Machine Learning Approaches for Efficient Indoor Positioning: A Comparative Analysis

1st Şakir Bingöl
Electrical Engineering
Marmara University
Istanbul, Turkey
sakirbingol@marmara.edu.tr

3rd Ozan Durgut
Electrical Engineering
Marmara University
Istanbul, Turkey
ozandurgut@marun.edu.tr

3rd İsmail Emre Alçı
Electrical Engineering
Marmara University
Istanbul, Turkey
emrealci@marun.edu.tr

4th Emre Emer
Electrical Engineering
Marmara University
Istanbul, Turkey
emreemer@marun.edu.tr

5th İbrahim Tınas
Electrical Engineering
Marmara University
Istanbul, Turkey
ibrahimtinas@marun.edu.tr

6th Oğuz Berk Akdeniz
Electrical Engineering
Marmara University
Istanbul, Turkey
oberkakdeniz@marun.edu.tr

7th Zeliha Nur Çabuk
Electrical Engineering
Marmara University
Istanbul, Turkey
zelihacabuk@marun.edu.tr

8th Abdulkerim Biz
Electrical Engineering
Marmara University
Istanbul, Turkey
abdulkerimbiz@marun.edu.tr

9th Ege Okan Kalı
Electrical Engineering
Marmara University
Istanbul, Turkey
ege.okan@marun.edu.tr

***Abstract*—This paper covers the implementation of Indoor Positioning System (IPS) technology with machine learning and deep learning models and the comparison of the accuracy scores. Accurate location information is a very important need all over the world and in many sectors. Since systems such as the Global Positioning System (GPS) are not successful in locations such as multistory buildings, airports, parking garages, and underground, IPS technology is used where necessary. IPS is a network of devices used to determine the location of people and objects in such situations. The focus is on the development of IPS with machine learning and deep learning. Artificial Neural Network (ANN), Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models were trained using the UJIIndoorLoc dataset. The accuracy score that is obtained with KNN is 98.57 percent. According to these results, previous studies were left behind.**

***Index Terms*—Indoor Positioning Systems, Machine Learning, Received Signal Strength Intensity, Deep Learning, Artificial neural network (ANN).**

I. Introduction

The world is still struggling to cope with the problems of poor signal transmission. The most prominent and longest-standing of these is the problem with the Global Positioning System (GPS). The main factors that cause GPS to be unable to provide a healthy enough signal are weather conditions, hard-to-reach points such as underground and indoor areas, and satellite location [1][2][3]. Overcoming these can only be solved by localizing the problem. In order to localize this problem, it is necessary to change the system of obtaining the signal.

The Indoor Positioning System (IPS) comes into play at this point. Thanks to this system, the signal is received from routers close to the user, and the distance and approximate location can be determined thanks to the collected Received Signal Strength Intensity (RSSI) data [10][24]. In order for this algorithm to work correctly and to be implemented in places where GPS is not available or may pose security problems, a stable method and good algorithms must be selected.

Research has shown that artificial intelligence algorithms can positively affect IPS and its operation. As a result of this conclusion, available algorithms were identified, and the most efficient one was determined. In this way, the investigations have found it logical to use machine and deep learning algorithms such as Random Forest, SVM, KNN, and ANN in the first place [10][4]. We searched for suitable data to train machine learning and deep learning models. The most suitable datasets were selected, and the selected models were trained using this data. The algorithms were compared with the metric data obtained as a result of this training, and this is how the article is planned to be concluded.

The main purpose of this paper is to accurately obtain location information for people and certain objects in locations where location information is important but difficult to detect [10][6][14]. In order to obtain this, the location of people can be determined with IPS. In this direction, it is aimed at using machine and deep learning methods with IPS.

This study delves into the implementation of IPS technology, integrating machine learning and deep learning algorithms such as ANN, Random Forest, KNN and SVM to increase accuracy. These models are trained with the UJIIndoorLoc dataset and obtain more successful results than other similar studies [8]. ANN is a fundamental deep learning algorithm that connects the input and output layers with interconnected layers. Each neuron in the layers has an associated weight and threshold value. The Random Forest algorithm is slightly more complex compared to other algorithms. This algorithm creates multiple decision trees using random subsets of data and features and then combines their results to increase predictive accuracy. Additionally, it classifies these decision trees according to a regression line to teach them to the machine [22][5]. The KNN algorithm classifies a data point by determining its proximity to neighbors in the feature space. To measure the proximity, it needs to define a distance function. The most common choice is the Euclidean distance, which is the straight-line distance between two points [7]. The SVM algorithm is a supervised learning model that classifies new data into two groups based on labeled training data. Using classification algorithms, it identifies the optimal hyperplane in an N-dimensional space by targeting the largest margin between the closest points of different groups.

II. Understanding the Data

It was decided to use The UJIIndoorLoc dataset. The dataset, as described in [8], encompasses Wi-Fi fingerprints collected from 19 different individuals and 25 Android mobile devices across various buildings and floors within Universitat Jaume I. This dataset comprises a total of 19,937 data points.

Furthermore, the dataset contains 529 descriptors, representing a wide range of features. These attributes involve Wi-Fi fingerprint data derived from Wireless Access Points (WAPs) and are expressed in numerical values denoting Received Signal Strength Intensity (RSSI). Alongside these Wi-Fi descriptors, the dataset includes additional information such as BUILDINGID, USERID, and geographical coordinates. A comprehensive listing of all 529 descriptors can be found in Table 1.

The primary objective in this problem is to make the most of Wi-Fi fingerprints for indoor localization of users, with LONGTITUDE, LATITUDE, and FLOOR being the key variables to predict. To increase prediction accuracy, BUILDINGID and RELATIVEPOSITION descriptors are also taken into account. On the other hand, descriptors like SPACEID, USERID, PHONEID, and TIMESTAMP are excluded from the prediction model as they do not provide relevant information for determining an individual's position. In essence, the analysis will focus on WAP001 to WAP520 descriptors to predict LONGTITUDE, LATITUDE, FLOOR, BUILDINGID, and RELATIVEPOSITION for each instance.

TABLE I
Dataset Descriptors & Explanations

Descriptors	Explanation
WAP001 … WAP520	The signal strengths of Wireless Access Points (WAPs) are expressed as negative integer values within the range of -104dB (indicating an extremely weak signal) to 0dB. Additionally, the value 100 is employed to indicate when a WAP is not detected. Throughout the process of constructing the database, a total of 520 distinct WAPs were identified. Consequently, the Wi-Fi fingerprint is constructed using 520 intensity values.
LONGITUDE	This descriptor associated with location of user and it encompasses values ranging from -7695.938754929929900 -7299.786516730871000
LATITUDE	This descriptor associated with location of user and it encompasses values ranging from 4864745.7450159714 to 4865017.3646842018
FLOOR	This descriptor associated with location of user and indicates the altitude inside the buildings. It can takes values from 0 to 4.
BUILDINGID	Dataset collected in three seperate buildings. This descriptor can take value from 0 to 2.
SPACEID	This identifier provides information about the specific space (such as offices or labs) where the data was collected
RELATIVEPOSITION	This descriptor indicates the relative position of the user this descriptor can take value 1 if the user inside otherwise it takes value as 2 this means that user is in front of the door
USERID	This identifer gives information about whose user data is used currently in use and it can takes value 0 to 18.
PHONEID	This descriptor specifies the type of Android device used by the user it can take values from 0 to 24
TIMESTAMP	This descriptor indicates the stime stamp when the data was collected in UNIX Time format.

III. Method

After analyzing the data, a literature analysis was started to determine the way forward. After the analysis, sample articles containing methods that overlapped with our ideas were examined. As a result of the reviews, the algorithms to be used, how to process the data, how to use the models, and how to analyze the results were decided. Artificial Neural Network (ANN), Random Forest, K Nearest Neighbor (KNN) and Support Vector Machine (SVM) were chosen as the algorithms to be used in this analysis since the results we will get overlap with the studies on IPS and analyzing their performance will add value to the studies on IPS[4][6][11][12][13][16][17].

The ANN algorithm is a common algorithm used in deep learning. It consists of an input layer, one or more hidden layers, and an output layer. Each node or artificial neuron is connected to another, and these connections are linked by an associated weight value and a threshold value. Any node becomes active and transmits data to the next layer of the network as soon as its output gets past the specified threshold value [15][19]. No data is sent to the network's next layer if the output value is below the threshold value. A linear regression model with input data, weights, bias (or equal), and output can be considered for each node separately [19]. If we want to represent each node in a formalized form, also see a neural network node in Fig. 1, we can represent it as follows:[15][20]

$$(\sum_{i=1}^{N} w_i x_i) + bias \tag{1}$$

The Random Forest algorithm is slightly more complex compared to other algorithms. In the random forest algorithm,

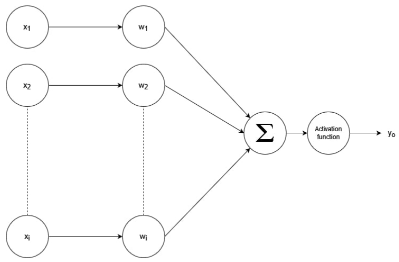

Fig. 1. Shows the neural network node structure.

multiple decision trees are created on the given data set. Decision trees created in large numbers take part in the decision mechanism of the algorithm [14][21]. The fact that each decision tree follows a different path allows all kinds of possibilities to be taken into account. Not all trees created in a classification or prediction process are used. A certain number of trees are selected in a random manner.[14][23][13] At the end of this selection, the average of the trees or the occurrence of a selected feature in greater numbers than others is the logic behind the random forest's decision. In Fig. 2, we observe the basic structure of the random forest algorithm.

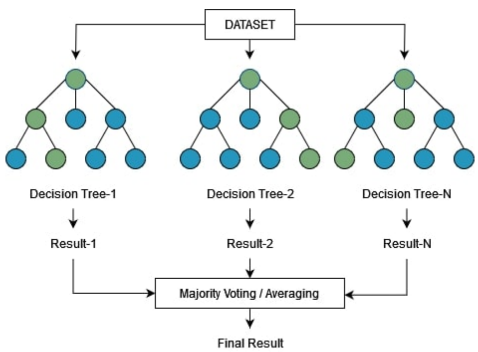

Fig. 2. Shows the random forest algorithm's schematic.

The KNN algorithm is a classifier that uses the concept of proximity to group a data point. It is a non-parametric and supervised learning method to find the k nearest neighbors of the data point in the feature space and look at their labels or values (Fig. 3).[26] For classification problems, where the labels are discrete, it assigns the most frequent label among the neighbors to the data point. For regression problems, where the values are continuous, it takes the average of the neighbors to predict the value of the data point. To measure the proximity, it needs to define a distance function. The most common choice is the Euclidean distance, which is the straight-line distance between two points.[16] Another way KNN works is weighted; KNN evaluates the weights. The weights are inversely proportional to the distances of neighbors. An estimate is made by comparing the weights. The estimate with the higher weight is selected.[17][18]

The SVM algorithm (Support Vector Machine) is a supervised machine learning model that can classify new data into two groups. It learns from sets of labeled training data for

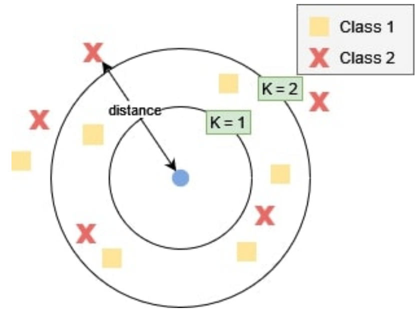

Fig. 3. Illustrates the KNN algorithm's decision stage.

each group, and then it can classify new data into one of the groups. It uses classification algorithms to find the optimal hyperplane in a space with N dimensions, where N is the number of features in the data (Fig. 4).[27][13][17][28] The optimal hyperplane has the largest possible margin between the closest points of different groups in the feature space.[29] The data points that are on the margin are called support vectors. If the optimal hyperplane exists, it is called the hard margin. Otherwise, some errors are allowed, and it is called the soft margin.[18][30]

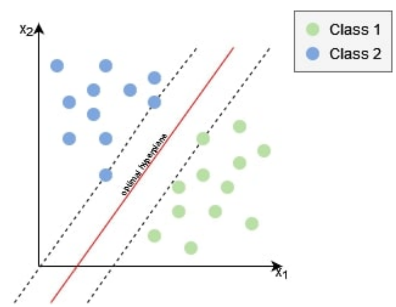

Fig. 4. Shows the SVM schematic for 2 dimensional space.

According to the dataset that is mentioned in Section II and the algorithms mentioned in this section, test code was written in Python using code libraries such as Pandas and Scikit-learn [32]. The dataset was analyzed using principal component analysis (PCA) to obtain a lower-dimensional and more compact dataset [9][19][31]. The data set was divided into an 80 percent training set and a 20 percent test set. Trained models were tested, and an accuracy score was obtained. This accuracy score is the ratio of predicted labels to true labels [20]. Predicted labels are the predictions obtained as a result of the test, and true labels are the correct ones of the predicted labels. The accuracy scores obtained are given in Table 2.

TABLE II
SHOWS THE ACCURACY RESULTS FOR EACH MODEL.

	Build ID	Floor	Rel Position
ANN	99.7492%	93.2296%	87.4122%
Random Forest	99.7993%	99.5737%	95.7873%
KNN	99.7241%	97.0411%	93.1795%
SVM	99.7993%	97.3671%	91.0481%

TABLE III
PERFORMANCE COMPARISON OF KNN AND RANDOM FOREST MODELS

	ML			
	KNN		**Random Forest**	
	PCA	*No-PCA*	*PCA*	*No-PCA*
FLOOR	97.1523%	97.0411%	97.0160%	99.5737%
BUILD ID	99.8495%	99.7241%	99.7743%	99.7993%
Mean	98.5687%	98.3826%	98.3951%	99.6865%

TABLE IV
PERFORMANCE COMPARISON OF SUPPORT VECTOR MACHINE AND ARTIFICIAL NEURAL NETWORK MODELS

	ML		**DL**	
	SVM		**ANN**	
	PCA	*No-PCA*	*PCA*	*No-PCA*
FLOOR	94.4583%	97.3671%	96.9157%	93.2296%
BUILD ID	99.6740%	99.7993%	99.87462%	99.7492%
Mean	97.0661%	98.5796%	98.3951%	96.4894%

Table 3 and 4 shows the accuracy values of 4 algorithms (KNN, Random Forest, SVM, ANN) with and without PCA transformation. Accuracy decreased slightly in methods using PCA transformation. Additionally, CPU usage has decreased. In this way, the system was used more efficiently.

IV. CONCLUSION

Indoor positioning continues to be a very successful technology where systems such as GPS are not available. Our research has shown that the use of artificial intelligence in conjunction with this technology will bring great advantages and increase productivity.

TABLE V
SHOWS THE RESULTS OF THE KNN ALGORITHM FROM OTHER RESEARCH (J. YANG ET AL., 2022).

Model	Accuracy Score
K-Nearest Neighbours	89.92%
Random Forest	89.92%
Support Vector Machine	92.44%
Decision Trees	85.60%
Yang et al.'s Solution	96.22%

A remarkable result has been obtained on indoor positioning related to the performance of these algorithms by using various machine learning and deep learning algorithms[25]. When the results of experiments using different algorithms such as KNN, Random Forest, SVM and ANN are examined, it is observed that the algorithm with the highest accuracy is Random Forest. Although the results of other algorithms have sufficient rates, the algorithm with the lowest success rate is SVM (Table 2).

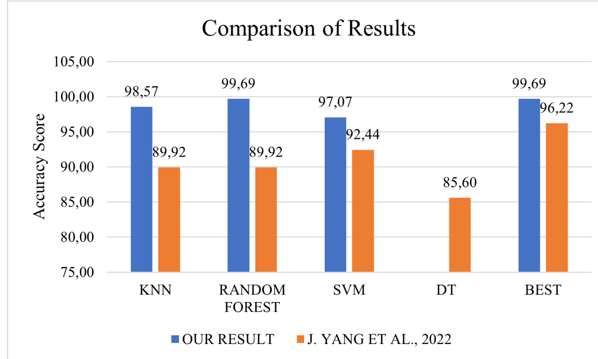

Fig. 5. Shows the comparison of the results between correlated study.

Since the density of the processed data set used the CPU at a high rate, the size of the data set was reduced by neglecting the least important feature variables with the PCA method [33] [34]. The most effective data, as the number of dimensions determined by the PCA method, was included in the progress. With the more efficient use of the data set, the number of WAPs from which data was obtained was reduced from 520 to 100. In return, CPU usage was reduced, and rearrangement was made for a more optimal system. In the system where the PCA method was used, there has been a negligible decline in the accuracy rate[32][35]. As a trade-off, CPU usage was significantly reduced, and a more optimal system was transitioned.After these procedures, there was no change in the success ranking between the algorithms. Random Forest again gave the most successful result.

TABLE VI
SHOWS THE RESULTS TAKEN FROM THE "CALIBRATION-FREE 3D INDOOR POSITIONING ALGORITHMS BASED ON DNN AND DIFF" RESEARCH[36].

	Target 3-FLOOR	Target 4-BuildID	Mean
KNN	96.30%	100%	98.15%

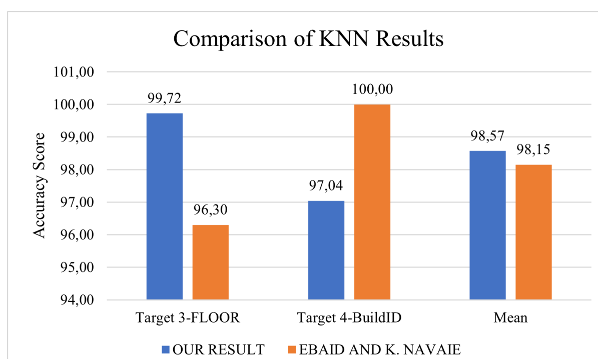

Fig. 6. Shows the comparison of the results between correlated study specificied to KNN.

It has been observed that we have achieved higher accuracy rates compared to the algorithms used in other articles.(Figure 5 and 6) For example, the highest accuracy score among the algorithms that are used is 99.6865%, while the higher value in another article is 96.22% (Table 6). Similarly, the maximum accuracy score received with the KNN algorithm is 98.5687%, compared to 98.15% in a different project (Figure 6).

Taking everything into account, when we look at our values, we have more successful results than other studies. From this point of view, the models that we have used and the machine learning methods that we have used are efficient.

REFERENCES

[1] N. Pritt, " Indoor positioning with maximum likelihood classification of wi-fi signals. In SENSORS, 2013 ", IEEE, pages 1–4, Nov 2013.

[2] Zhang, Jianjun, Lixiang Sun, and Hong Yuan. "Assessment and research on the Self-Interference of GPS weak signal acquisition in indoor location environment." 2009 IEEE Youth Conference on Information, Computing and Telecommunication. IEEE, 2009.

[3] Bachu, P., Dudala, T., Kothuri, S. M. (2001). Prompted recall in global positioning system survey:Proof-of-concept study.Transportation Research Record, (1768), 106 – 113.

[4] Yinhuan Dong, Francisco Zampella, Firas Alsehly, "Beyond KNN: Deep Neighborhood Learning for WiFi-based Indoor Positioning Systems", Edinburgh Research Centre Huawei Technologies Research and Development (UK) Ltd, Edinburgh, UK.

[5] Chia-Hsin Cheng & Tao-Ping Wang & Yung-Fa Huang, Indoor Positioning System Using Artificial Neural Network With Swarm Intelligence Published in: IEEE Access, PageStart: 84248, PageEnd: 84257, DOI:10.1109/ACCESS.2020.2990450, Digital Object Identifier: 10.1109/ACCESS.2020.2990450, Publised Date: April 27, 2020.

[6] Tao Gu, Hongyu Wu, Shu Wu, and Jun Guo, "Evaluation of Machine Learning Prediction Models for Wifi based Indoor Positioning System" 2023 International Conference on Computer, Information and Telecommunication Systems (CITS), Computer, Information and Telecommunication Systems (CITS), 2023 International Conference.

[7] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[8] UCI Machine Learning Repository: UJIIndoorLoc Data Ser. [Online]. Available: https://archive.ics.uci.edu/dataset/310/ujiindoorloc [Accessed: 16-Oct-2023].

[9] I.T. Jolliffe, " Principal Component Analysis" Springer Verlag, 1986.

[10] Jekaterina Novikova, Pavel Masek, Vladimir Marik, and Jiri Hosek, "Indoor Positioning Technologies, Solutions, and Challenges" International Conference on Indoor Positioning and Indoor Navigation (IPIN), 2017.

[11] J. Wang, F. Shi, P. Wan, M. Chen, and F. Jiang, "An Improved Particle Swarm Optimization Indoor Positioning Method Based on the Weighted Adaptive KNN Algorithm," in Indoor Positioning and Indoor Navigation (IPIN).Xi'an University of Posts and Telecommunications, Xi'an 710121, International Conference on, pp. 261–270, Oct 2014.

[12] S. Bozkurt, G. Elibol, S. Gunal and U. Yayan, "A comparative study on machine learning algorithms for indoor positioning," 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), Madrid, Spain, 2015, pp. 1-8, doi: 10.1109/INISTA.2015.7276725.

[13] X. Cai, L. Zhu, X. Zheng and H. Zhang, "Research and Analysis of Indoor Positioning Accuracy Based on Machine Learning and Particle Filtering Algorithm," 2022 International Seminar on Computer Science and Engineering Technology (SCSET), Indianapolis, IN, USA, 2022, pp. 32-35, doi: 10.1109/SCSET55041.2022.00017.

[14] A. H. Salamah, M. Tamazin, M. A. Sharkas and M. Khedr, "An enhanced WiFi indoor localization system based on machine learning," 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Alcala de Henares, Spain, 2016, pp. 1-8, doi: 10.1109/IPIN.2016.7743586.

[15] Meysam Sadeghi, Ali Khaleghi, and Ebrahim Saberinia, "An Indoor Pseudolite Positioning Method Based on Measured and Simulated Fingerprints" ,IEEE Antennas and Wireless Propagation Letters, Antennas and Wireless Propagation Letters, IEEE, Antennas Wirel. Propag. Lett.,2023,IEEE Xplore Digital Library.

[16] Indoor Positioning Using Wi-Fi and Machine Learning for Industry 5.0 Published in: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 2023 IEEE International Conference on,20230313,IEEE Xplore Digital Library.

[17] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967, doi: 10.1109/TIT.1967.1053964.

[18] H. Yigit, and A. Kavak, "A Learning Approach in Link Adaptation for MIMO-OFDM Systems," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 21, no. 5, pp. 1465-1478, 2013.

[19] Dastres, Roza & Soori, Mohsen."Artificial Neural Network Systems. International Journal of Imaging and Robotics". 2021, 21. 13-25.

[20] Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. "Understanding neural networks through deep visualization." arXiv preprint arXiv:1506.06579 (2015).

[21] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[22] Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996).

[23] Jedari, Esrafil, et al. "Wi-Fi based indoor location positioning employing random forest classifier." 2015 international conference on indoor positioning and indoor navigation (IPIN). IEEE, 2015.

[24] Martin Klepal, Stéphane Beauregard, et al., A novel backtracking particle filter for pattern matching indoor localization, in: Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPSless Environments, ACM, San Francisco, USA, 2008, pp. 79–84.

[25] D. V. Le, N. Meratnia and P. J. M. Havinga, "Unsupervised Deep Feature Learning to Reduce the Collection of Fingerprints for Indoor Localization Using Deep Belief Networks," 2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Nantes, France, 2018, pp. 1-7, doi: 10.1109/IPIN.2018.8533790.

[26] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3-62.

[27] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[28] L. Wang, Support Vector Machines: Theory and applications. Berlin, Heidelberg: Springer, 2005.

[29] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[30] Burges, C.J. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998).

[31] Holland, Steven M. "Principal components analysis (PCA)." Department of Geology, University of Georgia, Athens, GA 30602 (2008): 2501.

[32] Kramer, Oliver, and Oliver Kramer. "Scikit-learn." Machine learning for evolution strategies (2016): 45-53.

[33] Fengxi Song, Zhongwei Guo, Dayong Mei, "Feature selection using principal component analysis".

[34] Ranak Roy Chowdhury, Muhammad Abdullah Adnan, and Rajesh K. Gupta, "Real Time Principal Component Analysis".

[35] Ebaid, Emad, and Keivan Navaie. "Optimum NN Algorithms Parameters on the UJIIndoorLoc for Wi-Fi Fingerprinting Indoor Positioning Systems." 2022 32nd International Telecommunication Networks and Applications Conference (ITNAC). IEEE, 2022.

[36] Yang, Jingmin, et al. "Calibration-free 3D indoor positioning algorithms based on DNN and DIFF." Sensors 22.15 (2022): 5891.