Analyzing the Performance of Radiolocalization Algorithms using Data Augmentation

Matheus R. B. Godinho and Daniel C. Cunha

***Abstract*— This work aims to investigate and compare the impact of different synthetic data generation techniques on the performance of fingerprint-based localization models. We conducted experiments on two databases, exploring conditional, non-conditional, selective, and non-selective synthetic data generation methods. Three machine learning-based localization models were utilized, resulting in a total of 90 models being trained—some using only real data while others incorporated synthetic data. The results indicate that synthetic data generation can enhance the performance of machine learning prediction models, particularly for those based on support vector regression. Additionally, the conditional and selective generation methods outperformed their non-conditional and non-selective counterparts.**

***Keywords*— Radiolocalization, fingerprinting, machine learning, adversarial neural networks, data augmentation.**

I. Introduction

Location-based services for mobile users have become essential, offering numerous benefits and business opportunities for society. The growing prevalence of mobile and wireless devices underscores the importance and potential of these services. Nevertheless, their effectiveness relies heavily on the accuracy of mobile position estimation and considerations of energy consumption and implementation costs [1].

The popularity of satellite-based navigation systems, such as the global positioning system (GPS), is undeniable [2]. These systems can accurately estimate the position of objects within tens of meters, which meets the requirements of various location-based services [3]. However, GPS has limitations in indoor environments and high-density urban areas. In this context, location solutions based on radio frequency (RF) signals present viable alternatives, offering a good cost-benefit ratio [4], [5].

The fingerprinting (FP)-based radio localization algorithm is a pattern recognition technique that serves as a major alternative to GPS. Moreover, machine learning (ML) models have frequently been utilized in their implementation [6], [7]. The FP-based technique involves two steps: the offline step and the online step. During the offline step, parameters of RF signals, such as received signal strength (RSS), are collected in a wireless communication network (namely Wi-Fi, Bluetooth, and cellular networks) and associated with specific positions in a coordinate system. Each record, a combination of measured signal levels and location coordinates, is called a reference fingerprint. These reference fingerprints are stored in a database known as a radio map and are utilized to train ML models. At the end of the training, the ML models are used in the online stage to determine the location of a mobile device based on the signal levels received at its antenna, commonly referred to as the target fingerprint.

Matheus R. B. Godinho and Daniel C. Cunha, Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife-PE, e-mails: [mrbg, dcunha]@cin.ufpe.br.

Although the FP-based localization technique performs well, collecting reference fingerprints can be costly. ML models need a significant amount of data for effective training and optimization [8]. Additionally, the radio map requires frequent updates due to changes in the RF signal propagation environment, particularly in large-scale and long-term deployments [9]. Finally, while localization systems offer many advantages, there is also a risk to users' privacy when their data is accessed, such as when reference fingerprints are collected through *crowdsourcing* [10], [11].

Data augmentation has been applied to solve the problem of the high cost of collecting training data for FP-based radiolocation systems that employ ML models [12], [13]. In other words, including synthetic data generated from actual data collected allows the training of localization algorithms to be performed with a smaller amount of collected data, reducing the costs associated with obtaining measurements. Additionally, reducing the number of real measurements helps preserve user data privacy. Given the above, this work aims to evaluate how different synthetic data generation methods impact the performance of FP- and ML-based radio localization algorithms, considering signal measurement scenarios in indoor and outdoor environments.

The remainder of this paper is structured as follows: Section II introduces a background of the main concepts and techniques applied. Section III presents the framework and experimental setup designed and followed in this paper. In Section IV, the results gathered are shown and discussed. Finally, Section V summarizes the main conclusions of this work.

II. Background

A. Hierarchical Clustering Algorithm

Hierarchical cluster analysis (HCA) is a clustering algorithm that groups similar objects from a dataset into subsets called clusters [14]. The primary objective of HCA is to create clusters where the elements within each cluster are highly similar, while elements from different clusters are as dissimilar as possible.

The algorithm operates in a bottom-up, agglomerative manner. It starts by treating each data point as an individual cluster. Then, it follows two main steps iteratively. First, it identifies the pair of clusters that are the most similar based on a selected distance metric and linkage method. Next, it merges these two clusters into a single new cluster. This process continues until

all data points are combined into one overarching cluster. The progression of this hierarchical merging is visually represented by a dendrogram, which is a tree-like diagram that illustrates the nested grouping of clusters and the distances at which each merge occurs.

The dendrogram is a visualization tool that helps researchers determine the optimal number of clusters to extract from their data. By choosing a cutoff point along the vertical axis of the dendrogram, they can decide at what level of similarity to separate the data into distinct clusters. In this context, HCA is used to categorize a set of mapped locations in the test dataset into distinct zones [15]. These zones provide a framework for the conditional generation of synthetic data, enabling the model to effectively capture and replicate location-specific patterns.

B. Conditional Tabular Generative Adversarial Networks

Generative adversarial networks (GANs) are a well-established method of two adversarial components for training generative models [16]. The first one is a generative model G that learns to capture the underlying data distribution. The second component is a discriminative model D that estimates the probability that a given sample originates from the training data rather than G. Both G and D can be represented as non-linear mapping functions, such as multi-layer perceptrons.

To learn a generator distribution p_g over data x, the generator creates a mapping function $G(z;\theta_g)$ from a prior noise distribution $p_z(z)$ to the data space. The discriminator $D(x;\theta_d)$ outputs a single scalar value representing the probability that x comes from the training data rather than the generator distribution p_g.

GANs can be expanded to a conditional model, usually named as conditional tabular GANs (CT-GANs). This occurs if the generator and discriminator are conditioned on additional information y, which can be any auxiliary information, such as class labels [17]. We can perform the conditioning by providing y as an additional input layer to the discriminator and the generator. Throughout the training, we adjust the parameters for G and D to minimize the expression $\log(1-D(G(z))$ and $\log D(x)$, respectively. This process resembles a two-player min-max game with a value function $V(G,D)$, such that

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z|y))] \ . \quad (1)$$

III. Framework and Experimental Setup

This section will describe the proposed framework and the experiments carried out in this research.

A. Framework Description

Fig. 1 illustrates the complete RF FP-based localization system framework using ML and data augmentation. The initial step of the offline stage involves collecting real data in the field. RF FP-based localization techniques typically create radio maps from RSS measurements gathered from multiple base stations (BSs) or access points (APs) within a designated coverage area [18]. The FP-based technique collects RSS measurements at known reference positions throughout the coverage area to build a radio map. Each reference point is associated with an RSS vector, representing the signal strength received from each BS or AP. These RSS-position pairs can then be used directly to train a regression model that estimates the user's position from a given RSS fingerprint.

After collecting the RSS measurements, we perform cleaning and pre-processing procedures to adjust the dataset's dimensionality for training and generating synthetic data using a CT-GAN architecture. Before submitting the cleaned and preprocessed data for CT-GAN training, we apply the HCA to split the collected data into zones based on the coordinates of the RSS measurements.

Succeeding the CT-GAN synthetic data generation, we can then apply a selection filter to the output data. This filter works by keeping only the generated fingerprints that resemble the original data the most, based on Euclidean distance. This selection aims to ensure that the synthetic data used in the prediction model's training does not negatively impact its ability to predict locations accurately [13].

Having introduced the elements of the data augmentation process, we will outline the types of synthetic data generation that can be employed. The first type is called *indiscriminate generation*. Indiscriminate synthetic data generation produces synthetic data without considering the zones created before the CT-GAN's training. In other words, we effectively use only block 2 of the data augmentation process (see Fig. 1).

The second type is known as *stratified generation*. This method involves creating synthetic data that maintains the same proportion of samples per zone found in the original dataset. We believe that generating synthetic data in this way, aligned with the original distribution, could be beneficial for training localization models. This approach prevents excessive synthetic data from being added to poorly represented zones. Theoretically, if models are trained mostly on synthetic data from these zones, it could hinder their ability to predict in those areas correctly. In this case, blocks 1 and 2 are used in the data augmentation process.

The third type of synthetic data generation is called *balanced generation*. This approach creates data to ensure that each zone has equal data added to the original training set. The aim is to determine whether balancing the number of samples for each zone and providing equal representation during the training of the prediction model helps it generalize better and avoid bias towards zones with a majority of data. This method is similar to stratified generation, utilizing blocks 1 and 2.

After each type of synthetic data generation mentioned so far, in block 3, a selective filter may be applied, resulting in an additional type of generation called *selective generation*. For this generation, the number of synthetic samples to be initially generated was increased by a factor of 20. Then, a selection method was applied to retain only the best synthetic samples, reducing them to the desired final quantity.

The resulting radio map will then be used to train the ML prediction model, which will be able to estimate the mobile device's location.

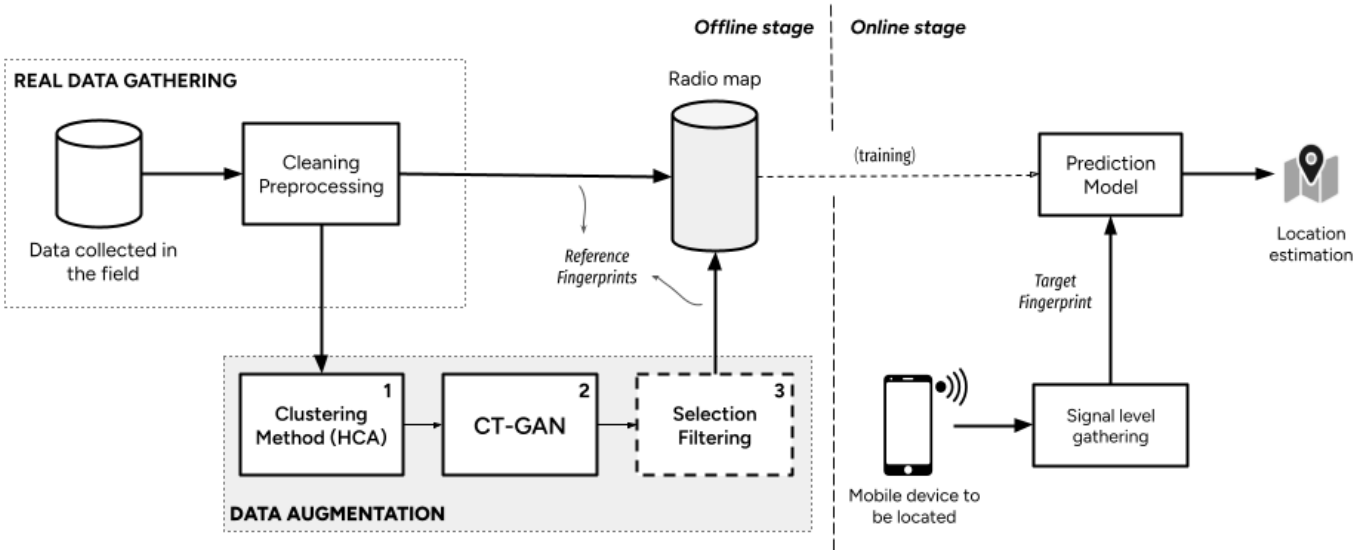

Fig. 1. Framework of the radio frequency (RF) fingerprinting (FP)-based localization techniques using machine learning (ML) and data augmentation.

B. Experimental Setup

In this study, we conducted four types of experiments to examine the effect of synthetic data generation on the training of predictive models. To this end, we considered the use of two RSS measurement datasets.

The first dataset utilized in this research was gathered in the vicinity of the UFPE [18]. This dataset consists of 6,775 samples that represent cellular network RSS measurements taken from the university's external and internal environments. In this dataset, we selected only the data collected in outdoor environments, as the indoor data was abundant but limited to a few locations, which could introduce bias during model training. Consequently, we created a subset consisting of 2,154 unique outdoor samples. From this subset, we allocated 90% of the data for training and 10% for testing, ensuring that the CT-GAN received sufficient data for effective training.

The second dataset was UJIIndoorLoc, shortly called UJI, a database of Wi-Fi RSS measurements from various buildings and floors of Universitat Jaume I [19]. This dataset contains 19,937 samples for training and 1,111 samples for testing and validation collected from 520 APs distributed throughout the university. We exclusively selected samples from the ground floor to simplify the prediction problem to two dimensions. This choice reduced the number of training samples to 4,305 and the test samples to 132. The training set was further reduced to 1,188 records through random sampling to achieve a proportion of 90% training data and 10% test data, to match the proportions of the UFPE dataset. Additionally, it was necessary to reduce the dimensionality of the dataset to enable the training and generation of synthetic data using the CT-GAN. Therefore, the principal component analysis algorithm was applied to the RSS columns, lowering the dimensionality from 520 to 12.

After completing the preprocessing pipeline, the HCA algorithm was applied to partition the datasets into spatial zones, inspired by the approach used by [15]. This clustering was based solely on the geographic coordinates (latitude and longitude) from the training data, which were first normalized. A dendrogram was created to assist in selecting an appropriate cutoff point for cluster formation. For the UFPE dataset, a Euclidean distance threshold of 2 resulted in the creation of 42 zones, which corresponded to the merging of approximately 97.88% of clusters. Following the same methodology, the UJI dataset was processed, where a distance threshold of 1.27 produced 27 zones that retained a similar percentage of cluster merges.

Two CT-GAN models were trained, one for each dataset, over the course of 7,000 epochs using zone information as a categorical variable to enable conditional data generation. The quality of the synthetic data was evaluated using three metrics: Kullback-Leibler divergence, Wasserstein distance, and Fréchet inception distance [20]. These metrics assess the similarity between the distributions of real and synthetic data. Low values for these metrics indicated that the models successfully generated realistic synthetic data.

After training the CT-GAN models, we conducted experiments on generation and filtering techniques. These strategies were tested by adding synthetic data to the training sets in three different proportions: 20%, 40%, and a dataset-specific maximum based on balanced generation—approximately 54% for the UJI dataset and 62% for the UFPE dataset. The resulting datasets were then used to train prediction models, allowing us to analyze how each generation strategy affected model accuracy and generalization. Fig. 2 shows examples of each type of generation method for specific zones within the UFPE dataset. As expected, balanced generation equalizes sample counts across zones, while stratified generation preserves the original distribution, and indiscriminate generation does not follow any particular pattern regarding zones.

The predictor models utilized in this study—MLP, SVR, and XGBoost—were trained with hyperparameter optimization using the Optuna library. This process involved five-fold cross-validation, with the mean of the negative mean squared error serving as the objective function. After identifying the optimal hyperparameters, a total of 90 models were trained: six using only real data, while the remaining models were trained with

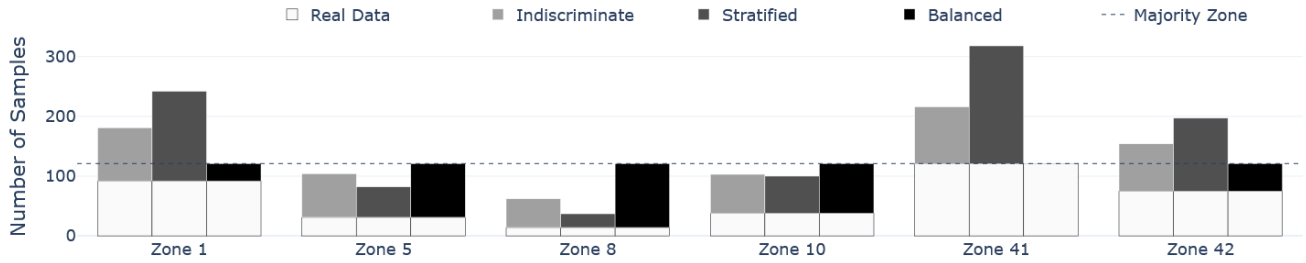

Fig. 2. Example of each type of generation method for specific zones within the UFPE dataset.

synthetic data generated for each type of experiment. Model evaluation focused on the distance error in meters, which was calculated using either the Euclidean distance or the Haversine formula, depending on the coordinate system of each dataset.

IV. Results

This section investigates the impact of incorporating synthetic data into model training by addressing four key questions: (a) how models trained with both real and synthetic data compare to those trained solely on real data; (b) how different proportions of synthetic data influence performance; (c) which type of synthetic data generation—indiscriminate, stratified, or balanced—yields the best outcomes; and (d) whether applying a selection filter to synthetic samples enhances results.

Tab. I presents the mean value $\bar{\epsilon}$ and standard deviation σ of the distance prediction error for each model and dataset, comparing the baseline experiments, trained solely with real data, with the best generative ones, trained with real and synthetic data. Only 12 models had their results presented due to space constraints. For brevity, the experiments have been given shortened names. The first letter of each shortened name indicates the generation type: 'S' for stratified, 'I' for indiscriminate, and 'B' for balanced. The second letter signifies whether the selective filter was applied, with 'S' indicating it was applied and 'NS' indicating it was not. Lastly, the number represents the percentage of synthetic data in the resulting training set. The *Rank* column in Tab. I shows how each experiment performed compared to the others from the same dataset and model, ordered by the mean distance prediction error, measured in meters.

As shown in Tab. I, several configurations utilizing synthetic data outperformed their respective baselines. This indicates that performance can improve with careful tuning of the generation type, the proportion of synthetic data, and the filtering methods used. When we extend our analysis to all 90 experiments, the proportion of synthetic data demonstrates model-specific effects. Moderate levels (e.g., 40%) occasionally outperformed both lower and higher proportions, suggesting that finding a balance is crucial for enhancing generalization without introducing excessive noise. Furthermore, when comparing different generation types, balanced and stratified approaches generally outperformed indiscriminate methods, with balanced generation achieving the best results in most comparisons.

TABLE I
Results for each ML model and dataset comparing the baseline with the best generation experiment.

Dataset	Model	Experiment	$\bar{\epsilon}$[m]	σ[m]	Rank
UFPE	SVR	Baseline	122.88	53.83	13
		S-S-62	103.28	58.91	1
	MLP	Baseline	95.38	53.00	2
		I-S-62	91.75	62.13	1
	XGBoost	Baseline	23.59	18.64	1
		B-S-62	25.41	19.22	2
UJI	SVR	Baseline	39.11	27.31	4
		B-S-54	38.50	27.85	1
	MLP	Baseline	26.33	24.03	2
		I-S-20	24.99	17.59	1
	XGBoost	Baseline	14.92	12.06	2
		B-S-54	14.15	10.24	1

However, these performance gains were not consistent across the board, and optimal strategies varied depending on the context.

Additionally, selective filtering improved model performance in roughly 69% of cases, with zone-aware selection proving particularly beneficial. While both balanced and stratified approaches consistently benefited from filtering, indiscriminate generation showed less consistent improvement, sometimes introducing spatial biases that hampered generalization. This highlights the importance of pairing synthetic generation with spatially informed selection to preserve distributional integrity.

One way to evaluate the effectiveness of localization techniques is by assessing their precision using the cumulative distribution function (CDF) of the distance prediction error [21]. Fig. 3 illustrates the CDFs of the distance prediction errors for the localization techniques using SVR and XGBoost, based on the UFPE and UJI datasets, respectively. A steeper curve and, to a lesser extent, a shorter range in the CDF indicate better precision of the localization system. From this analysis, we can conclude that the precision of the synthetic-based approaches is superior to that of the baseline, particularly for the SVR-based models in the UFPE dataset, but also for the XGBoost-based models in the UJI dataset, due to a shorter range of distance errors, even though the slope of the lines is similar.

Overall, synthetic data can enhance model accuracy and robustness, but its benefits depend heavily on the strategy used, making it crucial to tailor the generation and selection processes to each specific application.

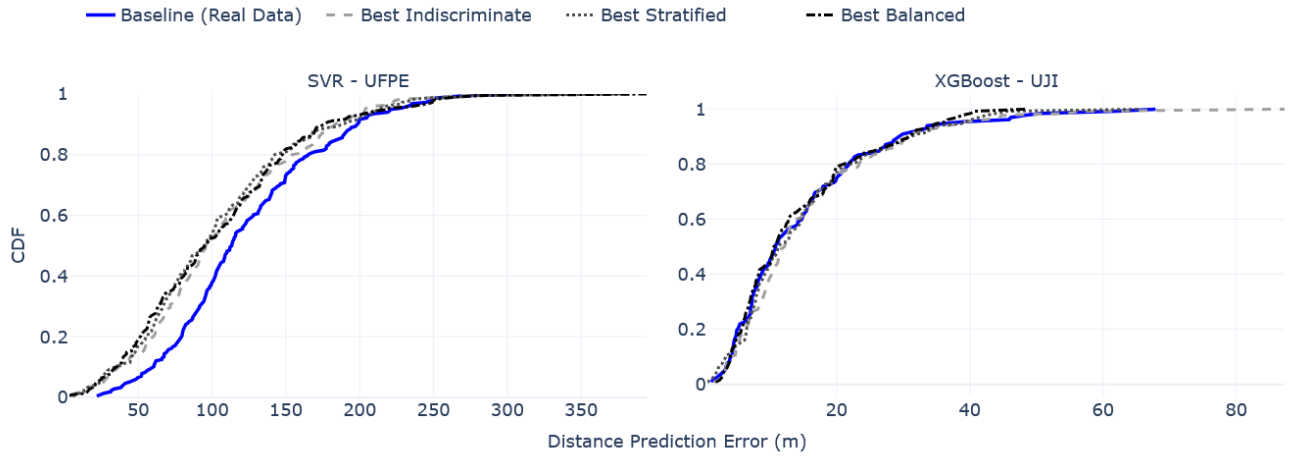

Fig. 3. Cumulative distribution function of the average distance prediction error considering the baseline ad the best types of synthetic generation for both datasets.

V. Conclusions

This study investigated the impact of different CT-GAN synthetic data generation strategies on the effectiveness of a fingerprint-based localization technique using machine learning models. Experiments were conducted using a dataset of outdoor fingerprint samples collected at UFPE and a Wi-Fi RSS measurement database named the UJI dataset. The analysis employed three predictive models: multilayer perceptron (MLP), support vector regression (SVR), and XGBoost. Furthermore, location zones defined through hierarchical clustering analysis (HCA) were used as conditioning variables in the synthetic data generation process.

Four generation strategies were tested: `(a)` indiscriminate generation, which ignores zone structure; `(b)` stratified generation, which reflects sample density per zone; `(c)` balanced generation, which equalizes the number of samples across zones; and `(d)` selective generation, where synthetic data from the previous methods is filtered to retain only the most realistic samples. The goal was to evaluate how these strategies and the proportion of synthetic data added affect prediction accuracy. Model performance was compared across different scenarios to assess each method's relative effectiveness and filtering impact.

The results indicated that synthetic data can enhance model performance, especially for SVR-based models. Among the various generation techniques, balanced generation generally outperformed both stratified and indiscriminate methods. Additionally, applying selection filters improved the quality of the synthetically generated data when using zone-based approaches. However, filtering data generated indiscriminately could lead to worse outcomes due to the introduction of spatial bias. Future research could investigate these techniques with smaller training datasets or utilize newer model architectures, such as transformers or convolutional neural networks, to assess their broader applicability.

References

[1] L. Wang et al., "Energy efficiency on location based applications in mobile cloud computing: A survey", *Computing*, v. 96, pp. 569–585, 2014.

[2] C. J. Hegarty, *The Global Positioning System (GPS)*. Springer, pp. 197–218, 2017.

[3] D. Dao et al., "Location-based services: Technical and business issues", GPS Solutions, v. 6, pp. 169–178, 2002.

[4] A. -M. -S. Chong et al., "Integration of UWB RSS to Wi-Fi RSS fingerprinting-based indoor positioning system", *Cogent Eng.*, 9: 2087364, pp. 1–23, 2022.

[5] X. Feng et al., "A Wi-Fi RSS-RTT indoor positioning model based on dynamic model switching algorithm", *IEEE J. Indoor and Seamless Position. Navig.*, v. 2, pp. 151–165, 2024.

[6] X. Zhu, et al., "Indoor intelligent fingerprint-based localization: Principles, approaches and challenges", *IEEE Commun. Surv. Tutor.*, v. 22, n. 4, pp. 2634–2657, 4th quarter 2020.

[7] N. Singh et al., "Machine learning based indoor localization using Wi-Fi RSSI fingerprints: An overview", *IEEE Access*, v. 9, pp. 127150 – 127174, 2021.

[8] R. Shahbazian et. al., "Machine learning assists IoT localization: A review of current challenges and future trends", *Sensors*, 23, 3551, pp. 1–20, 2023.

[9] X. Liu et. al., "An adaptive fingerprint database updating method for room localization", *IEEE Access*, v. 7, pp. 42626 – 42638, 2019.

[10] X. Du et. al., "CRCLoc: A crowdsourcing-based radio map construction method for WiFi fingerprinting localization", *IEEE Internet Things J.*, v. 9, n. 14, pp. 12364–12377, 2022.

[11] A. Fathalizadeh et al., "Indoor location fingerprinting privacy: A comprehensive survey", *arXiv preprint arXiv:2404.07345*, 2024.

[12] M. Nabati et al., "Using synthetic data to enhance the accuracy of fingerprint-based localization: A deep learning approach", *IEEE Sens. Lett.*, v. 4, n. 4, pp. 1-4, 2020.

[13] W. Njima et. al., "Indoor localization using data augmentation via selective generative adversarial networks", *IEEE Access*, v. 9, pp. 98337–98347, 2021.

[14] A. K. Jain et al., "Data clustering: A review", *ACM Computing Surveys (CSUR)*, v. 31, pp. 264–323, 1999.

[15] H. Grira et al., "Enhancing fingerprinting indoor positioning systems through hierarchical clustering and GAN-based CNN", *Proc. of the IEEE Symp. on Computers and Communications*, pp. 1054–1057, 2023.

[16] I. J. Goodfellow et al., "Generative adversarial nets", *Adv Neural Inf Process Syst*, v. 27, Curran Associates, Inc., 2014.

[17] M. Mirza et al., "Conditional generative adversarial nets", *CoRR*, abs/1411.1784, 2014.

[18] R. D. A. Timoteo and D. C. Cunha, "A scalable fingerprint-based angle-of-arrival machine learning approach for cellular mobile radio localization", *Comput. Commun.*, v. 157, pp. 92 – 101, 2020.

[19] J. Torres-Sospedra et. al. UJIIndoorLoc [Dataset]. *UCI Machine Learning Repository*, 2014.

[20] E. Eken, "Determining overfitting and underfitting in generative adversarial networks using Fréchet distance", *Turk J Elec Eng & Comp Sci*, v. 29, n. 3, p. 1524-1538, 2021.

[21] Federal Communications Comission, "Revision of the Commission's Rules To Ensure Compatibility with Enhanced 911 Emergency Calling Systems - CC Docket No. 94-102, RM-8143," 1997.