RSS Fingerprinting Dataset Size Reduction Using Feature-Wise Adaptive k-Means Clustering

Lucie Klus[*,†], Darwin Quezada-Gaibor[†,*], Joaquín Torres-Sospedra[‡,†],
Elena Simona Lohan[*], Carlos Granell[†] and Jari Nurmi[*]
[*]*Electrical Engineering Unit*, *Tampere University*, Tampere, Finland
[†]*Institute of New Imaging Technologies*, *Universitat Jaume I*, Castellón, Spain
[‡]*UBIK Geospatial Solutions S.L.*, Castellón, Spain

***Abstract*—Modern IoT devices, that include smartphones and wearables, usually have limited resources. They require efficient methods to optimize the use of internal storage, provide computational efficiency, and reduce energy consumption. Device resources should be used appropriately, especially when employed for time-consuming and energy-intensive computations such as positioning or localization. However, reducing computational costs usually degrades the positioning methods. Therefore, the goal of this article is to propose and compare compression mechanisms of the fingerprinting datasets for energy-saving without losing relevant information, by using adaptive k-means clustering. As a result, we achieved a compression ratio of up to 15.97 with a small decrease (1%) in position error.**

***Index Terms*—clustering, compression ratio, data compression, fingerprinting, indoor positioning, k-means, k-nearest neighbors**

I. Introduction and motivation

Mass-market wearables are steadfastly developing as one of the many future markets of Internet of Things (IoT) applications. Main characteristics of most wearables are that they are typically power-constrained, size-constrained, and cost-constrained devices, integrating several mass-market sensors with various capabilities, ranging from measuring physiological parameters to ensuring low-cost wireless communications and positioning solutions. Many wearables do not have Global Navigation Satellite System (GNSS) chipsets embedded and must perform indoor localization based on non-GNSS sensors, such as Bluetooth Low Energy (BLE), WiFi, ZigBee, Ultra Wide-Band (UWB) chipsets, accelerometers, gyroscopes, and/or barometers [1].

Indoor Positioning Systems (IPSs) based on wearables are attracting the attention of the research community more and more. For instance, Belmonte et al. [1] compared several IPS solutions for smart homes in Ambient Assisted Living (AAL) in terms of cost, scalability, obtrusiveness, connectivity, interoperability, and extensibility. WiFi-based positioning using fingerprinting with Received Signal Strength (RSS) measurements was found to offer the best trade-off in terms of considered criteria. However, energy consumption was not included in their study.

Corresponding Author: Lucie Klus (lucie.klus@tuni.fi)

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Sklodowska Curie grant agreement No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints, http://www.a-wear.eu/).

The authors in [2] discussed the most common wireless technologies on wearables and pointed towards BLE as a lower cost solution than WiFi. Wireless positioning aspects were only briefly addressed, which was assumed to come primarily from GNSS chipsets.

Wearable-based positioning with BLE sensors and RSS measurements has also been addressed recently in [3], by using Machine Learning (ML) techniques. A 5-layered Artificial Neural Network (ANN) was 8.5 times faster than k-NN at the expense of decreasing the accuracy of the label-based (classification) positioning system by 5%.

The above-mentioned recent research efforts show the increased interest towards building more accurate, more energy efficient, and robust IPS involving wearable devices. Our paper focuses on fingerprint-based methods and, in particular, on the representation of the RSS values in the radio map. The authors propose a novel compression method to reduce the data storage requirements for the RSS value, while benefiting the IPS positioning accuracy in wearable devices.

The main contributions of this paper are:

- A novel approach on applying clustering on RSS datasets.
- A new method for data compression utilizing k-means clustering and the substitution of RSS measurements with reduced alphabet representation.
- The validation of the method on 16 different RSS datasets
- The source code for its implementation offered in open access for the research community.

This article is divided into the following sections. Section II gives a general overview of the related work. Section III describes the method developed for fingerprinting data compression. Section IV presents the experiments and results. Finally, Section V provides the main conclusions of this work.

II. Related work

Given that computing efficiency, dimensionality reduction, and data compression are highly demanded in IPS, different authors have proposed multiple methods based on clustering [4], radio-map reduction [5], [6] and other complex-search algorithms [7]. These methods may be executed in dedicated servers, smartphones, and even in low profile devices. Additionally, the combination of IoT, ML, and wearable devices requires efficient algorithms in order to reduce energy consumption [8].

WiFi fingerprinting is commonly used for indoor positioning due to the fact that WiFi is widely deployed in multiple environments (indoor and outdoor). However, this method requires having one or more datasets (radio maps) that are necessary to estimate the user's position. In many cases, they are large datasets with thousands of samples which are not suitable for some IoT devices. Moreover, computing the distances to all fingerprints in the radio map might be too inefficient [9], especially in large operational areas.

To reduce the size of WiFi radio maps, some authors have proposed the dimensionality reduction. For instance, Abed *et al.* [10] exploit the feature of some APs to transmit more than one signal (Multiple Service Set Identifiers, MSSID), and they propose a new dimensionality-reduction technique based on it. Their objective is to identify the most relevant APs, improve computational efficiency, and reduce the effects of multipath propagation. Their approach is divided into two phases - online and offline phase. The offline phase is devoted to combining the MSSID vectors for each AP reducing the multipath effect. Additionally, in this step, a location based clustering process groups samples in small zones. In the online phase, the operational fingerprint is compared with the centroids of each cluster to find the most similar one. Finally, the author estimates the position by using the k-nearest neighbour algorithm (or k-NN) with the reference fingerprints falling into the selected cluster.

Other researchers are focused on the minimal description length of the data (data compression), for instance, by using Symbolic Aggregate ApproXimation (SAX) [11], [12], XOR-based compression, simple 8-b, etc [13]. These algorithms also provide computational efficiency. For example, Baldini *et al.* [14] study the use of SAX approach in RF Fingerprinting, demonstrating that this algorithm is more computationally efficient by reducing the execution time to 30% compared to the original time series.

Doan *et al.* [15] proposed a new framework based on lossless compression in order to provide efficient data storage and data indexing. This framework was divided into six blocks which are: data encoding, splitting, zigzag encoding, bit conversion, aggregation, and padding aggregate record. As a result, they saved 97% of the storage space, which is almost 3% more than the other techniques used for data compression.

Azar *et al.* [16] studied the effects of using lossy data compression techniques on time series by using deep learning. Their main approach is the combination of error-bound compressor (Squeeze) and Discrete Wave Transform (DWT) lifting scheme obtaining a high data compression ratio.

Based on the related work, it is obvious that there are benefits of applying data compression or dimensionality reduction over the datasets. In most cases, they provide high computational efficiency while extending the life time of IoT devices. However, when we use these techniques, the compressed data cannot be restored to its original form, and therefore, some amount of data is lost in the compression or dimensionality reduction process. This may lead to the decreased positioning accuracy of IPS.

The approach of utilizing clustering for the purpose of data compression or improving the performance of the system was explored in the past, e.g., in [17]. Traditionally, clustering on the fingerprinting data is realized by finding similarities in the fingerprints across all features and assigning a cluster to each fingerprint sample. The assigned clusters are then utilized to speed up the process of localizing the user by faster finding similar fingerprints. This paper explores the utilization of clustering on each measurement separately, thus substituting the actual measured RSS value with the cluster index. As the result, the size of the whole dataset is significantly reduced without reducing the amount of measurements and without the significant degradation of the dataset quality for localization purposes. Additionally, the proposed method is able to operate online, efficiently shifting cluster centroids with each newly measured fingerprint.

III. Proposed data compression algorithm

The symbols and notations used in this paper are captured in Table I.

TABLE I
SYMBOLS AND NOTATIONS USED IN THIS PAPER

$ceil()$	Function rounding a number up to the nearest integer
C_n	nth cluster's centroid coordinate
CR	Compression ratio
δ_{MSE}	Mean squared error difference
i	Iteration
K	Number of clusters
MSE	Mean squared error
n_{bits}	Number of bits used to represent the dataset
N_n	nth cluster's count in the dataset
N_{UNIQUE}	Number of unique values in the dataset
s_m^t	tth sample's value of the mth feature
X	Initial dataset used in the first stage
ξ_{KNN}	3D positioning error ratio

The method proposed in this paper aims to reduce data storage requirements, and it is based on clustering. Unsupervised learning method k-means was targeted due to its low complexity and good viability to find patterns in the non-complex data included in RF fingerprinting datasets. The novelty of our method comes from the fact that the proposed method reduces storage requirements by substituting the measured values with a reduced "alphabet", which represents the centroids assigned to each feature of each sample. This is different from the traditional approaches which typically reduce the number of features in the dataset, using principal component analysis (PCA), autoencoders (AEs), or other dimensionality reduction machine learning approaches.

This section includes a short introduction of k-means clustering, followed by the description of the proposed model for data compression.

Lloyd's algorithm, or k-means, is the most commonly and frequently used clustering algorithm worldwide [18]. There, each cluster is represented only using the coordinates of its centroid. The method requires the initial dataset, which is clustered in two repeating steps similar to expectation-maximization algorithm used in more complex, stochastic

methods. The algorithm is initiated by selecting the initial centroids of the clusters, either at random, using given coordinates or using e.g. k-means++ algorithm [19]. In the first step, each sample is assigned to the nearest cluster centroid, based on the chosen distance metric, usually Euclidean. The second step consists of shifting each centroid's coordinates to better represent the assigned data, as the mean coordinate of the assigned samples. These two steps are repeated until the samples no longer change the assigned clusters, or until a maximum iteration is reached.

In the first stage our proposed method applies k-means with k-means++ initiation [19] on the referenced dataset, namely the radio map with access point (AP) measurements. In the second stage, which is designed to operate online, new samples are assigned to the existing clusters and the cluster centroids are adjusted based on the new sample coordinates. Because of that, the clusters always represent the whole dataset at the given time and adjust accordingly with each new sample. The following paragraphs describe the proposed approach.

A. First Stage

The data in the fingerprinting dataset consist of individual samples. Each sample consists of a feature vector (power level measurements from considered APs), and a target vector (a set of values, usually spatial coordinates). A feature refers to power level measurement from a single, specific AP across all samples. In here, we assume that the data is stored per AP. Alternatively, the data can be also stored per each measurement point [17].

At the beginning of the first stage, the multiple features are either merged, if they represent the same physical entity, e.g. RSS measurements from separate APs or antennas, or are kept separate if they represent differing attributes, for example, time of arrival and angle of arrival. The merged features then share the same cluster centroids. In this paper, we consider datasets with only RSS measurements. Under the assumption of equivalent APs, all features can be merged for the clustering.

Next, the number of clusters is calculated from the initial RSS data. In this paper, we calculate the number of clusters for each dataset (or for each group of features in case of more than one group of merged features) by linearising the two dimensions of the radio map (samples and features) and applying formula shown in Eq. 1 to all RSS single values.

$$K = \sqrt{unique(X(:))} \tag{1}$$

Where K refers to the number of clusters, $X(:)$ refers to the whole radio map reshaped into a single vector and $unique()$ is a function finding the number of unique values in its input.

Based on the required compression ratio (CR) and tolerance of the method, the number of clusters can be adjusted. The k-means clustering is then applied to the dataset, resulting in each feature of each sample (every single measurement) being assigned to a single cluster. The dataset is then stored as the set of cluster indexes, instead of the measured values. Along with the clustered dataset, the centroid coordinates and the number of RSS measurements in each cluster are recorded and stored. Due to the limited number of clusters, each dataset entry can be stored using a significantly reduced number of bits, as shown in Eq. 2, instead of using e.g. 64 bits, as is the standard for double floating point format. The table that converts cluster indexes to coordinates and the array with the number of measurements assigned to each cluster must be stored as the necessary overhead.

$$n_{bits} = ceil(\log_2 K) \tag{2}$$

Where n_{bits} refers to the minimum number of bits required to store one feature of the sample and $ceil$ function rounds a number up to the nearest integer, if necessary.

B. Second Stage

In the first stage, the initial dataset was clustered and compressed. The second stage of the method is fed with a set of independent fingerprints not used in the first stage, e.g. samples from testing set or newly measured ones. In the second stage of the algorithm, the new sample is obtained and processed, to be added to the existing dataset. After the assignment of the features to clusters based on the given distance metric, the sample is added to the existing compressed dataset. Next, each cluster centroid and its count, assigned to the sample are updated as shown in Eq. 3 and Eq. 4.

$$C_n^{i+1} = \frac{C_n^i \cdot N_n^i + s_m^t}{N_n^i + 1} \tag{3}$$

$$N_n^{i+1} = N_n^i + 1 \tag{4}$$

where i and $i+1$ refer to the current and the following iteration, respectively, C_n refers to the n^{th} cluster's centroid coordinate, N_n refers to the n^{th} cluster's count in the dataset and s_m^t refers to the t^{th} sample's value of the m^{th} feature. In other words, each new sample's feature updates its assigned cluster's centroid coordinate based on its distance from the last centroid coordinate and number of features assigned to that cluster.

The centroid updates enable the dataset to best represent all the assigned data, rather than only the initial dataset, as would be the case without the updates. For computational efficiency, the updates can be performed in batches, instead of with each sample. The algorithmic description is described in Algorithms 1 and 2, whereas the workflow is depicted in Fig. 1.

Algorithm 1: First Stage

1: Load the initial dataset
2: Calculate the number of clusters (as in Eq. 1)
3: Apply k-means clustering to all RSS measurements in the dataset
4: Count the number of RSS measurements in each cluster
5: Save the clustered dataset, centroid coordinates and RSS measurement counts per cluster

Algorithm 2: Second Stage

1: Acquire new sample
2: Calculate the distances of the sample's RSS measurements to each centroid
3: Find the closest centroids and cluster the sample
4: Update the centroid coordinates and RSS measurement counts of the used clusters (as in Eq. 3 and Eq. 4)
5: Add newly labeled sample to the dataset

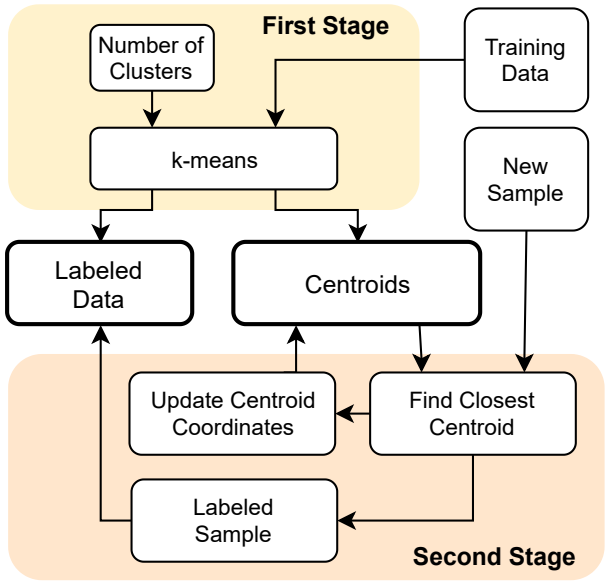

Fig. 1. Workflow of the proposed method

To summarize, the proposed method enables the efficient dataset compression using a reduced "alphabet" of values, without reducing the number of samples or features themselves. The method considers the features representing the same physical entities together, reducing the required overhead of the conversion table. The trade-off between the degree of compression and data distortion due to the compression can be adjusted by increasing or reducing the number of clusters.

IV. Experiments and Results

Nowadays, it is important to fulfill three main considerations which are repeatability, replicability, and reproducibility. They are essential in the research area, while repeatability is also mentioned in the ISO/IEC 18305:2016 [20]. In this section, we provide all of the information required to reproduce the experiment. Also, the source code is available online on Zenodo [21].

A. Datasets

This work uses 16 Wi-Fi fingerprinting datasets [22] for evaluation, created by Tampere University, Finland (TUT 1&2 [23], [24], TUT 3&4 [25], TUT 5 [26] and TUT 6&7 [27]), Universitat Jaume I, Spain (UJI 1&2 [28] and LIB 1&2 [29]), University of Minho, Portugal (DSI 1&2 [30]), and University of Mannheim, Germany (MAN 1&2 [31], [32]).

The datasets consist of Wi-Fi RSS measurements in dBm in different environments. Each dataset is separated into training and testing dataset. For the purposes of this work, the training dataset was used for the first stage including k-means clustering. The testing dataset served as the source of individual samples for the second stage of the algorithm. Additionally, all datasets include position references for each sample, containing the coordinates, building and floor indexes.

B. Evaluation metrics

To evaluate the performance and viability of the proposed method, the following metrics are considered. First, the mean squared error (MSE) between the original dataset samples and the recovered dataset was calculated in two instances. MSE_{S1} evaluates the MSE between the original and recovered data from the initial dataset after the first stage. MSE_{S2} evaluates the MSE between the original of the testing dataset and its recovered version after the second stage of the algorithm. Second, δ_{MSE}, representing the difference between MSE_{S1} and MSE_{S2} as shown in Eq. 5, is utilized to evaluate the capability of the method to adapt to the new data.

$$\delta_{MSE} = MSE_{S1} - MSE_{S2} \tag{5}$$

The impact of the compression on the data quality and the amount of information it contains was evaluated by comparing the positioning accuracy based on the k-Nearest Neighbor (k-NN) classifier. Each dataset was evaluated using 10-Nearest Neighbour (10-NN) classifier both before and after compression, and the mean 3D positioning error ratio ξ_{KNN} before and after compression was calculated, as shown in Eq. 6. The same classifier (i.e., 10-NN) was used to evaluate each dataset under the same conditions and hyperparameters. Finding the optimal value for the number of considered neighbors for each dataset is outside of the scope of this paper, as the classifier only compares the performance of the uncompressed and the compressed data.

$$\xi_{KNN} = \frac{MSE_{reconstructed}}{MSE_{original}} \tag{6}$$

Where $MSE_{original}$ refers to the mean positioning error of the original dataset and $MSE_{reconstructed}$ refers to the mean positioning error of the recovered dataset, using the 10-NN classifier after the second stage. 3D positioning error ratio larger than 1 represents the increase of the positioning error, while ξ_{KNN} lower than 1 means the positioning error decreased due to the compression.

Finally, the compression ratio (CR) of the method was evaluated to reflect the efficiency of the proposed model to reduce the storage requirements of the method. It was calculated as the ratio between the original dataset size and its reduced size using optimum coding (see Eq. 2). Smartphones usually provide quantized RSS values within range $[-105, \ldots, -30]$ dBm, which can be represented with 7 bits. In case of RSS post-processing, such as averaging the measurements over a specified area in datasets TUT 1, TUT 2, TUT 5 and MAN 2, the resulting RSS are stored in double (64 bits) format.

C. Numerical results

In our experiment, k-means clustering was independently executed over 16 selected datasets to create the new alphabets. We used Eq. 1 to set the value of k and the remaining k-means' hyperparameters include, for all datasets, Euclidean distance metric, a maximum number of 100 iterations, 100 replicates and the initialization proposed in k-means++ [19]. Then, the 10-NN algorithm was executed using original datasets and the reduced ones to evaluate the proposed radio map reduction. The results are reported in Table II.

TABLE II
RESULTS COMPARISON USING DIFFERENT DATASETS

Dataset	Num. of Samples	Num. of Clusters	MSE_{S1}	MSE_{S2}	δ_{MSE}	ξ_{KNN}	CR
DSI 1	1717	8	0.568	0.566	0.003	0.992	2.33
DSI 2	924	8	0.560	0.548	0.012	1.010	2.33
LIB 1	3696	7	0.250	0.541	-0.291	0.998	2.33
LIB 2	3696	8	0.290	0.299	-0.009	1.007	2.33
MAN 1	14760	8	1.311	1.361	-0.050	0.984	2.33
MAN 2	1760	37	0.065	0.069	-0.004	0.958	10.50*
SIM	11710	7	1.004	1.000	0.004	0.932	2.33
TUT 1	1966	29	0.037	0.030	0.007	0.992	12.78*
TUT 2	760	14	0.181	0.102	0.079	1.014	15.97*
TUT 3	4648	9	0.162	0.160	0.002	0.999	1.75
TUT 4	4648	9	0.160	0.161	-0.001	0.996	1.75
TUT 5	1428	80	0.003	0.002	0.001	0.999	10.96*
TUT 6	10385	9	0.152	0.149	0.002	0.993	1.75
TUT 7	9291	9	0.094	0.097	-0.002	0.997	1.75
UJI 1	20972	11	0.107	0.092	0.014	0.979	1.75
UJI 2	26151	11	0.107	0.110	-0.003	0.979	1.75
Average			0.316	0.330	-0.015	0.989	2.04 12.55*

* 64-bit representation

First, the number of clusters vary depending on the dataset being around 7–9 in most of cases. Datasets UJI 1&2 have larger number of clusters because one device reported unusual RSS values in smartphones, above -20 dBm. Datasets with RSS post-processing have the largest number of RSS unique values as they are Real-valued numbers ($\mathbb{R}$).

The table shows the varying MSE_{S1} and MSE_{S2} values across the datasets, which are in all cases but two below 1 dB. The results also show that δ_{MSE} is less than -0.015 dB, proving the property of the method to adapt well to new data. ξ_{KNN} is 0.989 on average, which corresponds to 1% decrease in positioning error across the datasets due to compression. The results show that although the compression reduces the number of values in the dataset, the quality of the dataset for positioning purposes actually increases. A CR of 12.55 was achieved across all real-valued datasets (64-bit representation) on average and 2.04 across the integer valued datasets (7-bit representation). The repeating values of compression ratios in integer valued datasets are caused by constant ratio between the original and the reduced bit representation, as the overhead is negligible (e.g. CR of 1.75 is achieved by compressing 7-bit values into 4-bit representation). The trade-off between the CR and ξ_{KNN} (as well as all MSE metrics) can be controlled by increasing the number of clusters of the method.

D. Discussion

The authors have presented the proposed dataset compression method and validate its usability on 16 different datasets. In comparison to e.g. SAX, the method does not require the assumption of Gaussian distribution of the data, nor any prior knowledge about the data statistics. However, this fact may lead to underfitting of the dataset by choosing the number of clusters too low, and resulting in significant information loss due to the compression. The method is also vulnerable to changes in the environment, which is the common problem of the fingerprinting datasets, as the changes in sample distributions will lead to the decrease of accuracy. In such cases, the new initial dataset should be created as the original samples from the first dataset do not reflect the reality anymore.

The significant differences in the number of clusters between the datasets are caused mostly by dataset post-processing of TUT 1, TUT 2, TUT 5 and MAN 2 datasets due to the larger number of unique values in them [22]. It is also worth mentioning, that the datasets LIB 1&2 contain measurements from the same area and device, measured 10 months apart. Despite this, each of the datasets got assigned different number of clusters, probably caused by rounding of the $ceil()$ function in the first stage. Also, the presence of outlier devices providing untrusted RSS values might degrade the IPS. Regulating the automatic selection of the number of clusters will be studied and improved in the future.

The authors acknowledge the lack of the validation set in the datasets, which will be added later as the paper presents the results of the preliminary study.

Future work will also concentrate on thorough comparison of the method with the current state-of-art methods, as well as combining the feature-reduction methods such as PCA or AE with the proposed one, to further increase the compression efficiency without losing positioning accuracy.

V. CONCLUSIONS

This paper explores a novel approach on clustering of RSS datasets for RSS-based indoor positioning on wearables, towards more energy efficient solutions. It introduces a new and efficient method of data compression based on k-means clustering and substitution of RSS measurements with a reduced alphabet representation.

The developed compression method allows for optimizing the storage space used by the WiFi fingerprinting datasets, resulting in reduced computational load on the online phase of this positioning technique. The proposed method achieved significant dataset compression, as well as slightly improved the accuracy of the position estimation (see Section IV B). As a result, the proposed method acquired a CR of 12.55 in the real-valued datasets, 2.04 in the integer-valued datasets, and the positioning error was reduced by 1% on average.

Finally, the paper discusses the shortcomings of the current method, highlighting the challenge of automatic selection for the number of clusters. In future work, this method will be compared with other existing compression methods in order to test the efficiency and robustness of the proposed work.

REFERENCES

[1] O. Belmonte Fernández, A. Puertas-Cabedo, J. Torres-Sospedra, *et al.*, "An indoor positioning system based on wearables for ambient-assisted living," *Sensors*, vol. 17, p. 36, Dec. 2016.

[2] G. Aroganam, N. Manivannan, and D. Harrison, "Review on wearable technology sensors used in consumer sport applications," *Sensors*, vol. 19, p. 1983, Apr. 2019.

[3] M. D'Aloia, A. Longo, G. Guadagno, *et al.*, "Iot indoor localization with ai technique," in *2020 IEEE International Workshop on Metrology for Industry 4.0 IoT*, 2020, pp. 654–658.

[4] Y. Zhong, F. Wu, J. Zhang, *et al.*, "Wifi indoor localization based on k-means," Jul. 2016, pp. 663–667.

[5] A. Abusara, M. Hassan, and M. Ismail, "Rss fingerprints dimensionality reduction in wlan-based indoor positioning," Apr. 2016, pp. 1–6.

[6] I. A.-Q. Ahmed Abed, "Rss-fingerprint dimensionality reduction for multiple service set identifier-based indoor positioning systems," *Applied Sciences*, vol. 9, no. 15, p. 3137, 2019.

[7] N. Brunelle, G. Robins, and A. Shelat, "Compression-aware algorithms for massive datasets," in *2015 Data Compression Conference*, 2015, pp. 441–441.

[8] A. R. Dargazany, P. Stegagno, and K. Mankodiya, "WearableDL: Wearable Internet-of-Things and Deep Learning for Big Data Analytics—Concept, Literature, and Future," *Mobile Information Systems*, vol. 2018, G. De Pietro, Ed., p. 8 125 126, 2018.

[9] T. J. Gallagher, B. Li, A. G. Dempster, *et al.*, "A sector-based campus-wide indoor positioning system," in *2010 International Conference on Indoor Positioning and Indoor Navigation*, IEEE, 2010, pp. 1–8.

[10] A. Abed and I. Abdel-Qader, "Rss-fingerprint dimensionality reduction for multiple service set identifier-based indoor positioning systems," *Applied Sciences*, vol. 9, p. 3137, Aug. 2019.

[11] P. Senin and S. Malinchik, "Sax-vsm: Interpretable time series classification using sax and vector space model," in *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 1175–1180.

[12] J. Lin, E. Keogh, L. Wei, *et al.*, "Experiencing sax: A novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, pp. 107–144, Aug. 2007.

[13] D. W. Blalock, S. Madden, and J. V. Guttag, "Sprintz: Time series compression for the internet of things," *CoRR*, vol. abs/1808.02515, 2018. arXiv: 1808.02515.

[14] G. Baldini, R. Giuliani, G. Steri, *et al.*, "The application of the symbolic aggregate approximation algorithm (sax) to radio frequency fingerprinting of iot devices," in *2017 IEEE Symposium on Communications and Vehicular Technology (SCVT)*, 2017, pp. 1–6.

[15] Q. Doan, A. S. M. Kayes, W. Rahayu, *et al.*, "Integration of iot streaming data with efficient indexing and storage optimization," *IEEE Access*, vol. 8, pp. 47 456–47 467, 2020.

[16] J. Azar, A. Makhoul, R. Couturier, *et al.*, "Robust iot time series classification with data compression and deep learning," *Neurocomputing*, vol. 398, pp. 222–234, 2020.

[17] A. Cramariuc, H. Huttunen, and E. S. Lohan, "Clustering benefits in mobile-centric wifi positioning in multi-floor buildings," in *2016 International Conference on Localization and GNSS (ICL-GNSS)*, 2016, pp. 1–6.

[18] A. Saxena, M. Prasad, A. Gupta, *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[19] A. David, "Vassilvitskii s.: K-means++: The advantages of careful seeding," in *18th annual ACM-SIAM symposium on Discrete algorithms (SODA), New Orleans, Louisiana*, 2007, pp. 1027–1035.

[20] ISO, "Information technology – Real time localization System – test and evaluation of localization and tracking systems," International Organization for Standardization, Geneva, CH, Standard, Nov. 2016.

[21] L. Klus et al. (2020). Supplementary materials for "RSS fingerprinting dataset size reduction usingfeature-wise adaptive k-means clustering". version v1, 01.09.2020, [Online]. Available: https://doi.org/10.5281/zenodo.4026370.

[22] J. Torres-Sospedra, P. Richter, A. Moreira, *et al.*, "A comprehensive and reproducible comparison of clustering and optimization rules in wi-fi fingerprinting," *IEEE Transactions on Mobile Computing*, 2020.

[23] A. Razavi, M. Valkama, and E.-S. Lohan, "K-means fingerprint clustering for low-complexity floor estimation in indoor mobile localization," in *IEEE Globecom Workshops (GC Wkshps)*, 2015.

[24] A. Cramariuc, H. Huttunen, and E. S. Lohan, "Clustering benefits in mobile-centric WiFi positioning in multi-floor buildings," in *2016 International Conference on Localization and GNSS*, 2016.

[25] E.-S. Lohan, J. Torres-Sospedra, H. Leppäkoski, *et al.*, "Wi-fi crowdsourced fingerprinting dataset for indoor positioning," *MDPI Data*, vol. 2, no. 4, Oct. 2017.

[26] P. Richter, E. S. Lohan, and J. Talvitie. (Jan. 2018). WLAN (WiFi) rss database for fingerprinting positioning, [Online]. Available: https://zenodo.org/record/1161525.

[27] Lohan. (May 2020). Additional TAU datasets for Wi-Fi fingerprinting-based positioning. version v1, 11.05.2020, [Online]. Available: https://doi.org/10.5281/zenodo.3819917.

[28] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proceedings of the Fifth Conference on Indoor Positioning and Indoor Navigation*, 2014, pp. 261–270.

[29] G. M. Mendoza-Silva, P. Richter, J. Torres-Sospedra, *et al.*, "Long-term wifi fingerprinting dataset for research on robust indoor positioning," *Data*, vol. 3, no. 1, 2018.

[30] A. Moreira, I. Silva, and J. Torres-Sospedra, *The DSI dataset for Wi-Fi fingerprinting using mobile devices*, version 1.0, Zenodo, Apr. 2020.

[31] T. King, S. Kopf, T. Haenselmann, *et al.*, *CRAWDAD dataset mannheim/compass (v. 2008-04-11)*, Downloaded from https : / / crawdad.org/mannheim/compass/20080411, Apr. 2008.

[32] T. King, T. Haenselmann, and W. Effelsberg, "On-demand fingerprint selection for 802.11-based positioning systems," in *2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks*, Jun. 2008, pp. 1–8.