e-ISSN: 2618-575X

Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES
and
ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

International
Open Access

Volume 06
Issue 02

August, 2022

Research Article

Research on the success of unsupervised learning algorithms in indoor location prediction

Fatma Önay Koçoğlu [a,*]

[a]*Muğla Sıtkı Koçman University, Faculty of Engineering, Software Engineering Department, Muğla, Turkey*

ARTICLE INFO

Article history:
Received 31 March 2022
Accepted 09 June 2022
Published 15 August 2022

Keywords:
Clustering
External indices
Fuzyy c-Means
Indoor location prediction
Indoor positioning
Internal indices
k-Means
Position prediction
Unsupervised learning

ABSTRACT

With location-based smart applications, the flow of life can be facilitated and support can be provided in case of security and emergency situations. Indoor location detection provides various conveniences in complex structures such as hospitals, schools, shopping centers, etc. Indoor location detection studies are carried out by using data related to location and signal and machine learning methods. Machine learning has become frequently used as a solution method in this field, as in many other fields. When the studies in the literature are examined, it is seen that the studies are mainly focused on producing solutions with supervised machine learning algorithms. Unsupervised algorithms are frequently used to determine the labels of data groups that do not have labels. In this direction, it can be seen as the first step in labeling the data collected in indoor positioning studies and then using it for training predictive models to be developed with supervised learning methods. For this reason, the results to be obtained regarding the success and usefulness of cluster analysis will constitute an important basis for further studies. In this study, it is aimed to examine the success of unsupervised learning, in other words, clustering algorithms. The Wireless Indoor Localization Data Set and well-known k-Means and Fuzzy c-Means algorithms have been used with different distance measure. The obtained methods performances have been evaluated with internal and external indices. The results show that the clustering algorithms can cluster correctly data points in the range of 93-95% according to the accuracy and F measure value. Although performances indicators are very close to each other according to the internal indexes, it can be stated that the model obtained using the Manhattan distance measure and the k-Means algorithm has higher performance in terms of clustering success.

1. Introduction

Positioning technologies are among the important technological study topics of today. These technologies are important in terms of determining the location, enabling active monitoring and routing. Positioning techniques are applied in two areas, indoor and outdoor. Global Positioning Systems (GPS), which uses satellite data and can detect position with a very low error, is widely used today. Position estimation is performed by a GPS receiver by measuring the difference between the arrival time of the satellite signals [1]. However, there are elements that prevent or attenuate line of sight and signal transmission, such as buildings, walls, roofs. Since these factors reduce the power and efficiency of satellite and radio signals, GPS is not as effective in determining location indoors and in high-rise urban areas. On the other hand, as a result of the increase in the use of mobile devices such as mobile phones, mobile communication systems enable location detection and monitoring indoors with Global System for Mobile Communications (GSM) signals. Although they differ in scope, method and type of location, technologies used via mobile devices or equipment for indoor localization include ultrasound, infrared, Wi-Fi, Bluetooth, Zig-Bee, Ultrawide Band, inertial navigation, magnetic-based methods, and Radio Frequency Identification (RFID) [2].

With location-based smart applications, the flow of life can be facilitated and support can be provided in case of security and emergency situations. Indoor location detection provides various conveniences in complex structures such as hospitals, schools, shopping centers, etc. In applications such as directing people, especially in closed areas such as museums and airports, tracking products in areas such as factories and warehouses, following elderly patients in need of support in areas such

* Corresponding author. Tel.: +90 252 211 5581.
E-mail addresses: fonaykocoglu@mu.edu.tr
ORCID: 0000-0002-1096-9865
DOI: 10.35860/iarej.1096573

as hospitals or nursing homes, finding a specific store in shopping centers, offering location-based advertisements, detecting of abnormal situations in units where security measures are intense etc. indoor localization technologies are used. In today's world, extracting information from data and using that information is very valuable. Although location data does not make sense on its own, transforming this data into new applications with dynamically personalized content can be transformed into important profits with a small additional bandwidth usage [3]. Moreover, along with various IoT protocols such as Bluetooth and WiFi, the connection between various devices has led to the emergence of more integrated systems. So, especially with the development of the IoT, it is clear that indoor location detection will appear in more applications and will become even more important in the next days [4].

In this study, it is aimed to examine the success of unsupervised learning methods, one of the machine learning methods, in indoor location estimation. The rest of this paper is as follows in “Literature Review” studies in literature have been given which point out the problem, in “Methodology” section the data, methods, performance indicators have been detailed, in “Results” section results has been presented, and “Discussion” and “Conclusion” sections include assessment of results.

2. Literature Review

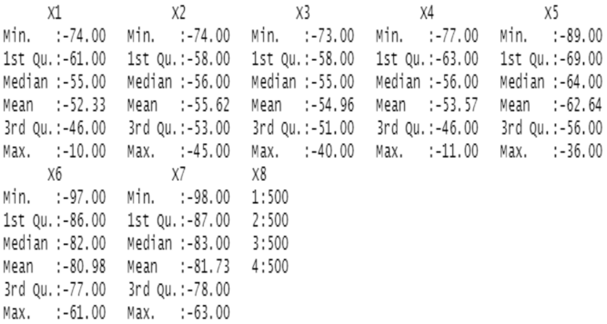
Positioning studies are carried out by using data related to location and signal and machine learning methods. Machine learning has become frequently used as a solution method in this field, as in many other fields. [5] have stated that supervised learning techniques deal with labeled data in the data collection stage of indoor localization, and she counted SVM among the most frequently used algorithms. [6] have examined the success of machine learning approaches with k-nearest neighbor (k-NN), rule-based classifier and random forest (RF) algorithms to predict indoor location using RSSI-based fingerprint method. [7] have presented an indoor positioning system (IPS) and motion tracking system for the elderly. Using the obtained data sets and Weka software tool, SVM, k-NN, RF and DT machine learning algorithms have been tested and the best classifier has been determined. [8] have presented an indoor location algorithm with the characteristics of WIFI fingerprint signals and a Naive Bayes machine learning algorithm. [9] has developed a mobile application that allows users to capture and create their own RSSI maps using the generated models to obtain the current indoor location. The models were obtained with Non-Nested Generalized Exemplars (NNge), Instance Based Learner (Ibk), Random Tree, RF and Random Committee machine learning algorithms. [10] have proposed an indoor localization approach based on fingerprints of Received Signal Strength Indicator (RSSI) measurements using Long Short-Term Memory (LSTM) Neural Networks. [11] have developed an integrated system for indoor location fingerprinting using Deep Neural Network (DNN) and improved k-NN algorithm. [12] have proposed a model for indoor localization with machine learning (ML) and deep learning (DL) algorithms in non-line-of-sight (NLoS) conditions using partial knowledge of channel state information (CSI). [13] have aimed to explore the possible improvement of system accuracy based on radio technology Bluetooth Low Energy through k-NN, Support Vector Machines (SVM), RF and Artificial Neural Network (ANN) machine learning approaches. [14] have pointed out a large number of mobile phone models causing changes in the measured received signal strength (RSS) in indoor positioning, they propose a deep learning-based system using cellular metrics to provide consistent performance in invisible tracking phones. [15] have used machine learning algorithms to developed a sensing platform consisting of a sensor toolkit with an environmental data server to provide indoor location awareness. These algorithms include k-NN, SVM, Decision Tree (DT), Adaptive Boosting, RF, Lightgbm, Xgboost, Gaussian Naive Bayes, and Gradient Boosting Classifier. [16] have developed an indoor positioning algorithm based on Back Propagation Neural Network (BPNN) to solve the low position calculation efficiency and positioning accuracy due to the complexity of indoor environments.

When the studies in the literature are examined, it is seen that the studies are mainly focused on producing solutions with supervised machine learning algorithms. In this study, it is aimed to examine the success of unsupervised learning, in other words, clustering algorithms in determining indoor location. Clustering algorithms are frequently used to determine the labels of data groups that do not have labels. In this direction, it can be seen as the first step in labeling the data collected in indoor positioning studies and then using it for training predictive models to be developed with supervised learning methods. For this reason, the results to be obtained regarding the success and usefulness of cluster analysis will constitute an important basis for further studies.

3. Methodology

3.1 Data

The Wireless Indoor Localization Data Set, which is open access in the UCI Machine Learning Repository, was used in the study [17]. The use of an open data set allows comparison with studies to be developed by different researchers. In the data set, there are 8 attributes one of which is a class attribute and 2000 records. There are 4 different class values in the dataset, which includes the signal strength of seven WiFi



```
      X1            X2            X3            X4            X5
 Min.   :-74.00  Min.   :-74.00  Min.   :-73.00  Min.   :-77.00  Min.   :-89.00
 1st Qu.:-61.00  1st Qu.:-58.00  1st Qu.:-58.00  1st Qu.:-63.00  1st Qu.:-69.00
 Median :-55.00  Median :-56.00  Median :-55.00  Median :-56.00  Median :-64.00
 Mean   :-52.33  Mean   :-55.62  Mean   :-54.96  Mean   :-53.57  Mean   :-62.64
 3rd Qu.:-46.00  3rd Qu.:-53.00  3rd Qu.:-51.00  3rd Qu.:-46.00  3rd Qu.:-56.00
 Max.   :-10.00  Max.   :-45.00  Max.   :-40.00  Max.   :-11.00  Max.   :-36.00
      X6            X7          X8
 Min.   :-97.00  Min.   :-98.00  1:500
 1st Qu.:-86.00  1st Qu.:-87.00  2:500
 Median :-82.00  Median :-83.00  3:500
 Mean   :-80.98  Mean   :-81.73  4:500
 3rd Qu.:-77.00  3rd Qu.:-78.00
 Max.   :-61.00  Max.   :-63.00
```

Figure 1. Descriptive statistics of the data

signals received by a smartphone in an indoor area, and these values indicate four different rooms. The class attribute field will not be included in the analysis, it will be used to measure clustering success after clusters are obtained. The descriptive statistics of attributes has been given in the Figure 1.

All values have been converted to values in the range of [0,1] by using the linear data transformation method as in Equation (1):

$$x_{normal\ value} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

3.2 Clustering

Classification and clustering are the basic functions of machine learning. In classification, groups must reflect some reference class. In clustering, the categories are discovered within the dataset itself [18]. Data clustering is the work of bringing together similar records, in other words generate homogeneous sub-groups, in multidimensional data and revealing relationships based on some similarity criteria and model [19]. The important thing in clustering is to bring together similar records. Therefore, the most important measure is similarity. The similarity of records in the same cluster should be maximum (maximum), while the similarity of records in different clusters should be minimum (minimum). Within the scope of the study, models were developed with k-Means, Fuzzy c-Means and algorithms.

k-Means Algorithm:

Records in the data set are assigned to a number of clusters determined by the user according to the similarity measure. In a dataset of numeric values, the measure of similarity is the distance between two data points. The distance is calculated according to the Euclidean, Manhattan and Minkowski distance formulas. The steps of the algorithm are given below [20]:

- The number of clusters is determined (k).
- Cluster centers as many as the determined number of clusters are determined randomly.
- The distance of each observation in the data set to the determined cluster centers is calculated and assigned to the cluster to which it is closest.
- Cluster centers are recalculated after all observations have been assigned.
- The third and fourth steps are repeated for the specified number of iterations, until the cluster centers do not change, and it falls below a predetermined very small threshold value.

Fuzzy c-Means:

Fuzzy c-Means is the extension of k-Means with Fuzzy logic approach. Accordingly, each data point does not belong to only one cluster. Therefore, each data point has membership degrees for the specified clusters. Clusters are determined by considering these membership degrees. The steps of the algorithm are given below [21]:

- The number of clusters (c), turbidity parameter (m), stopping criterion (Ɛ) are determined.
- Initial membership degrees are randomly determined and a membership matrix is created.
- Cluster centers are calculated.
- New membership degrees and membership matrix are calculated according to cluster centers.
- By checking the stopping criterion, the algorithm is renewed or terminated with the second step.

3.3 Distance measure:

Let each record in the data set consist of values of *n* different attributes. In this case, each record is represented by the vector $x_k = [x_{k,1}, x_{k,2}, \ldots x_{k,n}]^T$ and if the data set consists of *N* observations, the data set will be represented by $X = \{x_k |\ k = 1,2, \ldots, N\}$. Distance measures have been used for numerical data points as a measure of similarity in cluster analysis. The data set used within the scope of the study consists entirely of numerical values. So, models have been prepared by using Euclidean (Equal (2)) and Manhattan (Equal (3)) distance measurements as distance measures.

$$d_E(x_i, x_j) = \left(\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2 \right)^{1/2} \quad (2)$$

$$d_M(x_i, x_j) = \sum_{k=1}^{n} |x_{i,k} - x_{j,k}| \quad (3)$$

3.4 Performance Evaluation

Various indicators are used to test the performance of the models developed in the artificial learning process. For cluster analysis, these indicators are called internal and external indices. External indexes are also used in the measurement of success of models in which the supervised learning method is used. A confusion matrix (Table 1) is created by comparing the labels produced by the model with the actual labels of the data set. The values of the indicators are calculated over this matrix. Internal indexes are calculated regarding the similarity of cluster elements to each other as a result of clustering.

Within the scope of the study, accuracy measure from external indices and Dunn, Silhouette, Davies-Bouldin, and C index indices from internal indices have been used. The calculations of these index measures are given in Appendix.

Table 1. Confusion matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (*TP*)	False Positive (*FP*)
	Negative	False Negative (*FN*)	True Negative (*TN*)

4. Results

The results of the clustering algorithm according to accuracy measure and F score have been given in Figure 2 and Figure 3. When the results are examined, Fuzzy c-Means can estimate correctly room 4 with an accuracy value of 99.6%, regardless of the distance measure, and this value is the highest accuracy value obtained. The models developed with Fuzzy c-Means have predicted correctly room 1 with 98% accuracy and room2 with 98.6% accuracy rates and the highest success.

The accuracy values of each model have been averaged to determine the success of predicting all rooms correctly. According to the Manhattan distance, the average accuracy value obtained with the k-Means algorithm is 93.7% and the average accuracy value obtained with the Fuzzy c-Means algorithm is 95.2%. According to the Euclidean distance, the average accuracy value obtained with the k-Means algorithm is 94.3% and the average accuracy value obtained with the Fuzzy c-Means algorithm is 94.1%.

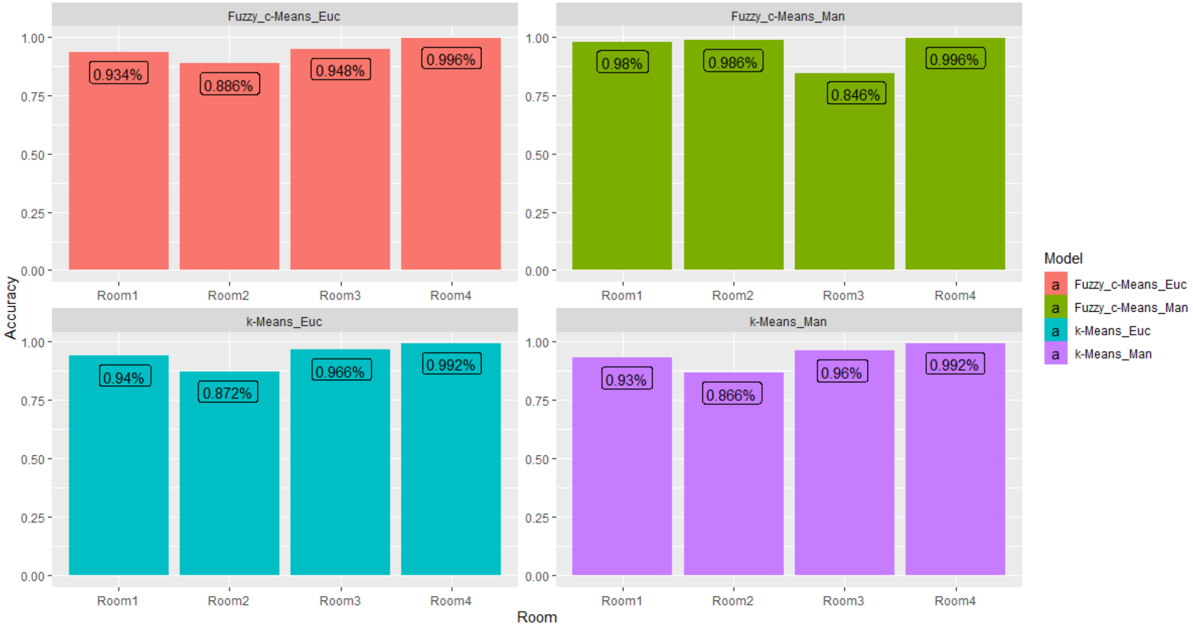


Figure 2. Results of clustering algorithms according to accuracy measure

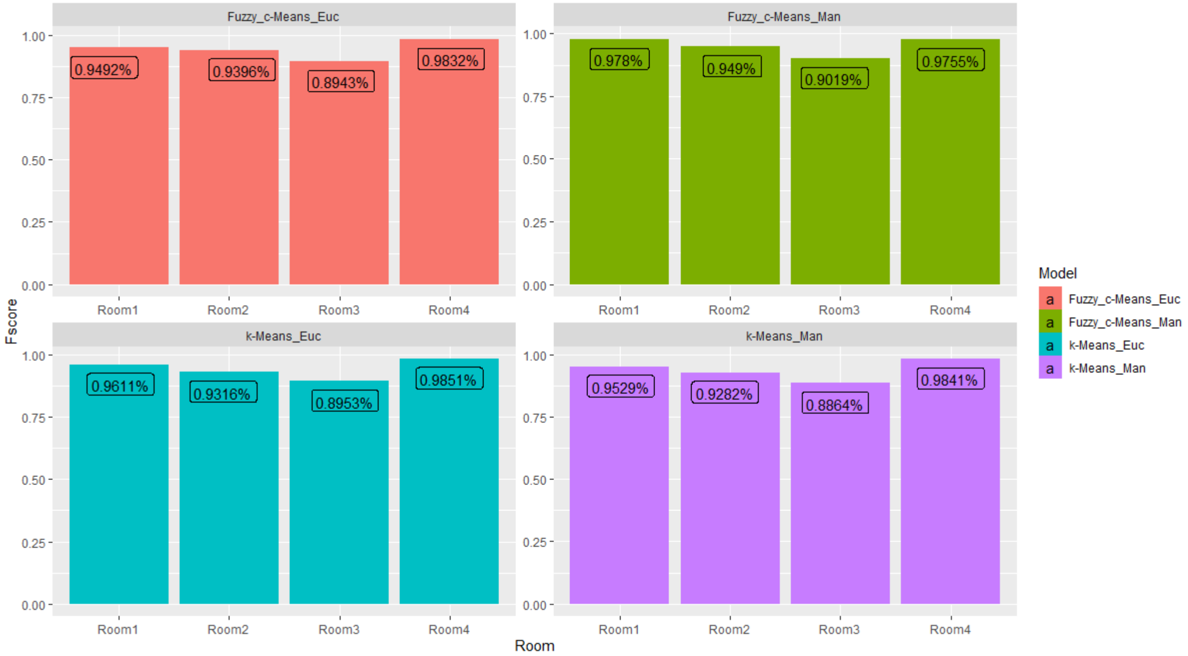


Figure 3. Results of clustering algorithms according to F score measure.

Table 2. Results of clustering algorithms according to internal indexes

	KM_ M	KM_E	FCM_M	FCM_E
DI	**0,042**	0,035	0,027	0,029
SI	0,421	**0,423**	0,378	0,420
DBI	0,891	0,868	0,927	**0,866**
CI	**0,059**	0,065	0,090	0,066
CI: C index, DBI: Davies-Bouldin Index, DI: Dunn Index, FCM_E :Fuzzy c-Means Euclidean Distance, FCM_M: Fuzzy c-Means Manhattan Distance, KM_E: k-Means Euclidean Distance, KM_M: k-Means Manhattan Distance				

The F score values of each model have been averaged to determine the success of predicting all rooms correctly. According to the Manhattan distance, the average accuracy value obtained with the k-Means algorithm is 93.8% and the average accuracy value obtained with the Fuzzy c-Means algorithm is 95.1%. According to the Euclidean distance, the average accuracy value obtained with the k-Means algorithm is 94.3% and the average accuracy value obtained with the Fuzzy c-Means algorithm is 94.2%.

As a result of successful clustering, Dunn and Silhouette should get maximum values, Davies-Bouldin and C index should get minimum values. When Table 2 regarding the internal index values is examined, the models produced with k-Means-Manhattan distance according to Dunn and C indices, k-Means-Euclidean distance according to Silhouette index, Fuzzy c-Means-Euclidean distance according to Davies-Bouldin index have been successful. In addition, it is seen that the results are close values.

5. Conclusions

As stated in the method section, clustering models have been obtained according to different distance measures by using the data set and the specified methods. As a result of running these models, all records in dataset have been assigned to four different clusters. The resulting cluster labels have been compared with the labels in the original data set expressing the location. Thus, it has been determined whether the clustering algorithms can make a correct clustering, and whether the users in the rooms and the users in the clusters determined by the algorithms match each other. The determined performance indicators have been calculated and the success of the clustering method has been evaluated according to these indicators.

When the results are evaluated according to internal and external indexes, different situations arise. While the external indices show the success of the Fuzzy c-Means algorithm, the inner indices indicate the success of the k-Means algorithm. At this point, external indexes should be taken as a basis for classification success. In the study, not only the accuracy measure was considered, but the F measure was especially preferred in order to include the results of different performance measures in the evaluation. Both performance measures produced parallel results when the separate success values obtained for the classes were averaged, and the most successful model was determined as Fuzzy c-Means using Manhattan's distance.

When the similarities of the elements within the cluster are examined, it is seen that the success of the clustering analysis has changed. According to the distances of the elements in each cluster, the results point to the k-Means algorithm, which mainly uses the Manhattan distance. It can be natural for this situation to occur. This situation can be evaluated from two perspectives. The evaluation is made according to the data points within the cluster, not a reference point. In other words, higher intra-cluster similarities have been obtained in the clusters obtained with k-Means. However, this situation can be interpreted as that although the data patterns in the main data set show similar features, this high similarity cannot produce a fully effective result in spatial clustering. On the other hand, average values for each model have been obtained for external indices. However, when the prediction accuracies on the basis of rooms are examined, it will be seen that there are variations in the prediction success.

When it is desired to perform location prediction on a data set without a class label, the correct number of clusters must be determined absolutely. Internal indexes will need to be used when determining this number. Even if this situation causes performance losses in the next location estimation stage, in general terms, clustering analysis can also achieve results that can compete with classification models.

Declaration

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

F.Ö. Koçoğlu is responsible for all sections of the study.

References

1. Hazas, M., J. Scott, and J. Krumm, *Location-aware computing comes of age*. Computer, 2004. **37**(2): p. 95–97.
2. Oguntala, G., R. Abd-Alhameed, S. Jones, J. Noras, M. Patwary, and J. Rodriguez, *Indoor location identification technologies for real-time IoT-based applications: An inclusive survey.* Computer Science Review, 2018. **30**: p. 55–79.
3. Curran, K. E. Furey, T. Lunney, J. Santos, D. Woods, and A. McCaughey, *An evaluation of indoor location determination technologies*. Journal of Location Based Services, 2011. **5**(2): p. 61–78.
4. Nath, R.K., R. Bajpai, and H. Thapliyal, *IoT based indoor location detection system for smart home environment*, in *IEEE International Conference on Consumer Electronics* (ICCE). 2018. Las Vegas, USA: p. 1-3.
5. Roy P. and C. Chowdhury, *A survey of machine learning*

techniques for indoor localization and navigation systems. J Intell Robot Syst, 2021. **101**(3): p. 63.

6. Jedari, E., Z. Wu, R. Rashidzadeh, and M. Saif, *Wi-Fi based indoor location positioning employing random forest classifier*, in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2015. Calgary, Canada: p. 1–5.
7. Tabbakha, N.E., W.-H. Tan, and C.-P. Ooi, I*ndoor location and motion tracking system for elderly assisted living home*, in *International Conference on Robotics, Automation and Sciences (ICORAS)*, 2017. Melaka, Malaysia: p. 1–4.
8. Chao C. and M. Xiaoran, *An innovative indoor location algorithm based on supervised learning and wifi fingerprint classification*, in *Signal and Information Processing, Networking and Computers*, 2018. Singapore: pp. 238–246.
9. Nuño-Maganda, M.A., H. Herrera-Rivas, C. Torres-Huitzil, H. Marisol Marín-Castro, and Y. Coronado-Pérez, *On-device learning of indoor location for wifi fingerprint approach*. Sensors, 2018. **18**(7).
10. Elbes, M., E. Almaita, T. Alrawashdeh, T. Kanan, S. AlZu'bi, and B. Hawashin, An *indoor localization approach based on deep learning for indoor location-based services*, in *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019. Amman, Jordan: p. 437–441.
11. Dai, P., Y. Yang, M. Wang, and R. *Yan, Combination of DNN and improved KNN for indoor location fingerprinting*. Wireless Communications and Mobile Computing, 2019. p. e4283857.
12. Ouameur, M.A., M. Caza-Szoka, and D. Massicotte, *Machine learning enabled tools and methods for indoor localization using low power wireless network*. Internet of Things, 2020. **12**: 100300.
13. Polak, L., S. Rozum, M. Slanina, T. Bravenec, T. Fryza, and A. Pikrakis, *Received signal strength fingerprinting-based indoor location estimation employing machine learning*. Sensors, 2021. **21**(13): 4605.
14. Rizk, H., M. Abbas, and M. Youssef, *Device-independent cellular-based indoor location tracking using deep learning*. Pervasive and Mobile Computing, 2021. **75**: 101420.
15. Ge, H., Z. Sun, Y. Chiba, and N. Koshizuka, *Accurate indoor location awareness based on machine learning of environmental sensing data*. Computers & Electrical Engineering, 2022. **98**: 107676.
16. Xie, Y., T. Wang, Z. Xing, H. Huan, Y. Zhang, and Y. Li, *An improved indoor location algorithm based on backpropagation neural network*. Arab J Sci Eng, 2022. https://doi.org/10.1007/s13369-021-06529-z.
17. Rohra, J.G., B. Perumal, S. J. Narayanan, P. Thakur, and R. B. Bhatt, User localization in an indoor environment using fuzzy hybrid of particle swarm optimization & gravitational *search algorithm with neural networks*, in *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*. 2017. Singapore: p. 286–295.
18. Rokach L. and O. Maimon, *Clustering methods*, in *Data Mining and Knowledge Discovery Handbook,* O. Maimon and L. Rokach, Eds. 2005, Boston, MA: Springer, p. 321–352.
19. Omran, M.G.H., A. P. Engelbrecht and A. Salman, *An overview of clustering methods*. Intelligent Data Analysis, 2007. **11**(6): p. 583–605.
20. Ghosh S. and S. K. Dubey, *Comparative analysis of k-means and fuzzy cmeans algorithms*. International Journal of Advanced Computer Science and Applications, 2013. **4**(4): p. 35–39.
21. Izakian H. and A. Abraham, *Fuzzy C-means and fuzzy swarm for fuzzy clustering problem*. Expert Systems with Applications,2011. **38**(3): p. 1835–1838.

Appendix

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{A.1}$$

$$Dunn = \underset{1 \le i \le n}{min} \left\{ \left(\underset{\substack{1 \le j \le n \\ i \ne j}}{min} \frac{d(c_i, c_j)}{maks_{1 \le k \le n}(d'(c_k))} \right) \right\} \tag{A.2}$$

$$Silhouette(k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{maks\ (b_i, a_i)} \tag{A.3}$$

$$Davies - Bouldin = \frac{1}{n} \sum_{i=1}^{n} \underset{i \ne j}{maks} \left\{ \frac{\alpha_i + \alpha_j}{d(c_i, c_j)} \right\} \tag{A.4}$$

$$C\_Index = \frac{S_W - S_{min}}{S_{max} - S_{min}} \tag{A.5}$$

c_i and c_j, cluster centers
$d(c_i, c_j)$, distance between c_i ve c_j
$d'(c_k)$, distance between records in set *k*
a_i, average distance of record *i* in the cluster from all other records in the same cluster
b_i, minimum value of the mean distances of record *i* to the records in other clusters
α_i, average distance of records in cluster *i* from their cluster center
S_W is the sum of the NW distances between all the pairs of points inside each cluster
S_{min} is the sum of the NW smallest distances between all the pairs of points in the entire data set
S_{max} is the sum of the NW largest distances between all the pairs of points in the entire data set.