by examining the UJIIndoorLoc database, which is by far the most popular multi-building and multi-floor Wi-Fi fingerprint database, and providing practical guidance on the use of existing databases and future directions for the design and construction of new databases.

As a basis, we have carried out a comprehensive case study of the UJIIndoorLoc database, where we investigate the statistical characteristics of the UJIIndoorLoc database through both field-level and record-level analyses. We have obtained several key insights on the UJIIndoorLoc database through those analyses, i.e., (1) the uneven spatial distributions of records, (2) the lack of user and phone diversity during the measurements, (3) the existence of WAP-absent records, and (4) the identification of WAPs not providing unique information for fingerprinting; especially, we have assessed and demonstrated the effects of WAP-absent records and WAPs not providing unique information on indoor localization performance through preliminary experiments.

Based on the results of the case study with the UJIIndoorLoc database, we have provided our recommendations on the challenges in using existing fingerprint databases and constructing new ones for large-scale multi-building and multi-floor indoor localization, where we discuss in detail potential application of advanced ML techniques like data augmentation, data resampling, data imputation, and semi-supervised learning in addressing the challenges.

To the best of the authors' knowledge, this is the first work to extensively analyze the UJIIndoorLoc database, identify its issues, and provide recommendations on the various challenges in using the existing database and constructing new ones for large-scale multi-building and multi-floor indoor localization. Our findings presented in this paper serve as a valuable starting point to researchers new to this field and provide a practical guidance to those interested in using the UJIIndoorLoc database or creating their own large-scale Wi-Fi fingerprint databases. Note that we have been constructing our own large-scale multi-building and multi-floor Wi-Fi fingerprint database based on the results of this case study, whose preliminary results are presented in [51].

Acknowledgment

This work was supported in part by the Postgraduate Research Scholarships (under Grant PGRS1912001), the Key Program Special Fund (under Grant KSF-E-25), and the Research Enhancement Fund (under Grant REF-19-01-03) of Xi'an Jiaotong-Liverpool University.

References

[1] F. Liu, J. Liu, Y. Yin, W. Wang, D. Hu, P. Chen, and Q. Niu, "Survey on WiFi-based indoor positioning techniques," *IET Communications*, vol. 14, no. 9, pp. 1372–1383, 2020.

[2] S. He and S.-H. G. Chan, "Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 466–490, 2016.

[3] S. A. Golden and S. S. Bateman, "Sensor measurements for Wi-Fi location with emphasis on time-of-arrival ranging," *IEEE Trans. Mobile Comput.*, vol. 6, no. 10, pp. 1185–1198, 2007.

[4] X. Chen, S. Song, and J. Xing, "A ToA/IMU indoor positioning system by extended kalman filter, particle filter and MAP algorithms," in *Proc. PIMRC 2016*, 2016, pp. 1–7.

[5] H. Nawaz, A. Bozkurt, and I. Tekin, "A novel power efficient asynchronous time difference of arrival indoor localization system using CC1101 radio transceivers," *Microwave and Optical Technology Letters*, vol. 59, no. 3, pp. 550–555, 2017.

[6] R. Kumarasiri, K. Shamaileh, N. Tran, and V. Devabhaktuni, "An improved hybrid RSS/TDOA wireless sensors localization technique utilizing Wi-Fi networks," *Mobile Networks and Applications*, vol. 21, 06 2015.

[7] F. Wen and C. Liang, "An indoor AOA estimation algorithm for IEEE 802.11ac Wi-Fi signal using single access point," *IEEE Commun. Lett.*, vol. 18, no. 12, pp. 2197–2200, 2014.

[8] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system," in *Proc. NSDI'13*, Lombard, IL, USA, 2013, pp. 71–84.

[9] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A survey on wireless indoor localization from the device perspective," *ACM Comput. Surv.*, vol. 49, no. 2, Jun. 2016.

[10] L. Zhang, E. Ding, Y. Hu, and Y. Liu, "A novel CSI-based fingerprinting for localization with a single AP," *EURASIP Journal on Wireless Communications and Networking*, 2019.

[11] J. Torres-Sospedra, P. Richter, A. Moreira, G. M. Mendoza-Silva, E. S. Lohan, S. Trilles, M. Matey-Sanz, and J. Huerta, "A comprehensive and reproducible comparison of clustering and optimization rules in Wi-Fi fingerprinting," *IEEE Trans. Mobile Comput.*, vol. 21, no. 3, pp. 769–782, 2022.

[12] B. Wang, X. Gan, X. Liu, B. Yu, R. Jia, L. Huang, and H. Jia, "A novel weighted KNN algorithm based on RSS similarity and position distance for Wi-Fi fingerprint positioning," *IEEE Access*, vol. 8, pp. 30 591–30 602, 2020.

[13] K. S. Kim, S. Lee, and K. Huang, "A scalable deep neural network architecture for multi-building and multi-floor indoor localization based on WiFi fingerprinting," *Big Data Analytics*, vol. 3, no. 1, Apr 2018.

[14] J. Cha and E. Lim, "A hierarchical auxiliary deep neural network architecture for large-scale indoor localization based on Wi-Fi fingerprinting," *Applied Soft Computing*, vol. 120, 2022.

[15] X. Song, X. Fan, X. He, C. Xiang, Q. Ye, X. Huang, G. Fang, L. L. Chen, J. Qin, and Z. Wang, "CNNLoc: Deep-learning based indoor localization with WiFi fingerprinting," in *Proc. 2019 SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, 2019, pp. 589–595.

[16] M. Ibrahim, M. Torki, and M. ElNainay, "CNN based indoor localization using RSS time-series," in *Proc. ISCC 2018*, 2018, pp. 01 044–01 049.

[17] A. E. Ahmed Elesawi and K. S. Kim, "Hierarchical multi-building and multi-floor indoor localization based on recurrent neural networks," in *Proc. CANDARW 2021*, 2021, pp. 193–196.

[18] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. IPIN 2014*, 2014, pp. 261–270.

[19] A. Moreira, M. J. a. Nicolau, F. Meneses, and A. Costa, "WiFi fingerprinting in the real world - RTLS@UM at the EvAAL competition," in *Proc. IPIN 2015*, 2015, pp. 1–10.

[20] J. Niu, B. Wang, L. Cheng, and J. J. P. C. Rodrigues, "WicLoc: An indoor localization system based on WiFi fingerprints and crowdsourcing," in *Proc. ICC 2015*, 2015, pp. 3008–3013.

[21] E. S. Lohan, J. Torres-Sospedra, H. Leppäkoski, P. Richter, Z. Peng, and J. Huerta, "Wi-Fi crowdsourced fingerprinting dataset for indoor positioning," *Data*, vol. 2, no. 4, 2017.

[22] P. Richter, E. S. Lohan, and J. Talvitie, "WLAN (WiFi) RSS database for fingerprinting positioning," Jan 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1161525

[23] M. Mohammadi, A. Al-Fuqaha, M. Guizani, and J.-S. Oh, "Semisupervised deep reinforcement learning in support of IoT and smart city services," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 624–635, 2018.

[24] Apple iBeacon. Accessed: February 22, 2023. [Online]. Available: https://developer.apple.com/ibeacon/

[25] Z. Tang, S. Li, K. S. Kim, and J. S. Smith, "On the multi-dimensional augmentation of fingerprint data for indoor localization in a large-scale building complex based on multi-output Gaussian processes," *Sensors*, vol. 24, no. 3:1026, pp. 1–19, Feb. 2024.

[26] S. Sharma, J. Singh, and A. Gosain, "Experimental analysis of over-sampling techniques in class imbalance problem," in *Evolution in Computational Intelligence*, V. Bhateja, X.-S. Yang, J. C.-W. Lin, and R. Das, Eds. Singapore: Springer Nature Singapore, 2023, pp. 415–429.

[27] "Understanding RSSI." [Online]. Available: https://www.metageek.com/training/resources/understanding-rssi/

[28] W. Kirch, Ed., *Pearson's Correlation Coefficient*. Dordrecht: Springer Netherlands, 2008, pp. 1090–1091.

[29] Wikipedia contributors, "Pearson correlation coefficient — Wikipedia, the free encyclopedia," 2024, [Online; accessed 13-February-2024]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1195564869

[30] sci-kit learn: K-nearest neighbors. Accessed: November 22, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn-neighbors-kneighborsclassifier

[31] sci-kit learn: C-support vector classification. Accessed: November 22, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

[32] C50: C5.0 decision trees. Accessed: November 22, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn-tree-decisiontreeclassifier

[33] AI MAVERICK. (2021) UJI IoT analytics. Accessed: February 8, 2024. [Online]. Available: https://www.kaggle.com/code/samanemami/iot-analytics

[34] İ. Onat Korkmaz, T. Özateş, E. Koç, E. Aydın, E. Kor, D. Dilek, M. Alp Güngen, İ. Gökalp Köse, and Ç. Akman, "Indoor localization with transfer learning," in *Proc. SIU 2022*, 2022, pp. 1–4.

[35] X. Chen, H. Li, C. Zhou, X. Liu, D. Wu, and G. Dudek, "Fidora: Robust WiFi-based indoor localization via unsupervised domain adaptation," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9872–9888, 2022.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv e-prints*, 2015, arXiv:1503.02531 [stat.ML].

[37] R. S. Sinha and S.-H. Hwang, "Improved RSSI-Based data augmentation technique for fingerprint indoor localisation," *Electronics*, vol. 9, no. 5, 2020.

[38] W. Njima, M. Chafii, A. Chorti, R. M. Shubair, and H. V. Poor, "Indoor localization using data augmentation via selective generative adversarial networks," *IEEE Access*, vol. 9, pp. 98 337–98 347, 2021.

[39] M. Shahrokh Esfahani and E. R. Dougherty, "Effect of separate sampling on classification accuracy," *Bioinformatics*, vol. 30, no. 2, pp. 242–250, 11 2013.

[40] P. Efraimidis and P. Spirakis, *Weighted Random Sampling*. Boston, MA: Springer US, 2008, pp. 1024–1027.

[41] D. Quezada-Gaibor, L. Klus, J. Torres-Sospedra, E. S. Lohan, J. Nurmi, C. Granell, and J. Huerta, "Data cleansing for indoor positioning Wi-Fi fingerprinting datasets," in *Proc. MDM 2022*, 2022, pp. 349–354.

[42] S. Eisa, J. Peixoto, F. Meneses, and A. Moreira, "Removing useless APs and fingerprints from WiFi indoor positioning radio maps," in *Proc. IPIN 2013*, 2013, pp. 1–7.

[43] L. Klus, D. Quezada-Gaibor, J. Torres-Sospedra, E. S. Lohan, C. Granell, and J. Nurmi, "RSS fingerprinting dataset size reduction using feature-wise adaptive k-means clustering," in *Proc. ICUMT 2020*, 2020, pp. 195–200.

[44] J. Talvitie, M. Renfors, M. Valkama, and E. S. Lohan, "Method and analysis of spectrally compressed radio images for mobile-centric indoor localization," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 845–858, 2018.

[45] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, *Collaborative Filtering Recommender Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 291–324.

[46] X. Li, H. Li, H. K.-H. Chan, H. Lu, and C. S. Jensen, "Data imputation for sparse radio maps in indoor positioning," in *Proc. ICDE 2023*, 2023, pp. 2235–2248.

[47] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog Artif Intell*, vol. 5, pp. 221–232, 2016.

[48] H. Zheng, M. Gao, Z. Chen, X.-Y. Liu, and X. Feng, "An adaptive sampling scheme via approximate volume sampling for fingerprint-based indoor localization," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2338–2353, 2019.

[49] H. Rizk, M. Torki, and M. Youssef, "CellinDeep: Robust and accurate cellular-based indoor localization via deep learning," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2305–2312, 2019.

[50] J. Torres-Sospedra, D. Rambla, R. Montoliu, O. Belmonte, and J. Huerta, "UJIIndoorLoc-Mag: A new database for magnetic field-based localization problems," in *Proc. IPIN 2015*, Banff, Alberta, Canada, 2015, pp. 1–10.

[51] Z. Tang, R. Gu, S. Li, K. S. Kim, and J. S. Smith, "Static vs. dynamic databases for indoor localization based on Wi-Fi fingerprinting: A discussion from a data perspective," *ArXiv e-prints*, Feb. 2024, arXiv:2402.12756 [cs.LG].

[52] S. Li, Z. Tang, K. S. Kim, and J. S. Smith, "Exploiting unlabeled RSSI fingerprints in multi-building and multi-floor indoor localization through deep semi-supervised learning based on mean teacher," in *Proc. CANDAR 2023*, Matsue, Japan, Nov.–Dec. 2023, pp. 155–160.