Machine Learning-Based Region Segmentation for Improved Wi-Fi Fingerprinting in Campus Indoor Localization

Mohamad Anas Mohamad Nour
Department of Engineering Science
University West
Trollhättan 461 32, Sweden
mohamad-anas.mohamad-nour@student.hv.se

Saleh Alshami
Department of Engineering Science
University West
Trollhättan 461 32, Sweden
saleh.alshami@student.hv.se

Rashid Ali
Department of Engineering Science
University West
Trollhättan 461 32, Sweden
rashid.ali@hv.se

***Abstract*—GPS is unreliable for indoor positioning due to signal degradation in enclosed environments. As an alternative, Wi-Fi fingerprinting-based indoor positioning systems (FPIPS), which utilize Received Signal Strength Indicator (RSSI) values, have gained popularity due to the ubiquity of Wi-Fi infrastructure and the lack of need for additional hardware. Despite their effectiveness, improving localization accuracy and reducing computational cost remain ongoing challenges. Recently, researchers have proposed the use of a regions-based KNN (RB-KNN) algorithm for Wi-Fi FPIPS. However, the formation of regions and the selection of number of regions remains still a challenging task to optimize. In this study, we explore the use of four machine learning-based clustering algorithms—K-Means, DBSCAN, Gaussian Mixture and Agglomerative for optimizing region formation for a KNN algorithm in Wi-Fi FPIPS. By identifying the most suitable clustering approach for region formation, this study aims to further enhance the performance of RB-KNN algorithm for Wi-Fi FPIPS in campus environments.**

***Index Terms*—Wi-Fi fingerprinting, indoor positioning system, indoor localization, wi-fi, machine learning**

I. Introduction

Advancements in the Internet of Things (IoT) and automation have heightened the need for accurate indoor localization, where Global Positioning System (GPS) signals are often unreliable due to attenuation and multipath effects caused by physical obstructions [1]. Alternative technologies such as Bluetooth, ultrasonic, and Wi-Fi communication are employed to address this limitation. Among these, Wi-Fi is particularly suitable due to its widespread availability in indoor environments [2]. Wi-Fi fingerprinting relies on collecting Received Signal Strength Indicator (RSSI) values from multiple access points (APs) at known reference points (RPs) to build a fingerprint database. During localization, the current RSSI readings are compared with this database to estimate position.

Researchers are continuousness working to improve the efficiency of Wi-Fi FPIPS. A previous study [3] evaluated four K-Nearest Neighbour (KNN) variants and found the Region-Based KNN (RB-KNN) algorithm to be the most accurate and computationally efficient. However, that study used predefined fingerprint regions, which may introduce errors and inefficiencies. Several studies have explored machine learning-based (ML) clustering techniques to enhance indoor positioning by reducing the search space during localization. Ren et al. [4] proposed an improved public c-means clustering method. Ramires et al. [5] and Chen [6] used signal-based clustering to improve accuracy and efficiency. Park and Rhee [7] incorporated signal clustering within KNN frameworks, and others such as Anuwatkun et al. [8] and Neyaz et al. [9] applied K-Means for better performance.

This research proposes using unsupervised machine learning algorithms—K-Means, DBSCAN, Gaussian Mixture Model (GMM), and Agglomerative Clustering—to automatically generate fingerprint regions based on RSSI patterns. These generated clusters replace manual segmentation in RB-KNN, and the algorithms are compared in terms of localization accuracy and computational cost.

- How can the optimal number of regions for RB-KNN be determined?
- Can ML-based clustering effectively define these regions?
- Which algorithm offers the best trade-off between accuracy and computational cost?

II. Machine Learning-Based Region Segmentation

To investigate the effect of ML algorithms on region segmentation in indoor positioning, a quantitative research approach was adopted. This analytical study simulates clustering on a real-world dataset from University West [3], aiming to uncover meaningful patterns among RPs to create regions and evaluate impacts on accuracy and computational efficiency.

A. Dataset Description

This study uses a previously collected Wi-Fi fingerprint dataset from [3], consisting of RSSI values (in dBm) and BSSID identifiers measured at RPs spaced one meter apart across Blocks I and J. Each RP includes signal data from six

This work was conducted as part of an undergraduate thesis project within the Wireless Networking and AI research domain at the Computer Science Research Group, University West, Sweden. The authors sincerely acknowledge the use of OpenAI ChatGPT, which contributed to improving the written content.

APs, and corresponding spatial coordinates in image space (35.7 pixels per meter). Randomly sampled test points (TPs) are used to evaluate positioning performance. Data was pre-processed to fit ML requirements, with each RP represented as a 6-feature vector.

B. Clustering Algorithms

Four unsupervised algorithms clustering algorithms were selected to segment the dataset into regions:

1) K-Means: : A centroid-based algorithm that partitions data into k clusters by minimizing intra-cluster variance. The optimal k is selected using the Silhouette score.

2) DBSCAN: : A density-based method clustering high-density RPs without needing the number of clusters in advance, defined by parameters eps and $min_samples$.

3) Gaussian Mixture Model: : A probabilistic approach modeling overlapping clusters. Expectation-Maximization (EM) and Bayesian Information Criterion (BIC) are used to fit the model and prevent overfitting.

4) Agglomerative Clustering: : A hierarchical method that merges RPs based on RSSI similarity, using Ward linkage to minimize variance within clusters.

Each model was trained on 90% of the dataset and tested on the remaining 10%, with clusters serving as regions in the RB-KNN framework.

C. Evaluation Metrics

Three metrics were used to compare clustering approaches:

- Clustering Accuracy: Correct assignment of TPs to clusters.
- Computational Cost: Time required to generate clusters.
- Positioning Accuracy: Localization performance using RB-KNN with ML-based clustered regions.

III. Results

This section presents a comprehensive evaluation of the clustering algorithms implemented in this study. Their performance is compared against the predefined RB-KNN approach proposed in a prior study [3]. The evaluation focuses on three key metrics: clustering accuracy, positioning accuracy, and computational efficiency.

A. Clustering Accuracy Evaluation

Clustering accuracy was assessed using the Silhouette score, which evaluates the compactness and separation of clusters. As detailed in Table I, the K-Means algorithm achieved the highest Silhouette score of 0.4385, indicating well-defined clusters and minimal intra-cluster variation. Agglomerative clustering followed closely with a score of 0.4238, demonstrating strong structural adaptability and cohesive group formation, particularly suited to complex indoor environments. GMM also performed adequately (Silhouette score: 0.4094), leveraging its probabilistic framework to model overlapping regions effectively. Conversely, DBSCAN yielded the lowest Silhouette score (0.3314), suggesting poor cluster formation. Its density-based logic failed to adapt to the sparsity and irregular signal distributions of the dataset, resulting in only two clusters and limited regional specificity.

TABLE I
COMPARISON OF OPTIMAL NUMBER OF CLUSTERS (K) OF K-MEANS, GMM, AGGLOMERATIVE, AND DBSCAN CLUSTERING ALGORITHMS.

Algorithm	k	Silhouette Score	Computational Time (s)
K-Means	8	0.4385	0.0469
GMM	19	0.4094	0.0628
Agglomerative	16	0.4238	0.0214
DBSCAN	2	0.3314	0.0127

B. Positioning Accuracy Evaluation

To evaluate how each clustering algorithm influenced indoor positioning, experiments were conducted using the RB-KNN framework with region definitions derived from each method. Table II reports detailed error metrics: mean error, standard deviation, and RMSE.

Agglomerative clustering achieved the most accurate results with a mean error of 4.21 meters, standard deviation of 2.54 meters, and RMSE of 4.92 meters—a 41% improvement over the baseline RB-KNN (mean error: 7.13 meters). This substantial improvement highlights the advantage of data-driven region formation in reducing localization ambiguity.

Visual analysis in Figure 1 confirms these findings: Agglomerative clustering demonstrated the steepest CDF curve, with over 70% of predictions within a 5-meter error range. GMM, while slightly less accurate (mean error: 4.34 m, RMSE: 5.27 m), performed comparably well due to its fine-grained clustering (19 clusters), capturing signal variance effectively. DBSCAN, however, showed poor positioning accuracy (mean error: 16.95 m, RMSE: 21.11 m) due to coarse clustering, as evident in its flat CDF curve and wide boxplot distribution (Figure 2).

TABLE II
COMPARISON OF MEAN ERROR (ME), STANDARD DEVIATION (SD) AND ROOT MEAN SQUARE ERROR (RMSE) IN METERS (M) OF RB-KNN (REGIONS-BASED KNN), K-MEANS KNN, GMM KNN, AGGLOMERATIVE KNN, AND DBSCAN KNN POSITIONING ALGORITHMS.

Algorithm	ME (m)	SD (m)	RMSE (m)
RB-KNN	7.13	5.28	8.87
K-means KNN	5.46	4.05	6.80
GMM KNN	4.34	2.99	5.27
Agglomerative KNN	4.21	2.54	4.92
DBSCAN KNN	16.95	12.58	21.11

C. Computational Efficiency

All clustering algorithms exhibited efficient execution times, making them suitable for periodic updates in real-time systems. Agglomerative clustering achieved the best performance among effective models (0.0214 seconds), followed by K-Means (0.0469 s) and GMM (0.0628 s). DBSCAN, despite being the fastest (0.0127 s), compromised accuracy significantly, illustrating that speed alone is insufficient for practical use. In terms of online inference, all algorithms maintained low prediction latency (8.7 milliseconds per instance), confirming their viability for real-time positioning applications.

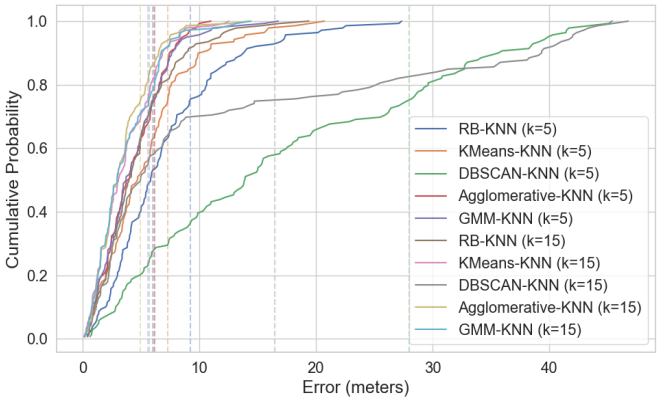

Fig. 1. Cumulative Distribution Function including 75th percentile (Q3) of accuracy errors for RB-KNN and clustering algorithms with k = 5 and k = 15.

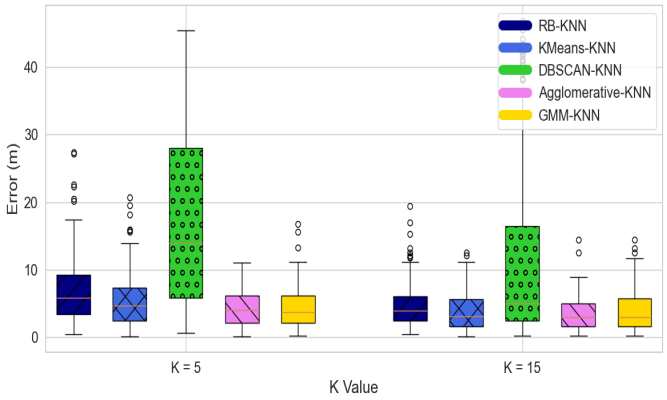

Fig. 2. Box plot showing the accuracy errors (m) of RB-KNN and clustering algorithms for k = 5 and k = 15. The plot includes average and variance values with outliers.

IV. Discussion and Analysis

This section presents a detailed interpretation and analysis of the data presented in Tables I and II, as well as the visual evidence provided in Figures 1 and 2.

A. Clustering Accuracy and Suitability for Wi-Fi Fingerprinting

As reflected in Table I, the K-Means algorithm attained the highest Silhouette score of 0.4385, indicating its effectiveness in forming compact, well-separated clusters. This result aligns with the theoretical strengths of K-Means, which assumes spherical and equally sized clusters. In this study, it produced 8 clusters, offering a moderate level of granularity suitable for many indoor environments.

Agglomerative clustering, while recording a slightly lower Silhouette score (0.4238), demonstrated significantly superior positioning performance. By forming 16 clusters via hierarchical merging, this method was better able to respect signal similarity and spatial proximity—key elements in real-world indoor localization. Crucially, Agglomerative clustering does not impose assumptions on cluster shape or size [2], making it particularly well-suited for the irregular and often non-uniform characteristics of RSSI data in indoor environments.

GMM ranked third in terms of clustering quality (Silhouette score: 0.4094) but excelled in its segmentation capabilities. By producing 19 clusters, GMM offered the finest partitioning of the dataset, with its probabilistic nature enabling the formation of elliptical and overlapping clusters. This trait proved advantageous in handling noisy or overlapping RSSI patterns commonly encountered in large-scale indoor deployments.

Conversely, DBSCAN struggled to adapt to the structure of the Wi-Fi fingerprinting dataset. It generated only 2 clusters and produced the lowest Silhouette score (0.3314), suggesting an inability to form meaningful regions. While DBSCAN had the fastest clustering time (0.0127 seconds), the resulting coarse segmentation failed to capture the spatial variability of the signal environment, ultimately leading to poor localization performance. In this case, computational efficiency came at the cost of practical utility.

B. Positioning Accuracy and Error Distribution

The impact of clustering methods on positioning accuracy is best understood through the results in Table II. Agglomerative clustering once again proved the most effective, achieving a mean positioning error of 4.21 meters, standard deviation of 2.54 meters, and an RMSE of 4.92 meters. Compared to the RB-KNN baseline (mean error: 7.13 meters), this reflects a 41% improvement, strongly validating the use of adaptive region formation.

The cumulative distribution of localization errors, shown in Figure 1, supports this conclusion. Agglomerative clustering had the steepest CDF curve, with approximately 70–75% of predictions falling below the 5-meter threshold—a critical benchmark in indoor positioning, as larger errors can span multiple rooms and severely limit usability.

GMM followed closely with a mean error of 4.34 meters, standard deviation of 2.99 meters, and an RMSE of 5.27 meters. Its higher cluster count enabled finer-grained representation of signal variation, although this came with slightly reduced precision and more outliers, as seen in Figure 2. Still, the GMM CDF curve closely tracks that of Agglomerative clustering, highlighting its robustness and real-world applicability.

DBSCAN, on the other hand, demonstrated significantly inferior positioning performance, with a mean error of 16.95 meters, standard deviation of 12.58 meters, and an RMSE of 21.11 meters. The flatness of its CDF curve and the wide distribution seen in the boxplot (Figure 2) underscore the failure of its density-based logic in capturing the necessary spatial and signal distinctions within the dataset. With only two clusters formed, DBSCAN was unable to provide any meaningful localization support.

Even the RB-KNN baseline, despite its rigid structure, outperformed DBSCAN in reliability. While not highly accurate, RB-KNN exhibited more consistent behavior, with a smoother CDF curve and tighter error distribution. This reinforces the

importance of algorithm selection tailored to data characteristics, and the limitations of generic clustering methods when applied to specialized tasks like indoor localization.

C. Computational Cost and Real-Time Applicability

In addition to accuracy, practical deployment requires that clustering algorithms be computationally efficient, making them suitable for periodic recalibration in real-time systems.

Among the top-performing methods, Agglomerative clustering offered the best trade-off, completing clustering in 0.0214 seconds, enabling rapid recalibration with minimal system overhead. K-Means and GMM also performed well, with times of 0.0469 seconds and 0.0628 seconds, respectively. These values affirm their viability for real-world use, where both accuracy and speed are critical.

Although DBSCAN recorded the fastest execution time (0.0127 seconds), this efficiency is rendered irrelevant by its substantial accuracy degradation. This result exemplifies the broader principle that execution speed must not come at the expense of effectiveness, especially in safety- or context-critical systems.

Lastly, online inference latency—the time required to compute a location prediction for each instance—remained consistently low for all algorithms, averaging around 8.7 milliseconds. This ensures that, despite differences in clustering strategies, the prediction pipeline can support real-time indoor positioning applications.

V. Conclusion

This study explored the integration of unsupervised machine learning techniques into Wi-Fi fingerprinting-based indoor positioning systems, focusing on replacing predefined region definitions in RB-KNN with dynamically generated clusters. Four clustering algorithms — K-Means, Gaussian Mixture Model, Agglomerative Clustering, and DBSCAN—were evaluated in terms of clustering structure, positioning accuracy, and computational efficiency.

Results showed that clustering quality, as measured by Silhouette score, does not directly correlate with positioning performance. While K-Means achieved the highest Silhouette score (0.4385), it was Agglomerative clustering that yielded the lowest positioning error (mean: 4.21 m), surpassing the predefined RB-KNN method by 41% in mean accuracy. GMM also demonstrated high reliability through fine-grained clustering, while DBSCAN underperformed due to its unsuitability for sparse or uneven signal patterns.

These findings underscore the importance of selecting clustering algorithms that align with the signal and spatial characteristics of indoor environments. In conclusion, ML-based region formation, especially using Agglomerative clustering, significantly improves the performance, scalability, and adaptability of Wi-Fi-based indoor positioning systems—representing a promising advancement for next-generation IPS solutions.

References

[1] GPS.gov, "Gps accuracy," accessed: Mar. 2, 2024. [Online]. Available: https://www.gps.gov/systems/gps/performance/accuracy/#how-accurate

[2] C. S. Álvarez Merino, E. J. Khatib, H. Q. Luo-Chen, A. T. Muñoz, and R. B. Moreno, "Evaluation and comparison of 5g, wifi, and fusion with incomplete maps for indoor localization," *IEEE Access*, vol. 12, pp. 51 893–51 903, 2024.

[3] R. Andersson, W. Tagesson, and R. Ali, "Dataset: Evaluating Wi-Fi Fingerprinting for Enhanced Indoor Positioning in Campus Environments," Apr. 2024. [Online]. Available: https://github.com/wirelessATwest/PEWFIPS-HV

[4] J. Ren, Y. Wang, C. Niu, W. Song, and S. Huang, "A novel clustering algorithm for wi-fi indoor positioning," *IEEE Access*, vol. 7, pp. 122 428–122 434, 2019.

[5] M. Ramires, J. Torres-Sospedra, and A. Moreira, "Accurate and efficient wi-fi fingerprinting-based indoor positioning in large areas," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–6.

[6] S. Chen, "Indoor localization based on fingerprint clustering," *Network and Communication Technologies*, vol. 5, p. 40, 2020.

[7] C. Park and S. H. Rhee, "Indoor positioning using wi-fi fingerprint with signal clustering," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, 2017, pp. 820–822.

[8] A. Anuwatkun, J. Sangthong, and S. Sang-Ngern, "A diff-based indoor positioning system using fingerprinting technique and k-means clustering algorithm," in *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2019, pp. 148–151.

[9] H. Neyaz, M. Inamullah, and M. S. Beg, "Machine learning based indoor positioning system using wi-fi fingerprinting dataset," in *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, 2024, pp. 1–5.