Bi *et al. Satellite Navigation* (2022) 3:25
https://doi.org/10.1186/s43020-022-00086-y

Satellite Navigation
https://satellite-navigation.springeropen.com/

ORIGINAL ARTICLE Open Access

Supplementary open dataset for WiFi indoor localization based on received signal strength

Jingxue Bi[1], Yunjia Wang[2], Baoguo Yu[3], Hongji Cao[1*], Tongguang Shi[1] and Lu Huang[3]

Abstract

Several Wireless Fidelity (WiFi) fingerprint datasets based on Received Signal Strength (RSS) have been shared for indoor localization. However, they can't meet all the demands of WiFi RSS-based localization. A supplementary open dataset for WiFi indoor localization based on RSS, called as SODIndoorLoc, covering three buildings with multiple floors, is presented in this work. The dataset includes dense and uniformly distributed Reference Points (RPs) with the average distance between two adjacent RPs smaller than 1.2 m. Besides, the locations and channel information of pre-installed Access Points (APs) are summarized in the SODIndoorLoc. In addition, computer-aided design drawings of each floor are provided. The SODIndoorLoc supplies nine training and five testing sheets. Four standard machine learning algorithms and their variants (eight in total) are explored to evaluate positioning accuracy, and the best average positioning accuracy is about 2.3 m. Therefore, the SODIndoorLoc can be treated as a supplement to UJIIndoorLoc with a consistent format. The dataset can be used for clustering, classification, and regression to compare the performance of different indoor positioning applications based on WiFi RSS values, e.g., high-precision positioning, building, floor recognition, fine-grained scene identification, range model simulation, and rapid dataset construction.

Keywords: WiFi, Indoor localization, Open dataset, RSS, AP, Machine learning

Introduction

The indoor positioning has thrived for more than twenty years, and various technologies and methods have emerged to meet the requirements of location-based services in Global Navigation Satellite System (GNSS) denied environments. According to the principles of indoor positioning technologies, they can be divided into wireless signal positioning (Alvarez-Merino et al., 2021; Chen et al., 2021; Li et al., 2020; Poulose et al., 2020; Ye et al., 2022), inertial navigation (Feng et al., 2020; Liu et al., 2021), computer visual positioning (Maheepala et al., 2020; Morar et al., 2020), and others (Huang et al., 2022; Kunhoth et al., 2020; Li et al., 2016; Ruiz et al., 2011; Xu et al., 2021). Wireless Fidelity (WiFi) plays a vital role in wireless signal positioning technologies because of its wide applications and high commercial value (Liu et al., 2020; Zhuang et al., 2015).

Received Signal Strength (RSS) (Poulose et al., 2020; Tao & Zhao, 2021; Torres-Sospedra et al., 2014), Channel State Information (CSI) (Gönültaş et al., 2021; Rocamora et al., 2020; Tian et al. 2020), and Round Trip Time (RTT) (Cao et al., 2020; Guo et al., 2019) can be extracted from WiFi signals for fingerprint-based, range-based, and angle-based indoor localization. Due to the advantages of their diversity and easy access, WiFi indoor positioning methods, especially RSS-based, have attracted a great attention. RSS is the superposition of multipath signals at the same time. It is a kind of coarse-grained data. It is simple and easily accessible, while RTT and CSI are fine-grained information that requires specialized equipment. Although RTT and CSI can achieve better positioning accuracy than RSS, they are still in the laboratory stage or limited by some devices, so they can't be widely promoted and applied at present.

*Correspondence: caohongji22@sdjzu.edu.cn

[1] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China
Full list of author information is available at the end of the article

To improve network performance, security, and battery life, the Android Operating System (OS) restricted the permissions and the frequency of WiFi scans from Android 8.0, i.e. Application Programming Interface (API) level 26. Foreground application running on Android smartphone can scan four times in two minutes. All background apps can scan one time in thirty minutes. Android 9.0 (API level 28) and higher versions tightened permission requirements limiting the frequency of WiFi scans. It seems that RSS is not available on Android native OS. However, running the developed app on many Android smartphones, we found that 29 phones could collect RSS data normally after granting permission in December 2021. These phone brands are Huawei, Oppo, Vivo, and Xiaomi. The OS of both Oppo and Xiaomi is Android, for Vivo is OriginOS, and for Huawei is HarmonyOS. The details of the smartphones that can be used to collect RSS data are shown in Table 1.

The resources section of the official website of Indoor Positioning and Indoor Navigation (IPIN) listed several collected or crowdsourced WiFi RSS-based datasets in different environments. Lohan summarized twelve WiFi fingerprint datasets and reported the corresponding limitations of these datasets (Lohan et al., 2017). Montoliu et al. (2017) released an IndoorLoc platform to compare and evaluate several kinds of indoor positioning methods by using different datasets. These existing open-source datasets help researchers quickly carry out indoor localization experiments. However, they are not enough to deal with all the RSS-based positioning problems and some datasets are not accessible.

Table 1 The details of the smartphones

Brands	Models	OS
Huawei	Honor 10	HarmonyOS
	Honor 20	
	Honor 30	
	Mate 30	
	Mate 40	
	Nova 4	
	Nova 5	
	Nova 7	
	Nova 8	
	P 30	
Oppo	A9	Android
	A93	
	R15	
Vivo	iQOO	OriginOS
	S1	
	X27	
Xiaomi	Redmi Note 10 Pro	Android
	Redmi K20 Pro	
	Redmi K40	
	Xiaomi 8	
	Xiaomi 10	
	Xiaomi 11	

To the best of our knowledge, sampling points in the existing datasets were sparse, and the distance between two adjacent sampling points was usually too large to achieve high precision positioning. Besides, most of the datasets include only corridor and hall scenes, and a few datasets contain rooms with a small number of sampling points. In addition, all the open-source and available WiFi RSS-based datasets were constructed before September 2017, where most of the wireless Access Points (APs) were single-band, not the dual-band used nowadays, i.e., an AP contains multiple Media Access Control (MAC) addresses. In this case, combining all MAC addresses in one dataset will cause a dimensional disaster. Moreover, APs are usually fixed within the building, whose locations are not provided in the previous datasets.

Therefore, the SODIndoorLoc was created, covering three buildings with multiple floors where there are corridors, office rooms, and meeting rooms. The SODIndoorLoc can be found on the GitHub website (Bi, 2022). And it can be treated as a supplement of UJIIndoorLoc. Expecting to record the same information as UJIIndoorLoc, the locations and channel information of pre-installed APs and Computer-Aided Design (CAD) drawings of each floor are provided. Layouts of the above rooms are also preserved in CAD drawings. Most importantly, the average distance between two adjacent Reference Points (RPs) is less than 1.2 m, which is smaller than those in existing datasets, and the locations of Testing (or validation) Points (TPs) are different from those of RPs. The main characteristics of the dataset are:

- It covers a total area of 8000 m^2, including three buildings with one or three floors.
- 105 APs are pre-installed in three buildings, among which 56 are single-band and 49 are dual-band. The locations and channel information of these APs are summarized.
- 1802 points at different locations are arranged, and the number of RPs and TPs are 1630 and 272, respectively.
- 23,925 samples are recorded, among which 21,205 for training/learning and 2720 for testing/validation.
- The dataset contains three kinds of scenes, office room, meeting room, and corridor. Hall and corridor are seamless in these buildings, so there is no distinction between the two scenes.

- The distance between two adjacent sampling points is about 1.2 m in two buildings with one floor, while about 0.5 m in a three-story building.
- At each RP in the two buildings with one floor, training data has thirty samples. In contrast, training data in the three-story building only contains one sample, a vector of average values throughout sampling time. All the testing data at each TP possesses ten samples.

The main contributions are as follows. We expect that the dataset can become a reference dataset and help researchers delve into WiFi RSS-based indoor localization.

- It is a dataset with dense RPs. The average interval of two adjacent RPs is smaller than 1.2 m, which could be used for high-precision positioning, scene identification, and dataset construction.
- The locations and channel information of pre-installed APs are provided in the dataset. This information can be used for range model simulation, high-precision positioning, and dataset construction.

The rest of the paper is organized as follows: Section heading "Related work" introduces the related work of WiFi RSS-based datasets. Section heading "Description of SODIndoorLoc dataset" describes the presented dataset in detail. Several experiments using the dataset with different machine learning methods are shown in Section heading "Experiments based on the SODIndoorLoc dataset". Discussions and conclusions are given in Section heading "Discussions and conclusions".

Related work

Public WiFi RSS-based datasets have greatly facilitated the development of WiFi indoor localization to some extent. The well-known WiFi RSS-based dataset is UJIIndoorLoc (Torres-Sospedra et al., 2014), covering multiple buildings and floors, which was published in the University of California, Irvine (UCI) machine learning repository in 2014. It is available and utilized as a dataset for competition by IPIN. More than 300 articles cited the UJIIndoorLoc dataset. And thousands of researchers carried out experiments by using the UJIIndoorLoc (Cao et al., 2021; Qin et al., 2021).

The coverage and data of UJIIndoorLoc are too large. Many researchers therefore studied the private dataset by collecting the local WiFi RSS data from small areas (Tao & Zhao, 2021). The IPIN 2016 tutorial provided a dataset named IPIN2016 Tutorial, focusing on the study of a small scenario covering a small corridor with an area of approximately 120 m^2. The dataset consists of 927 training records and 702 testing ones with 177 attributes. As a simplified version of the IPIN2016 Tutorial dataset, the Alcala Tutorial 2017 dataset is published with 670 training records and 405 testing ones with 154 attributes. These two datasets are evaluated by the IndoorLoc platform (Montoliu et al., 2017), and only the training datasets are open source. If you want to obtain the testing datasets, you should contact the authors. The limitation has made many researchers split the training dataset into two parts, one for training and the other for localization (Qin et al., 2021).

During the off-site competition track three (smartphone-based) of IPIN 2016, accelerations, angular velocity, magnetic field strength, pressure, ambient light, orientation, sound level, WiFi RSS, and other information from external devices were logged with a dynamic strategy in four buildings by multiple users and devices. Everyone can download the datasets and supporting materials from the IPIN resource section or the long-term repository Zenodo, such as log files, files of floor maps and the visualization of training routes, evaluation scripts, and some codes for reading and processing these files. The IPIN resource section contains the competition results that Zenodo does not. The Zenodo contains the ground truth of evaluation scripts, while the IPIN resource section does not. The number of MAC addresses detected at different points varies greatly, and few MAC addresses can be detected at many points. The best positioning accuracy of the IPIN2016 competition track three is about 5.85 m based on the 578 evaluation points by integrating all the collected information. The competition score metric is based on the 75th percentile of the point error (Potortì et al., 2022). WiFi plays a minor role in the track three competition, even if IPIN 2017 competition provided the locations of seven MACs, and IPIN 2019 competition recorded WiFi frequency corresponding to the detected MAC addresses.

Before the release of the UJIIndoorLoc dataset, Lohan published a WiFi measurement dataset in the repository of her homepage (Lohan, 2013). The dataset covers two four-floor buildings, and all the measurements were done in 2013. But the format of the dataset is too complex to be conveniently used and not as straightforward as the UJIIndoorLoc. The crowdsourced WiFi dataset (Lohan et al., 2017) covering a four-floor building was released in the Zenodo repository in 2017. The measurements were conducted using a visualized benchmark software in a crowdsourced mode via 21 different devices and users. Therefore, the testing data (3951 records) is about 5.75 times more than the training data (687 records). Unlike other datasets, it provided elevations rather than floor numbers. A supplementary dataset (Richter et al., 2018) was published in 2018 for the crowdsourced WiFi dataset.

An interesting dataset (Moreira et al., 2017) with multiple simultaneous WiFi interfaces was released at the IPIN 2017 conference. Many independent WiFi interfaces simultaneously collected WiFi measurements with a total area of around one thousand square meters. The experimental results showed that positioning accuracy was greatly improved. An enhanced dataset (Torres-Sospedra et al., 2019) based on the above one by exploiting the combinations of complementary sensor's data was displayed at the IPIN 2019 conference with a significant improvement in positioning accuracy. Moreover, some other WiFi RSS-based datasets (Nahrstedt & Vu, 2012; Parasuraman et al., 2016) are published on crawdad and are restricted to download.

Description of SODIndoorLoc dataset

In this section, the proposed dataset is described in detail. Section heading "Description of the testing area" briefly introduces three buildings, CAD drawings of each floor, and layouts. Section heading "Description of data collection" interprets how to collect data. Then, Section heading "Description of data sheets" ultimately shows the information on the records in the dataset.

Description of the testing area

The total indoor area is about 8000 m^2, covering three buildings in different cities. As shown in Fig. 1, a floor plan with a simple layout is overlaid on the satellite image of each building. And all the WiFi collecting areas are rendered in orange.

The CETC331 building has not been used for a long time, and it is an old building with three floors in Shijiazhuang city. The area of the 1st and 2nd floor is about 830 m^2 each, and the area of the 3rd floor is about 140 m^2. The total area is about 1800 m^2. The floor plan in Fig. 1a is for the 2nd floor. The collecting area includes corridor, office, and meeting rooms. There were no APs in the building until we deployed 26 dual-band ones in September 2018, seven APs for the 1st floor, twelve APs for the 2nd floor, and seven APs for the 3rd floor. All APs were fixed to the ceiling. The deployment scheme of APs is supported by using Cramér-Rao lower bound as the metric of positioning error, which could ensure high positioning accuracy in the whole building.

The HCXY building is an office building at a college in Xuzhou city. The floor plan in Fig. 1b is for the 4th floor. The total area of the 4th floor is about 3600 m^2. The corridor with the width of 2.4 m was chosen as the WiFi collecting area in orange color. The length of the corridor is about 211 m. Single-band APs were symmetrically deployed and distributed at a uniform height with the same distance interval. All of them were displayed on both sides of the walls. Because several APs were broken

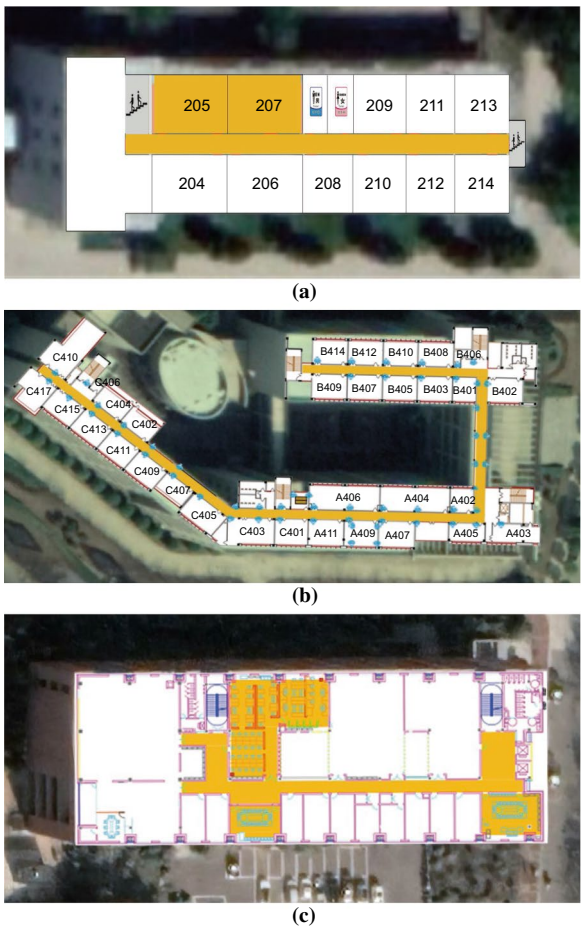

Fig. 1 The testing area covers three buildings, **a** CETC331 building with the 2nd floor, **b** HCXY building, and **c** SYL building with the 4th floor

before the WiFi collection in July 2017, only 56 APs were available and labeled in the floor plant. Four APs were in an office room, and two APs were in a meeting room. There are many offices on both sides of the corridor, and the rest are glass curtain walls. Due to the difference in wall material and structures, the corridor can be divided into four parts.

The WiFi collecting area is also set on the 4th floor of the SYL building in Jinan city, as shown in Fig. 1c. The total area is about 2600 m^2. It contains office rooms, meeting rooms, and corridors. 23 dual-band APs are deployed. The office room is the graduate student's laboratory with a complex layout. Six APs are fixed on the walls of the office room, and two APs are put on an iron chest and a wooden desk. Two APs are fixed on the wall of one meeting room, but no AP is in the other meeting room. The remaining thirteen APs are installed on both sides of the walls in the corridor; adjacent three APs can

form a triangle. WiFi RSS data were collected in January 2022. The data were collected at each grid point in the corridor while they were on the path in the office and meeting rooms.

Unlike the other datasets, clear floor plans are provided in three CAD files, which are named depending on the buildings. An independent coordinate system is utilized in a CAD drawing. The coordinates of training and testing points are processed. Everyone could render their own floor plan by themselves.

Description of data collection

Figure 2 clearly illustrates the procedure of data collection in a testing area. Several APs are installed on the walls or the ceiling. The pre-planned green solid point denotes RP or TP. The design location can be found with the help of tiles on the floor. The distance between two adjacent RPs is less than 1.2 m. The precise locations of APs and points are obtained using an electronic total station, a precise surveying instrument that can measure a distance of one kilometer with an error less than 3 mm. So, the coordinates and localization results can be represented in the order of a millimeter, i.e., three decimals.

A user holding smartphone in hand moves from one point to another in the forward direction. During the data collection, the smartphone faces up to the ceiling. The forward direction is consistent with the orientation of the pedestrian movement. When the user walks to the RP or TP, the user should stay at the location for a while and operate the self-developed application to collect the data at the sampling frequency of 1 Hz. The application could set up the count parameter. Once the count operation is complete, the user can conduct a similar process to the next point until all data is collected in the testing area.

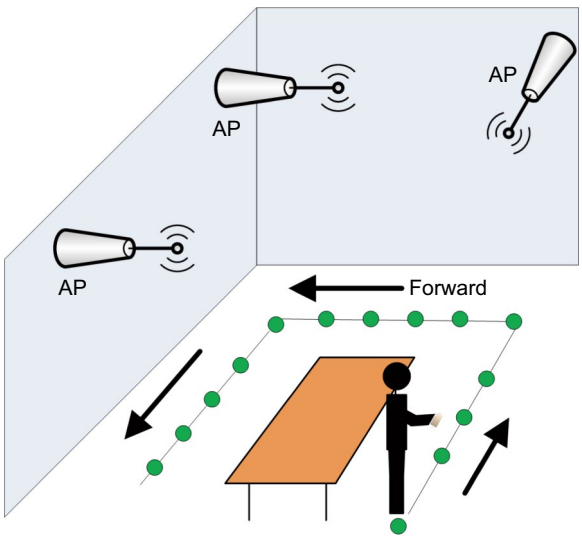

Fig. 2 The illustration of data collection in a testing area

An uncertainty distance is the interval between RP and the smartphone, which is easy to ignore. In the process of data collection, the position of RP is usually taken as that of the smartphone. In addition, due to the high degree of freedom of the smartphone held by the user, the uncertainty distance is dynamically changing and difficult to be compensated for, which leads to poor reliability and introduces positioning errors. However, the indoor positioning accuracy based on RSS is about 3–5 m, and the uncertainty distance is less than 0.5 m. So, the influence of the uncertainty distance on RSS-based localization can be ignored.

Description of data sheets

Like the UJIIndoorLoc dataset, the proposed SODIndoorLoc dataset also adopts sheets to store WiFi fingerprint records and other supplementary information. The most significant differences from the other existing datasets are the dense RPs, the addition of several attributes, and a sheet file containing APs' location and transmission frequency.

The division of data sheets

Three buildings are in different cities. The pre-installed APs operate three models of TP-Link, and the model of the pre-installed APs is the same in each building. For the convenience of presentation, the data sheets are discriminated according to buildings. The advantage of the division in this way is that the dimensions in each sheet are not large enough to cause dimension disaster. For example, the UJIIndoorLoc dataset is an aggregation of the data from three buildings, the dimension of the RSS vector is 520. If the researchers want to utilize the data from one of three buildings, they should be further conducted to filter and discard unwanted RSS elements. Table 2 shows the dimensions of the RSS vector in different data sheets, which are less than the UJIIndoorLoc dataset. The last volume denotes the dimensions of the RSS vector in the UJIIndoorLoc dataset. If three data sheets are integrated into the form of the UJIIndoorLoc dataset,

Table 2 Dimensions of RSS vector in different sheets

Buildings	Dimensions
CETC331	52
HCXY	347
SYL	363
ESTCE-TI	520

Table 3 Number of samples in different sheets

Buildings	Training sheets	Testing sheets
CETC331	955	840
HCXY	11 370	1 020
SYL	8 880	2 720

Table 4 Number of points in different sheets

Buildings	Training sheets	Testing sheets
CETC331	955	84
HCXY	379	86
SYL	296	102

the dimension will reach 762, which will be a significant challenge in the data processing. And suppose you want to utilize these data sheets for buildings identification. In that case, data enhancement is required by increasing the attributes of the RSS vector and setting the empty value as 100 dB·m. In practice, data enhancement is much less complicated than data reduction.

Depending on the purpose, data sheets are divided into two different sets: the training set and the testing set. The training set contains large amounts of RSS vectors and other relevant information at RPs for offline model learning. The testing set provides the same records at arbitrary TPs, which are usually located differently from RPs, to evaluate the performance of different methods.

In the process of WiFi collection, the sampling frequency is 1 Hz. The sampling time at each RP is thirty seconds and ten seconds at each TP. For a multi-story building, the volume of the CETC331 sheet will be very large if all thirty samples at each RP are stored. So, the CETC331 training sheet saves one sample, i.e., the average of thirty samples, corresponding to each RP. Nevertheless, all thirty samples at each RP are stored in HCXY and SYL training sheets. The number of TPs is smaller than that of RPs in three buildings. Ten samples at each TP are stored in three testing sheets. Table 3 shows the number of samples in the training and testing sheets. The total number of samples in three training sheets of the corresponding three buildings is 21,205, and that in three testing sheets is 2720. Both are larger than those in the UJIIndoorLoc dataset. Another HCXY and SYL training sheets also store the average sample corresponding to each RP, i.e., the average of thirty samples.

The corresponding numbers of RPs and TPs in different sheets are shown in Table 4. Because the distance between two adjacent RPs is about 0.5 m, the number of RPs in the CETC331 building is huge. The total number of RPs and TPs are 1670 and 272, respectively. The number of RPs in the proposed dataset is larger than that in the UJIIndoorLoc dataset. And most of the RPs and TPs are at different locations.

Table 5 Dimensions of RSS vector in simplified sheets

Buildings	Dimensions
CETC331	52
HCXY	56
SYL	46

Table 6 Summary of training and testing sheets

Buildings	Training sheets	Testing sheets
CETC331	Training_CETC331	Testing_CETC331
HCXY	Training_HCXY_All_30	Testing_HCXY_All
	Training_HCXY_All_Avg	
	Training_HCXY_AP_30	Testing_HCXY_AP
	Training_HCXY_AP_Avg	
SYL	Training_SYL_All_30	Testing_SYL_All
	Training_SYL_All_Avg	
	Training_SYL_AP_30	Testing_SYL_AP
	Training_SYL_AP_Avg	

As mentioned above, there are many pre-installed APs in three buildings. Dimension reduction for the HCXY and SYL sheets is conducted by filtering known MAC addresses to obtain corresponding simplified sheets. Table 5 shows the dimensions of the RSS vector in these simplified sheets.

Therefore, in the proposed dataset, there are nine training sheets and five testing sheets, as summarized in Table 6. Only a pair of training and testing sheets about the CETC331 building is provided, while there are four training sheets and two testing sheets for both the HCXY building and the SYL building. The letter "All" indicates all detected APs are adopted to build a whole RSS vector, and the letter "AP" denotes only pre-installed APs are utilized for filtering the RSS vector. The number "30" means that thirty samples at each RP are stored in the sheet. The letter "Avg" indicates that the average of thirty samples at each RP is in the sheet.

The format of training and testing sheets

Each WiFi fingerprint is characterized by the detected MAC addresses and the corresponding RSS values. And most Wireless Access Points (WAPs) are with multiple bands. It is not appropriate to assign WAP as an attribute, as the UJIIndoorLoc dataset does. The detected

Table 7 The header of a sheet and an example, the 12th sample of the Training_CETC331

MAC address number 1 (dB·m)	…	MAC address number *n* (dB·m)	Number($n+1$) of coordinate value in east (*E*) direction (m)	Number($n+2$) of coordinate value in north (*N*) direction (m)	FloorId ($n+3$)	BuildingId($n+4$)	SceneId ($n+5$)	UserId ($n+6$)	PhoneId ($n+7$)	Counts ($n+8$)
-44	…	100	47.400	18.000	1	1	1	4	3	1

MAC addresses are utilized for identifying RSS values in the proposed dataset. Regarding privacy, all detected MAC addresses in the building are sorted in the detected order. For example, the 1st attribute denotes RSS value of the 1st MAC address in the detected order, and the *n*th attribute is that of the *n*th MAC address. Because the numbers of detected MAC addresses in three buildings are different, *n* means different values in training and testing sheets for three buildings.

If the number of all detected MAC addresses is *n*, the range from the 1st attribute to the *n* th attribute can be expressed by *n* RSS values. The $(n+1)$th and the $(n+2)$th attributes indicate the coordinates in the east and north directions, named ECoord and NCoord. Identifiers (Id) of floor level, building, scene, user, and phone are indicated from the $(n+3)$th attribute to the $(n+7)$th attribute, named as FloorId, BuildingId, SceneId, UserId, and PhoneId in sequence. The $(n+8)$th attribute indicates the counts of samples, named Counts. The header of a sheet is shown as the first row of Table 7. And the 2nd row of Table 7 is an example, i.e., the 12th sample of the Training_CETC331.

RSS vector

RSS vector is a set of RSS values in the order corresponding to the first *n* attributes. MAC addresses and corresponding RSS values can be obtained from the nearby APs in each scan. If the program aligns the scanned MAC addresses to the first *n* attributes and assigns the corresponding RSS values to the attribute values. In that case, it sets the attribute values of the undetected MAC addresses to a particular value, e.g., 100 dB·m, suggested by the UJIIndoorLoc dataset. A whole RSS vector can be obtained, as shown in Table 7.

The scanned RSS value is a negative integer in the unit of dB·m, where -100 dB·m is equivalent to a weak signal, whereas 0 indicates an excellent signal. The probability distribution of RSS values from the training sheet and the testing sheet in three buildings is shown in Fig. 3. The minimum RSS value is -104 dB·m, the same value in the UJIIndoorLoc dataset. The larger one is -3 dB·m. The probabilities of maximum and minimum RSS values are tiny. Large amounts of RSS values are concentrated from -85 to -50 dB·m. RSS values conform to a Gaussian distribution.

Local coordinates

The adopted coordinate system is a local independent coordinate system. All coordinates in the proposed dataset are not the same as those in CAD drawings. They have been transformed for privacy reasons. The unit of local coordinates is in meters with three decimals.

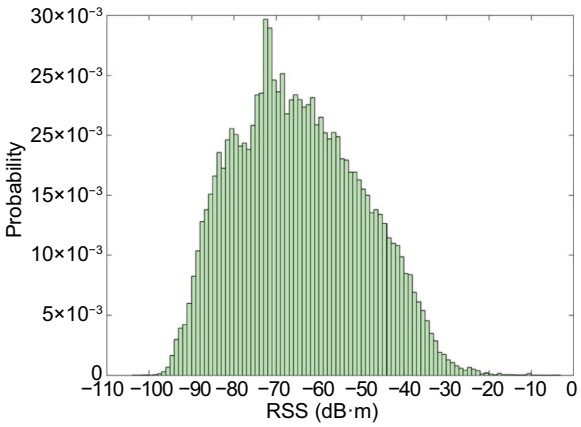

Fig. 3 The distributed probability of RSS values in three buildings

Space identifiers

FloorId, BuildingId, and SceneId are referred to as space identifiers. They are set as positive integer values from one to four. FloorId ranges from 1 to 3 in CETC331 sheets, and FloorId is four for HCXY and SYL sheets. BuildingId ranges from 1 to 3, and the CETC331, HCXY, and SYL buildings are set as 1, 2, and 3. There are three scenes in the WiFi collecting area, corridor, office, and meeting rooms. And they are set as 1, 2, and 3 in sequence. Hall and corridor are seamless in these buildings, so there is no distinction between the two scenes. Space identifiers in Table 7 mean that the WiFi collection is in a corridor on the 1st floor of the CETC331 building.

User identifier (UserId)

Ten students participated in the WiFi RSS collection in three buildings. They are marked with numbers from 1 to 10 instead of names. The height of each user is provided because we think this information might be necessary for range model simulation and range-based localization. The coarse height of the user holding a smartphone is also supplied, as shown in Table 8. The unit of height is centimeter.

Phone identifier (PhoneId)

Nine Android smartphones were utilized to collect WiFi data, among which two phones were in the same model and brand, e.g., Xiaomi 6, but in different memory sizes. Three brands were Xiaomi, Huawei, and Samsung, respectively. The detail can be found in Table 9.

Figure 4 shows the RSS ranges detected by different smartphones. The horizontal axis represents the identifier of smartphones, and the vertical axis denotes RSS values. Each bar represents the maximum and minimum RSS values detected by the smartphone. The

Table 8 Information of users

UserId	Height (cm)	Height of phone (cm)	UserId	Height (cm)	Height of phone (cm)
1	165	109	6	176	125
2	179	132	7	178	127
3	174	123	8	177	125
4	171	116	9	182	134
5	158	101	10	181	131

Table 9 Phone identifiers

Model	Id	Model	Id	Model	Id
Xiaomi 8	1	Huawei Mate 8	4	Xiaomi 6	7
Xiaomi 11	2	Xiaomi 6	5	Redmi 4	8
Xiaomi 4	3	Samsung S7	6	Xiaomi 5X	9

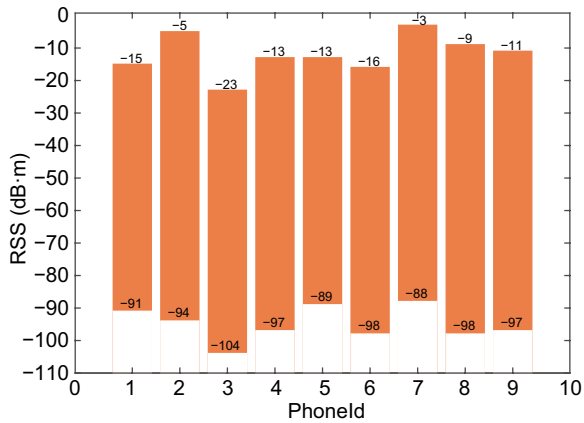

Fig. 4 RSS ranges vs. PhoneId

maximum value is at the top of each bar, and the minimum value is at the bottom. The number of RSS values larger than −20 dB·m is very small. Each bar has a different range of RSS values, even if smartphones are the same brand and model, e.g., both the 5th and 7th bars are the RSS values detected by Xiaomi 6. Obviously, there are differences in RSS values due to device heterogeneity. The statistics of RSS ranges with different smartphones might be helpful in solving device heterogeneity.

Counts

The timestamp register was introduced in the UJIIndoorLoc dataset in Unix time format to represent the time of the WiFi collection. But the count is adopted in the proposed dataset to record the times of WiFi samples, ranging from 1 to 30. In the Training_CETC331 sheet and training sheets labeled "Avg", the count is recorded as one at each sample. The counts range from 1 to 30 in the training sheets labeled by "30". The counts range from 1 to 10 for all the testing sheets.

Information of pre-installed APs

Table 10 is an example of the information on a pre-installed AP in the SYL building. It mainly contains space locations, MAC addresses, and channel frequencies. A sheet file is provided about the information on pre-installed APs for each building. This information is vital for range model simulation and range-based localization. Space locations are recorded in the format of RPs and TPs using ECoord, NCoord, and FloorId. Corresponding attributes replace MAC addresses in the order. In Table 10, the MAC addresses in 2.4 GHz and 5 GHz are the 125th and 340th ones, respectively. Channel frequency is the central frequency of the WiFi channel, which can reflect if the signal belongs to the 2.4 GHz band or 5 GHz band. The unit of channel frequency is MHz. For example, 2 437 MHz belongs to the 2.4 GHz band, while 5 220 MHz is the 5 GHz band. It is noted that the sheet of the HCXY building doesn't have the last two columns because the pre-installed APs are single-band. The height of an AP is not provided, and it can be customized.

Experiments based on the SODIndoorLoc dataset

Localization experiments based on the SODIndoorLoc dataset were conducted by using four kinds of machine learning methods (Ji et al., 2021; Maw et al., 2020; Salamah et al., 2016; Wu et al., 2019), i.e., K Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). Both classification and regression algorithms are adopted to evaluate the performance of localization based on the SODIndoorLoc dataset by using positioning accuracy and computational complexity. NN is implemented by the Multi-Layer Perceptron (MLP) algorithm, which trains the model using back propagation with no activation function in the

Table 10 An example of information of pre-installed APs

Id	Coordinate value in east (*E*) direction (m)	Coordinate value in north (*N*) direction (m)	FloorId	Attr_2.4	Freq_2.4 (MHz)	Attr_5	Freq_5 (MHz)
1	50.600	12.600	4	125	2 437	340	5 220

output layer. For the convenience of distinction, the above machine learning methods are referred to as K Nearest Classification (KNC), K Nearest Regression (KNR), Support Vector Classification (SVC), Support Vector Regression (SVR), Random Forest Classification (RFC), Random Forest Regression (RFR), Multi-Layer Perceptron Classification (MLPC), and Multi-Layer Perceptron Regression (MLPR). All simulations are carried out on MacBook (the central process unit is Intel Core M3 in 1.1 GHz) using Scikit-learn, an open-source machine learning toolkit. The coordinates are converted into labels to ensure that classification algorithms can be used for position estimation, and then the prediction labels obtained from the training model are converted into coordinates. Since the purpose is not to propose a new localization algorithm, no parameter optimization is carried out. The critical parameters of the above algorithms are artificially specified, and the rest are by default. The critical parameters of the same machine learning method are the same. For example, classification and regression algorithms of KNN are separately implemented by 'KNeighborsClassifier' and 'KNeighborsRegressor', where the integer value k is set as five, the weights are assigned as the inverse of the distance from the query point, and the distance metric is Euclidean. For SVC and SVR, the regularization parameter is set as one, the kernel is chosen as radial basis function, the kernel coefficient 'gamma' is 'scale', and the tolerance for the stopping criterion is specified as 0.01. The number of trees in the forest is set as 100 for 'RandomForestClassifier' and 'RandomForestRegressor'. MLP contains one input layer, one output layer, and one hidden layer, where there are 100 neurons. MLPC and MLPR are implemented by 'MLPClassifier' and 'MLPRegressor', respectively.

In this section, experiments with different training sheets are utilized for training localization models. The Mean Absolute Error (MAE), Root Mean Square Error (RMSE), the 50th percentile error, the 75th percentile error (suggested by IPIN competition as competition score), and 95th percentile error are introduced as positioning accuracy evaluation metrics, and the unit is meter. The computational complexity is indicated by the elapsed time, the summary of training and testing, and the unit is second.

Positioning experiments in the HCXY building

Positioning experiment with all MACs

Table 11 shows the statistics of positioning errors and elapsed time using the Training_HCXY_All_30 sheet, i.e., the 1st sheet. KNC, KNR, SVC, SVR, and RFR methods achieve good positioning accuracy with the MAE of around 2 m, while RFC, MLPC, and MLPR methods get poor positioning results. According to the suggestion of the IPIN competition, RFR will be treated as the best positioning algorithm because of the smallest 75th percentiles error. A set of boxes are utilized to show positioning errors of different methods in Fig. 5, where the red line in the blue box indicates the 50th percentiles error, the bottom and top edges of the blue box indicate the 25th and 75th percentiles error, the red square in the blue box denotes the MAE, and the red plus symbol outside of the blue box denotes the outliers with large errors. It seems that the RFR method may not have the best positioning accuracy, and SVC achieves better positioning accuracy than other methods by comparing Table 11 and Fig. 5.

Positioning experiment with pre-installed APs

Table 11 shows the statistics of positioning errors and elapsed time based on the Training_HCXY_AP_30 sheet, i.e., the 2nd sheet. The errors of the RFR method are obviously smaller than those of the other seven methods with the MAE of 3.198 m.

Although the elapsed time of methods with the training sheet only containing pre-installed APs are shorter than those containing all MAC addresses, the positioning errors basically become larger as the decrease of the number of MACs in the training sheet. The average positioning accuracy decreases by 41.6% from 347-dimension fingerprint data to 56-dimension one. It is hard to provide high-precision positioning results by vastly simplifying high-dimensional data in the HCXY building. If positioning accuracy and computational complexity are considered, the number of MAC addresses can be appropriately increased.

Positioning experiments in the SYL building

Only regression methods are used for investigating positioning performance based on the training sheets with different samples and MACs in the SYL building, i.e., the four training sheets in Table 6. They are distinguished in order by using numbers from 1 to 4. 32 sets of positioning results can be obtained, as well as the corresponding statistics of positioning errors and elapsed time, as shown in Table 12. The best positioning accuracy for each training sheet can be found by comparing with metrics of different methods.

For the 1st sheet in Table 12, the SVR method achieves the best positioning accuracy with the MAE of 3.844 m, expressed as SVR_All_30. For the 2nd sheet, the RFR method achieves the best positioning accuracy with the MAE of 3.787 m, expressed as RFR_AP_30. The KNR method achieves the best positioning accuracy with the MAE of 3.782 m for the 3rd sheet and with the MAE of 4.048 m for the 4th sheet. They are sequentially expressed as KNR_All_Avg and KNR_AP_Avg.

Table 11 Statistics of different methods based on the two sheets in the HCXY building

Sheets	Methods	MAE (m)	RMSE (m)	50th percentiles error (m)	75th percentiles error (m)	95th percentiles error (m)	Elapsed time (s)
1	KNC	2.339	2.096	1.342	3.059	6.686	0.44
	KNR	2.365	2.033	1.47	3.231	6.627	0.445
	SVC	2.259	1.966	1.342	3.081	6.627	17.201
	SVR	2.652	2.177	2.169	3.385	6.707	66.053
	RFC	6.242	10.057	1.879	5.433	30.606	11.458
	RFR	2.412	3.309	1.382	2.87	7.572	27.337
	MLPC	29.411	25.435	21.054	42.957	79.122	71.15
	MLPR	8.269	4.792	7.586	11.275	17.435	10.932
2	KNC	5.893	7.232	3.059	6.689	25.807	0.304
	KNR	5.813	7.22	3.048	7.341	25.807	0.286
	SVC	5.916	7.678	3	6.708	25.807	11.916
	SVR	5.108	4.659	3.094	6.876	15.62	35.678
	RFC	7.227	10.931	1.942	5.532	35.405	6.059
	RFR	3.198	3.724	2.101	4.129	8.294	7.941
	MLPC	6.955	9.531	3.547	7.58	30.606	40.914
	MLPR	15.091	15.714	7.675	26.663	47.184	9.332

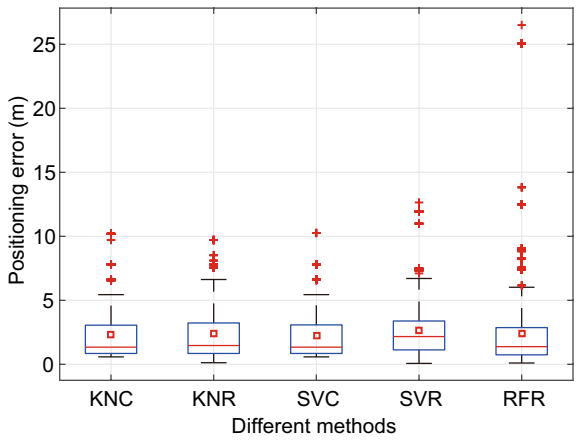

Fig. 5 Positioning errors of different methods based on the training sheet containing all APs and thirty samples at each RP in the HCXY building

The corresponding results can be illustrated as Cumulative Distribution Function (CDF) for a comparison of positioning accuracy. The maximum errors are larger than 50 m, so the positioning errors are limited to the range from 0 to 10 m. CDFs of these four methods are shown in Fig. 6. The differences among the four curves are very small. And the trends of the four curves are consistent.

However, the elapsed time of four regression methods based on the four training sheets are quite different. Compared with SVR_All_30 and RFR_AP_30, the elapsed time of KNR_All_Avg and KNR_AP_Avg are reduced by more than 99%. Adopting the average sample at each RP can significantly reduce the computational complexity with the consistent positioning accuracy in the SYL building.

Positioning experiments in the CETC331 building

Positioning experiments based on the training sheet in the CETC331 building were conducted using four regression methods to evaluate the impact of floor discrimination on positioning performance. According to whether floor discrimination is used for location estimation in the experiments, the experiments are divided into two cases. The 1st case is that the whole data are directly utilized in all calculations without floor discrimination. The 2nd case is that training models and further location estimation are carried out after floor discrimination. In this section, floor discrimination is not identified using the RSS vector but based on the FloorId attribute.

Table 13 summarizes the statistics of positioning errors and the elapsed time in two cases. The positioning effect of the KNR method is the same regardless of the adoption of floor discrimination. Minor changes in positioning errors occurred when the SVR method adopted floor discrimination. And the elapsed time of KNR and SVR methods are greatly reduced by 55.5% and 66.9%, respectively. There are slight changes in positioning errors and elapsed time when the RFR and MLPR methods adopt floor discrimination. MAEs and RMSEs decrease while the 50th, 75th, 95th, and elapsed time increase.

Figure 7 shows CDFs of positioning errors of four regression methods after floor discrimination. The red

Table 12 Statistics of regression methods based on training sheets containing different MAC addresses and samples in the SYL building

Sheets	Methods	MAE (m)	RMSE (m)	50th percentiles error (m)	75th percentiles error (m)	95th percentiles error (m)	Elapsed time (s)
1	KNR	4.656	5.291	3.842	6.078	8.521	0.368
	SVR	3.844	4.948	3.096	4.696	7.730	18.637
	RFR	4.461	5.890	3.058	4.585	15.241	14.785
	MLPR	6.624	5.153	6.318	8.209	12.143	7.474
2	KNR	4.752	5.369	3.667	5.532	9.232	0.238
	SVR	4.427	4.980	3.412	5.419	9.753	10.983
	RFR	3.787	5.094	2.785	4.551	8.122	4.199
	MLPR	5.687	4.964	4.782	7.007	12.066	6.229
3	KNR	3.782	4.975	3.087	4.435	7.751	0.017
	SVR	5.445	4.708	4.651	7.668	10.951	0.308
	RFR	3.832	4.897	3.149	4.178	8.272	0.947
	MLPR	11.007	6.392	10.284	14.018	22.954	0.764
4	KNR	4.048	4.970	3.544	4.810	7.922	0.014
	SVR	5.772	4.921	5.113	7.871	11.414	0.222
	RFR	4.011	5.625	2.655	4.607	11.736	0.473
	MLPR	5.679	5.110	4.731	7.230	11.466	0.543

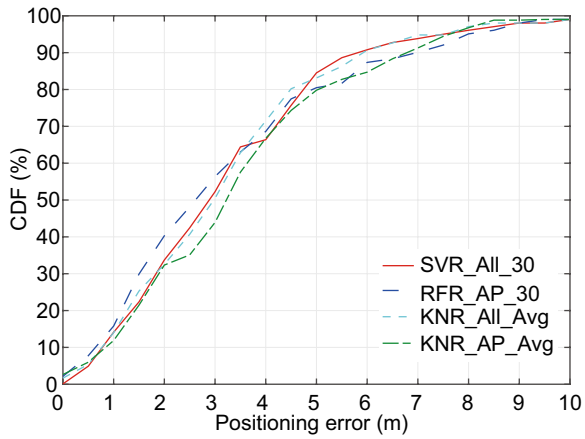

Fig. 6 Positioning errors of several regression methods based on the four training sheets containing different MACs and samples in the SYL building

line with circles indicates the CDF of the KNR method. It is always higher than the other lines. Considering Table 13 and Fig. 7 together, the KNR method achieves the best positioning performance with the MAE of 2.876 m and the elapsed time of 0.015 s based on the CETC331 training sheet.

Discussions and conclusions

Discussions

Some details, such as machine learning methods, precision, parameters, and the conversion between coordinates and labels, are not specified above. These issues are discussed and explained in this section.

Machine learning methods

Many machine learning methods can be used for fingerprint-based localization in indoor positioning and navigation. And the Scikit-learn toolkit provides abundant

Table 13 Statistics of regression methods based on the training sheet in the CETC331 building

Case	Methods	MAE (m)	RMSE (m)	50th percentiles error (m)	75th percentiles error (m)	95th percentiles error (m)	Elapsed time (s)
1	KNR	2.876	2.757	2.316	3.803	6.307	0.036
	SVR	3.522	3.009	3.084	4.304	8.834	0.829
	RFR	3.740	4.171	2.561	4.099	11.667	1.890
	MLPR	5.152	3.626	4.379	6.166	11.837	1.274
2	KNR	2.876	2.757	2.316	3.804	6.306	0.016
	SVR	3.567	2.991	3.143	4.421	8.643	0.274
	RFR	3.579	3.631	2.683	4.265	9.241	1.922
	MLPR	5.124	3.360	4.516	6.327	11.212	3.374

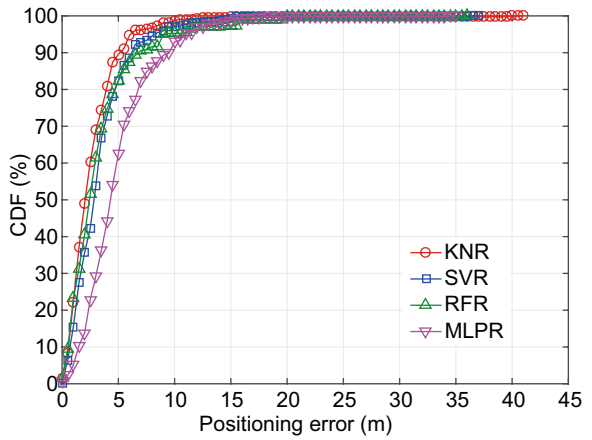

Fig. 7 Positioning errors of regression methods based on the training sheet after floor discrimination in the CETC331 building

interfaces, which can easily implement training models and predictions. In addition to the listed KNN, SVM, RF, and NN methods, several other methods have also been tried for location estimation, but their positioning accuracies are poor, and the corresponding training time is very long. Therefore, only four kinds of machine learning methods are listed.

Positioning accuracy and parameters

Positioning accuracy of each method depending to buildings or data sheets is separately provided. Due to the adoption of the default parameters for each method, the results obtained by many methods are not very accurate.

For example, the MLPR method achieves the MAE of 11.007 m based on the Training_SYL_AP_30 sheet in the SYL building. But the MLPR method achieves a smaller MAE based on the Training_SYL_AP_Avg sheet, which is the simple version of the Training_SYL_AP_30 sheet. And KNR, SVR, and RFR methods achieve better performance based on the Training_SYL_AP_30 sheet than the Training_SYL_AP_Avg sheet. The main reason for the MLPR method with large positioning errors is that the globally optimal model is not established within the finite iterations. The default value of iterations in the MLPR method is 2000. By constantly adjusting parameters and using optimization algorithms, these methods can obtain high-precision positioning results, but further research on positioning accuracy is not conducted because it is not the objective of this paper.

Positioning accuracy and MACs per unit area

Table 11 shows that positioning errors become very large when the training sheet changes from the Training_HCXY_All_30 to the Training_HCXY_AP_30. The number of MACs is changed from 347 to 56. For example, the MAEs of KNC, KNR, SVC, SVR, and MLPR methods are increased by 60.3%, 59.3%, 61.8%, 48.1%, and 45.2%, respectively. The RMSEs of KNC, KNR, SVC, SVR and MLPR methods are increased by 71%, 71.8%, 74.4%, 53.3% and 69.5%, respectively. Outwardly, the reason for large errors is the great decrease in the number of MACs. However, when the training sheet changes from the Training_SYL_All_30 to the Training_SYL_AP_30 in the SYL positioning experiment, the number of MACs is changed from 363 to 46, MAEs are increased by 2% and 13% for KNR and SVR methods, MAEs are decreased by 17.8% and 16.5% for RFR and MLPR methods, the trend of corresponding RMSEs is consistent. The coverage of the 4th floor in the HCXY building is much broader than in the SYL building. Therefore, the decrease in positioning accuracy is not due to the small amount of MACs but the small number of MACs per unit area.

Positioning accuracy and the conversion between coordinate and label Some classification methods don't support multiple outputs, which requires us to convert coordinates into labels. In Table 11, the number of positioning errors of classification methods larger than those of regression methods is five. In the SYL positioning experiments, the number is fourteen. In other words, there is a high probability that the positioning error of the classification method is greater than that of the regression method. MAEs of classification methods are usually hundreds based on the training sheets with the letter "Avg" in the HCXY building. It is also why not provide the statistics of positioning errors based on the training sheets with the average RSS vector at each RP in the HCXY building. However, high-precision positioning results are obtained by classification methods based on the training sheets with 30 samples at each RP in the HCXY building in Table 11. The problem has puzzled us for a long time. In future research, the provided classifier with multiple outputs by the Scikit-learn will be adopted for classification methods to estimate location, and the conversion between coordinate and label will be gradually abandoned.

Conclusions

A supplementary open dataset for WiFi indoor localization was created. The SODIndoorLoc is a dataset with dense RPs. And the locations and channel information of pre-installed APs are provided. Therefore, it can be treated as a supplement to the UJIIndoorLoc dataset. It covers three buildings and multiple floors where there are corridors, office rooms, and meeting rooms. The total covered area is about 8000 m^2. More than 1800 points at different locations were arranged,

and the number of RPs is about six times as many as TPs. 23,935 samples were recorded at these points with 21,205 samples for training/learning and 2720 for testing/validation. 105 single-band and dual-band APs were pre-installed in the three buildings. Given the differences in the number of samples and MACs in the training data, there are nine training sheets and five corresponding testing sheets in the dataset. The distance between two adjacent sampling points is about 1.2 m in two buildings, while about 0.5 m in a three-story building. Four kinds of machine learning methods (eight variants in total) are introduced to estimate the locations of TPs with default parameters. The best average positioning accuracy is about 2.3 m.

Because of dense RPs, the locations of pre-installed APs, and CAD drawings of each floor, the SODIndoorLoc dataset can be used for clustering, classification, and regression to compare the performance of different indoor positioning applications based on WiFi fingerprint, e.g., high-precision positioning, building, floor recognition, fine-grained scene identification, range model simulation, and rapid construction of fingerprint datasets.

Acknowledgements
We would like to thank Ye Tao, Meiqi Zhao, Hongxia Qi, Gang Yuan, Shenglei Xu, Jiapeng Zhou, Teng Wang, and Fengfeng Zhao for their help with the data collection. Jingxue Bi thanks Emilio Sansano from University Jaime I for the help in getting datasets from the IndoorLoc platform. We thank editors and reviewers for their helpful and constructive comments on our work.

Author contributions
BJX and CHJ proposed the idea and carried out the simulation; WYJ and YBG recommended the experiment part, WYJ, STG and HL assisted in the installation of APs in three buildings. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (No. 42001397), the National Key Research and Development Program of China (No. 2016YFB0502102), the Introduction and Training Program of Young Creative Talents of Shandong Province (No. 0031802), the Doctoral Research Fund of Shandong Jianzhu University (No. XNBS1985), and the National College Student Innovation and Entrepreneurship Training Program (No. S202110430036).

Availability of data and materials
The present dataset is released in the GitHub repository of the first author, and the URL is https://github.com/renwudao24/SODIndoorLoc.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China. [2]School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China. [3]Key Laboratory of Satellite Navigation System and Equipment Technology, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China.

Received: 15 May 2022 Accepted: 12 October 2022
Published online: 08 November 2022

References

Alvarez-Merino, C. S., Luo-Chen, H. Q., Khatib, E. J., & Barco, R. (2021). Opportunistic fusion of ranges from different sources for indoor positioning. *IEEE Communications Letters, 25*(7), 2260–2264.
Bi, J. (2022). SODIndoorLoc. https://github.com/renwudao24/SODIndoorLoc. Accessed 13 Jul 2022.
Cao, H., Wang, Y., Bi, J., Xu, S., Si, M., & Qi, H. (2020). Indoor positioning method using WiFi RTT based on LOS identification and range calibration. *ISPRS International Journal of Geo-Information, 9*(11), 627.
Cao, X., Zhuang, Y., Yang, X., Sun, X., & Wang, X. (2021). A universal Wi-Fi fingerprint localization method based on machine learning and sample differences. *Satellite Navigation, 2*(1), 1–15.
Chen, L., Zhou, X., Chen, F., Yang, L.-L., & Chen, R. (2021). Carrier phase ranging for indoor positioning with 5G NR signals. *IEEE Internet of Things Journal, 9*(13), 10908–10919.
Feng, D., Wang, C., He, C., Zhuang, Y., & Xia, X.-G. (2020). Kalman-filter-based integration of IMU and UWB for high-accuracy indoor positioning and navigation. *IEEE Internet of Things Journal, 7*(4), 3133–3146.
Gönültaş, E., Lei, E., Langerman, J., Huang, H., & Studer, C. (2021). CSI-based multi-antenna and multi-point indoor positioning using probability fusion. *IEEE Transactions on Wireless Communications, 21*(4), 2162–2176.
Guo, G., Chen, R., Ye, F., Peng, X., Liu, Z., & Pan, Y. (2019). Indoor smartphone localization: A hybrid WiFi RTT-RSS ranging approach. *IEEE Access, 7*, 176767–176781.
Huang, L., Chen, R., Ye, F., Liu, Z., Li, Z., Xu, S., Guo, G., & Qian, L. (2022). An indoor positioning system based on combined audio chirp/mems/floor map: Performance analysis of kepler A100. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 46*, 53–60.
Ji, W., Zhao, K., Zheng, Z., Yu, C., & Huang, S. (2021). Multivariable fingerprints with random forest variable selection for indoor positioning system. *IEEE Sensors Journal, 22*(6), 5398–5406.
Kunhoth, J., Karkar, A., Al-Maadeed, S., & Al-Ali, A. (2020). Indoor positioning and wayfinding systems: A survey. *Human-Centric Computing and Information Sciences, 10*(1), 1–41.
Li, B., Zhao, K., & Sandoval, E. B. (2020). A UWB-based indoor positioning system employing neural networks. *Journal of Geovisualization and Spatial Analysis, 4*(2), 1–9.
Li, Y., Zhuang, Y., Lan, H., Zhang, P., Niu, X., & El-Sheimy, N. (2016). Self-contained indoor pedestrian navigation using smartphone sensors and magnetic features. *IEEE Sensors Journal, 16*(19), 7173–7182.
Liu, F., Liu, J., Yin, Y., Wang, W., Hu, D., Chen, P., & Niu, Q. (2020). Survey on WiFi-based indoor positioning techniques. *IET Communications, 14*(9), 1372–1383.
Liu, G., Yu, B., Huang, L., Shi, L., Gao, X., & He, L. (2021). Human-interactive mapping method for indoor magnetic based on low-cost MARG sensors. *IEEE Transactions on Instrumentation and Measurement, 70*, 1–10.
Lohan, E. S. (2013). Open-source software and measurement data available at TLTPOS group, TUT. https://homepages.tuni.fi/elena-simona.lohan/pos.cs.tut.fi/pos//Software.htm. Accessed 13 Jul 2022.
Lohan, E. S., Torres-Sospedra, J., Leppäkoski, H., Richter, P., Peng, Z., & Huerta, J. (2017). Wi-Fi crowdsourced fingerprinting dataset for indoor positioning. *Data, 2*(4), 32.
Maheepala, M., Kouzani, A. Z., & Joordens, M. A. (2020). Light-based indoor positioning systems: A review. *IEEE Sensors Journal, 20*(8), 3971–3995.
Maw, M. M., Tint, H. M. N. M., & Duangsuwan, S. (2020). Analysis of indoor Wi-Fi localization using gaussian process regression and K-nearest neighbor algorithms. *UTK Research Journal, 14*(1), 30–39.
Montoliu, R., Sansano, E., Torres-Sospedra, J., & Belmonte, O. (2017). IndoorLoc platform: A public repository for comparing and evaluating indoor positioning systems. In 2017 International conference on indoor positioning and indoor navigation (IPIN), Sapporo, Japan, 2017, 1–8.
Morar, A., Moldoveanu, A., Mocanu, I., Moldoveanu, F., Radoi, I. E., Asavei, V., Gradinaru, A., & Butean, A. (2020). A comprehensive survey of indoor localization methods based on computer vision. *Sensors, 20*(9), 2641.
Moreira, A., Silva, I., Meneses, F., Nicolau, M. J., Pendao, C., & Torres-Sospedra, J. (2017). Multiple simultaneous Wi-Fi measurements in fingerprinting

indoor positioning. In 2017 International conference on indoor positioning and indoor navigation (IPIN), Sapporo, Japan, 2017, 1–8.
Nahrstedt, K., & Vu, L. (2012). The uiuc/uim dataset. https://crawdad.org/uiuc/uim/20120124/. Accessed 13 Jul 2022.
Parasuraman, R., Caccamo, S., Baberg, F., & Ogren, P. (2016). The kth/rss dataset. https://crawdad.org/kth/rss/20160105/. Accessed 13 Jul 2022.
Potortì, F., Torres-Sospedra, J., Quezada-Gaibor, D., Jiménez, A. R., Seco, F., Pérez-Navarro, A., Ortiz, M., Zhu, N., Renaudin, V., & Ichikari, R. (2022). Off-line evaluation of indoor positioning systems in different scenarios: the experiences from IPIN 2020 competition. *IEEE Sensors Journal, 22*(6), 5011–5054.
Poulose, A., & Han, D.S. (2020). Hybrid deep learning model based indoor positioning using Wi-Fi RSSI heat maps for autonomous applications, Electronics, 10:2.
Qin, F., Zuo, T., & Wang, X. (2021). Ccpos: Wifi fingerprint indoor positioning system based on cdae-cnn. *Sensors, 21*(4), 1114.
Richter, P., Lohan, E. S., & Talvitie, J. (2018). WLAN (WiFi) RSS database for fingerprinting positioning. https://zenodo.org/record/1161525. Accessed 13 July 2022.
Rocamora, J. M., Wang-Hei Ho, I., Mak, W. M., & Lau, A. P. T. (2020). Survey of CSI fingerprinting-based indoor positioning and mobility tracking systems. *IET Signal Processing, 14*(7), 407–419.
Ruiz, A. R. J., Granja, F. S., Honorato, J. C. P., & Rosas, J. I. G. (2011). Accurate pedestrian indoor navigation by tightly coupling foot-mounted IMU and RFID measurements. *IEEE Transactions on Instrumentation and Measurement, 61*(1), 178–189.
Salamah, A. H., Tamazin, M., Sharkas, M. A., & Khedr, M. (2016). An enhanced WiFi indoor localization system based on machine learning. In 2016 International conference on indoor positioning and indoor navigation (IPIN), Alcala de Henares, Spain, 2016, 1–8.
Tao, Y., & Zhao, L. (2021). AIPS: An accurate indoor positioning system with fingerprint map adaptation. *IEEE Internet of Things Journal, 9*(4), 3062–3073.
Tian, H., Zhu, L. (2020). MIMO CSI-based super-resolution AoA estimation for Wi-Fi indoor localization. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China, 2020, 457–461.
Torres-Sospedra, J., Montoliu, R., Martínez-Usó, A., Avariento, J. P., Arnau, T. J., Benedito-Bordonau, M., & Huerta, J. (2014). UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. In 2014 international conference on indoor positioning and indoor navigation (IPIN), Busan, South Korea, 2014, 261–270.
Torres-Sospedra, J., Moreira, A., Mendoza-Silva, G. M., Nicolau, M. J., Matey-Sanz, M., Silva, I., Huerta, J., & Pendão, C. (2019). Exploiting different combinations of complementary sensor's data for fingerprint-based indoor positioning in industrial environments. In 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Pisa, Italy, 2019, 1–8.
Wu, P., Imbiriba, T., LaMountain, G., Vilà-Valls, J., & Closas, P. (2019). WiFi fingerprinting and tracking using neural networks. In Proceedings of the 32nd international technical meeting of the satellite division of the institute of navigation (ION GNSS+ 2019), Florida, United States, 2019, 2314–2324.
Xu, Y., Cao, J., Shmaliy, Y. S., & Zhuang, Y. (2021). Distributed Kalman filter for UWB/INS integrated pedestrian localization under colored measurement noise. *Satellite Navigation, 2*(1), 1–10.
Ye, H., Yang, B., Long, Z., & Dai, C. (2022). A method of indoor positioning by signal fitting and PDDA algorithm using BLE AOA device. *IEEE Sensors Journal, 22*(8), 7877–7887.
Zhuang, Y., Syed, Z., Li, Y., & El-Sheimy, N. (2015). Evaluation of two WiFi positioning systems based on autonomous crowdsourcing of handheld devices for indoor navigation. *IEEE Transactions on Mobile Computing, 15*(8), 1982–1995.

Publisher's Note