International Journal of Innovative
Computing, Information and Control
Volume **20**, Number **6**, December **2024**

ICIC International ©2024 ISSN 1349-4198
pp. 1851–1866

ALGORITHM OF VARIABLE WEIGHT NEAREST NEIGHBOR POSITIONING UNDER MULTI-SCALE OPTIMIZATION

Shunyuan Sun[1,2], Jingyuan Yu[1,2,*] and Ningning Qin[1,2]

[1]School of Internet of Things Engineering
Jiangnan University
[2]Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education
No. 1800, Lihu Avenue, Wuxi 214122, P. R. China
{ hzrobin; qinnn }@jiangnan.edu.cn
*Corresponding author: 6211924127@stu.jiangnan.edu.cn

Received December 2023; revised May 2024

Abstract. *The conventional clustering method is influenced by the distribution shape, scale, dimension, and model parameters of the dataset, which may hinder accurate and effective positioning. To address the aforementioned issues, a novel algorithm for indoor positioning, which incorporates variable weight approximate nearest neighbor and multi-scale optimization, is proposed. During the offline stage, a one-to-many support vector machine partition model is established based on feature extraction, taking account of the distribution shape of the data set and sample data. This model aims to reduce computational complexity and enhance search efficiency in the online phase. In the online phase, the integration of historical data compensation and the similarity measure of variable weight between samples is proposed, along with a specialized Euclidean metric calculation method for the approximate nearest neighbor search of input samples. It can effectively address the issue of high search efficiency but poor accuracy in the traditional approximate nearest neighbor method. Experimental results indicate that the proposed method exhibits a 16.8% higher probability of positioning error within 1.5 m compared to the traditional improved clustering method. Additionally, the average positioning accuracy can achieve 0.802 m.*

Keywords: Indoor location, Multi-scale, Feature extraction, Variable weight similarity measure, Special-shaped Euclid

1. **Introduction.** In the context of indoor fingerprint positioning [1], partition clustering represents a widely used approach that aims to enhance the accuracy of positioning estimation by segmenting the fingerprint data into distinct regions. Partition clustering is based on a clustering algorithm that groups fingerprint data with similar features into the same region, thereby decreasing the complexity and accuracy demands of indoor positioning.

Currently, partition clustering is extensively utilized in indoor fingerprint positioning. The methods encompass K-means [3], DBSCAN density clustering [4], hierarchical clustering [5] and so on. Nevertheless, the aforementioned partition clustering methods are susceptible to the shape, size, dimensions, and model parameters of the dataset, and are unable to attain precise and efficient localization. Extensive research has been conducted by scholars both domestically and internationally on the decline of positioning accuracy resulting from the instability of clustering results.

During the offline stage, [6] employed heuristic techniques to set the initial values of the algorithm or iteratively restarts the K-means algorithm to address the persistent

DOI: 10.24507/ijicic.20.06.1851

degradation of clustering results. However, this approach consumes substantial computing resources during the clustering process. [7] improved the positioning algorithm through combining K-means and Wasserstein to obtain generative adversarial network (WGAN), resulting in an extended fingerprint database obtained through dual clustering in the offline phase. While the algorithm does to some extent balance the utilization of computing resources, it also adds complexity to the algorithm and overlooks the processing of outliers. [8] proposed a one-to-one multi-classification SVM and integrated transfer learning algorithm for locating mobile tags in indoor scenes. While the accuracy of positioning has been enhanced through data standardization, the one-to-one multi-classification method will have "grey areas", resulting in misjudgment of classification.

For the online phase, conventional location algorithms such as nearest neighbor search (NN), K nearest neighbor (KNN) [9], and weighted K nearest neighbor (WKNN) [10] require traversing the entire fingerprint database to identify targets similar to the data to be located. This process is time-consuming and does not ensure real-time performance in target prediction. The support vector machine (SVM) [12], Naïve Bayes (NB) [13], and random forest (RF) [14] are commonly used classification algorithms in machine learning. While these methods do not require traversal of the entire fingerprint database, their heightened sensitivity to outliers may result in significant location fluctuations. [15] introduced an approximate nearest neighbor search approach to expedite target search at the expense of some accuracy. However, in scenarios where high accuracy is essential, the compromised accuracy may not adequately demonstrate the benefits of expedited search.

In this paper, we propose a variable weight nearest neighbor positioning algorithm under multi-scale optimization (MSO-VWNN) to address the challenges of low clustering efficiency, time-consuming target prediction, and sensitivity to outliers.

The discrete nature of each sample data feature in relation to the central feature and the distribution of the largest eigenvalues of the sample are combined, and the one-to-many multi-classification support vector machine (MCSVM) is employed to construct the location scene partition model, compared with the traditional partition model, the proposed method maximizes the effective information from the sample data, reduces data complexity, facilitates the classification algorithm in identifying sample features, and diminishes the impact of irrelevant information on positioning accuracy.

The traditional on-line search method has been enhanced based on the calculation results of the heteromorphic approximate Euclidean metric (HAEM) and historical data compensation. This improvement involves the utilization of the approximate nearest neighbor search (ANN) algorithm to achieve on-line position estimation, compared with the traditional nearest neighbor search method, the proposed algorithm aims to penalize features with larger scales or differences, and also reduces the impact of outliers on the prediction results.

The rest of the paper is organized as follows. In the second section, the location scene model and algorithm of the system are introduced in detail. In the third part, the concept of feature extraction and the specific algorithm flow are introduced in detail. In the fourth part, a detailed explanation of the enhancement of the traditional approximate nearest neighbor algorithm is provided. The fifth part involves the experimental verification and analysis of the methods mentioned above. The sixth part is a summary of the thesis work.

2. **Scene Model and Algorithm.**

2.1. **Scene model.** In the localization domain, n wireless access points (AP) and m received signal strength (RSS) sampling reference points (RP) are organized, and the initial fingerprint database is established based on the RSS and physical coordinates of

each reference point. The necessary parameters of the algorithm are described as follows unless otherwise specified:

1) $\mathbf{H}' = \left\{\left(RSSI_i^1, RSSI_i^2, \ldots, RSSI_i^n\right) | 1 \leq i \leq m\right\}$ represents the RSS from n APs at m reference points within the location area, where $RSSI_i^j$ denotes the signal strength of the jth AP at the ith reference point. $j \in \{1, 2, \ldots, n\}$;

2) $\mathbf{X} = \{(x_i, y_i) | 1 \leq i \leq m\}$ represents the set of physical coordinates for two reference points within the positioning area. Here, the physical coordinates of the ith reference point are denoted as (x_i, y_i);

3) $\boldsymbol{q} = \left(RSSI_q^1, RSSI_q^2, \ldots, RSSI_q^n\right)$ represents the RSS vector of the target to be identified in the online phase.

2.2. **Algorithm idea.** In this paper, the wireless RSS is utilized as the signal fingerprint for indoor localization. The system is comprised of two stages: the construction of a fingerprint database and fingerprint recognition.

During the construction phase of the fingerprint database, intelligent mobile terminal equipment is utilized to gather the RSS of each beacon at various locations within the positioning area. The corresponding physical coordinates are then incorporated as the initial fingerprint library of the positioning area. The initial fingerprint database undergoes feature extraction to be converted into an eigenvalue matrix, which is subsequently input into the MCSVM algorithm for model training.

During the fingerprint recognition stage, the user gathers data from any location within the positioning area and subsequently cleans the data. During feature extraction, the sample data undergoes transformation into a feature vector, and the feature weight of the sample is subsequently calculated. The MCSVM model is employed for forecasting the partition of the input data. It determines whether to augment historical data compensation based on the disparity between the historical and current prediction results. Upon identifying the partition to which the sample data belongs, the weight ANN complete search method is employed to ascertain the location information of the target. The positioning procedure of the MSO-VWNN algorithm is depicted in Figure 1.

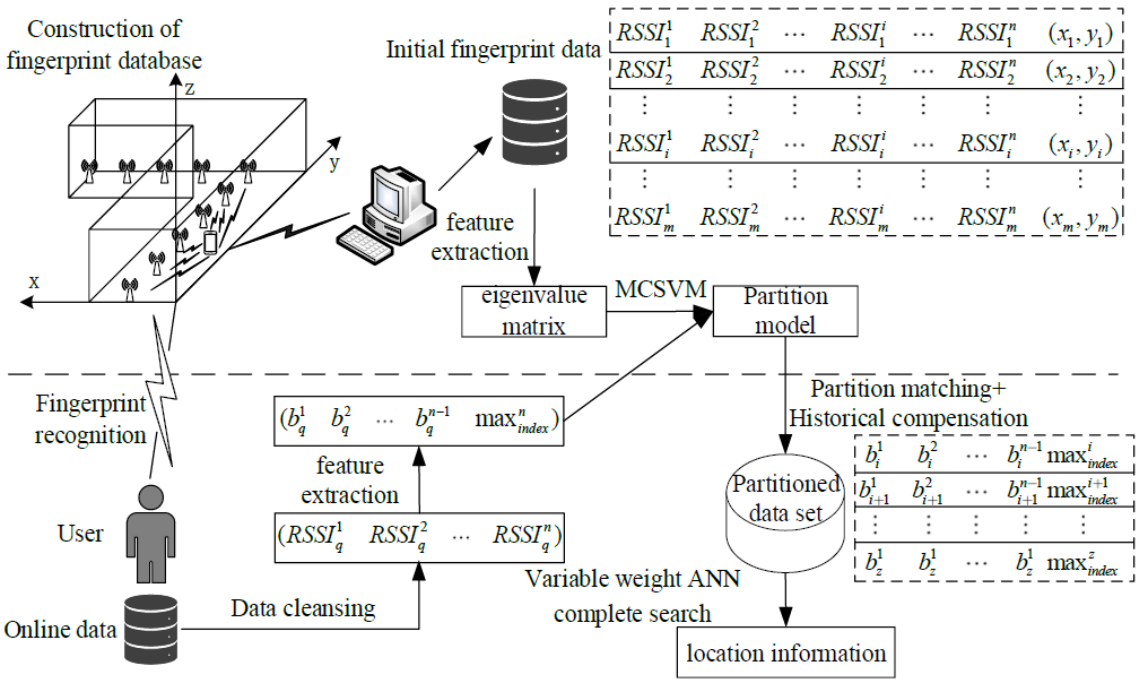

FIGURE 1. MSO-VWNN algorithm positioning flow

3. **Partition Clustering Method Based on MCSVM.** During the offline phase, it is essential to gather and determine the RSS associated with each reference point in the scene, and to classify the collected RSS data effectively. Traditional clustering methods are primarily categorized based on sample similarity. However, in complex indoor environments, various interference components, including obstacle attenuation, multipath effects, personnel movement, and electromagnetic interference from other wireless devices, contribute to the instability of RSS data and the presence of numerous outliers. Even in remote areas, similarity may emerge, leading to a significant decrease in the precision of the positioning algorithm.

Given the aforementioned issues, the paper focuses on the evolving trend of the sample data in the fingerprint database. It extracts the characteristics of the original fingerprint database based on the distinctions among the sample features and converts it into an eigenvalue matrix. Automatically add tags to eigenvalue matrix based on the optimal partition interval, employ the MCSVM algorithm to train the model, and subsequently save the model.

3.1. **MCSVM algorithm.** Given the efficient nonlinear classification, strong generalization, and excellent high-dimensional data processing abilities of support vector machines (SVM), they are capable of addressing issues related to uneven signal distribution and high noise. The RSS of each reference point within the location area exhibits similarity to a non-linearly separable multi-class signal with a small number of outliers. This characteristic can be leveraged to address the optimal partition problem of the location region.

In this study, the one-to-many (One-vs-Rest) two-classification approach in MCSVM is employed to create several two-classification subproblems. The eigenvalue matrix is clustered, and the optimal set of hyperplanes capable of multi-classification is determined. The appropriate kernel function and regularization parameters for each classifier should be selected, and SVM should be used to determine the optimal decision boundary. This will enable the division of the location region into different sub-regions and reduce the amount of data computation in the online phase.

3.2. **Feature extraction of sample distribution.** The signal fingerprint classification algorithm requires the initial establishment of the location area's original fingerprint database, followed by the optimization of its classification. The initial fingerprint database is presented below.

$$\mathbf{H} = [\mathbf{H}', \mathbf{X}] \tag{1}$$

In an effort to decrease the complexity of the initial fingerprint database and enhance the precision of the algorithm, the sample distribution feature of the original fingerprint database is extracted and utilized to simplify the intricate representation of the original data. First, the average median value of each feature in sample data set $\mathbf{H}'$ is computed, followed by the calculation of the difference between each feature in the sample data and the $\bar{M}$ coefficient. The weight, denoted as w_i^j, is calculated as the difference between the jth feature and the $\bar{M}$ value of the ith sample, divided by the total difference of the ith sample. The calculation formula is as follows.

$$w_i^j = \frac{RSSI_i^j - \bar{M}}{\sum_{j=1}^{n} \left(RSSI_i^j - \bar{M}\right)} \tag{2}$$

where w_i^j represents the weight of the jth feature of the ith sample data.

All values of w_i^j are computed, and the feature weight matrix $\boldsymbol{W}_{m\times n}$ is constructed as depicted in Equation (3). Given that assigning a minimum value without a signal would

render the weight invalid, the feature weight linked to the beacon without a signal is reset to 0.

$$\boldsymbol{W}_{m\times n}=\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\\ \vdots\\ \boldsymbol{w}_m\end{bmatrix}=\begin{bmatrix}w_1^1 & w_1^2 & \cdots & w_1^n\\ w_2^1 & w_2^2 & \cdots & w_2^n\\ \vdots & \vdots & \ddots & \vdots\\ w_m^1 & w_m^2 & \cdots & w_m^n\end{bmatrix} \tag{3}$$

where $\boldsymbol{w}_i$ represents the vector of distance weights for the ith reference point in relation to the median of the average feature.

The feature weight matrix, as depicted in Formula (3), represents the distribution of various features in relation to the feature center within each sample data. Subsequently, the original data is projected onto the feature weight space through the dispersion of sample features. Complete the scene mapping and simultaneously simplify its complex representation.

3.3. **Feature matrix of fusion quantization coding.** To simplify the calculation of the original feature weight matrix, it is quantized, encoded, and transformed into a binary matrix, thereby reducing computational complexity.

The weight matrix in Formula (3) is transformed into a binary matrix based on the size relationship of the adjacent elements in each row vector. This transformation involves comparing the sizes of the two adjacent elements in the matrix row vector sequentially. If the value of the former exceeds that of the latter, it will be designated as 1; otherwise, it will be designated as 0. The calculation formula is presented below.

$$b_i^t=\begin{cases}0, & w_i^j\le w_i^{j+1}\\ 1, & w_i^j>w_i^{j+1}\end{cases} \tag{4}$$

where b_i^t represents the size relation of two adjacent elements in each row of the weight matrix, denoted as $b_i^t\in\{0,1\}$, and t is the binary matrix column index, $t\in\{1,2,\ldots,n-1\}$.

In order to enhance the predictive accuracy of the algorithm, the peak feature of the sample is incorporated in addition to the binary matrix. The analysis of the relationship between signal strength and distance reveals a notable spike in the features present in each sample within the data collected from various location areas. The feature peak index vector, as defined in Formula (5), is created by extracting the index of the feature peak from each sample data. This vector is then incorporated into the binary matrix as a new feature. Subsequently, the quantized eigenvalue matrix, represented by Formula (6), is produced, with $\boldsymbol{B}_i$ denoting the eigenvalue vector of the ith RSS sample. The process of generating the eigenvalue matrix is given in Algorithm 1.

$$\boldsymbol{h}^{index}=\begin{bmatrix}\max_1^{index} & \max_2^{index} & \cdots & \max_m^{index}\end{bmatrix}^T \tag{5}$$

where $\boldsymbol{h}^{index}$ represents the characteristic peak index vector of the original fingerprint database, and $\max_i^{index}$ denotes the index of the largest eigenvalue of the ith sample.

$$\boldsymbol{B}=\begin{bmatrix}\boldsymbol{B}_1\\ \boldsymbol{B}_2\\ \vdots\\ \boldsymbol{B}_m\end{bmatrix}=\begin{bmatrix}b_1^1 & b_1^2 & \cdots & b_1^{n-1} & \max_1^{index}\\ b_2^1 & b_2^2 & \cdots & b_2^{n-1} & \max_2^{index}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ b_m^1 & b_m^2 & \cdots & b_m^{n-1} & \max_m^{index}\end{bmatrix} \tag{6}$$

After acquiring the eigenmatrix, it is essential to assign tags to it before employing MCSVM for model training. The paper aims to classify the eigenvalue matrix according

to different partition intervals. The matrix of eigenvalues following the process of tagging is presented below

$$\boldsymbol{S} = \{(\boldsymbol{B}_1, l_1), (\boldsymbol{B}_2, l_2), \ldots, (\boldsymbol{B}_m, l_m)\} \tag{7}$$

where, the label $l_i \in \{1, 2, \ldots, K\}$ denotes the ith eigenvector, which partitions the spatial region into K sub-regions $\{\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_K\}$.

The input for the model training in the MCSVM algorithm is denoted by $\boldsymbol{S}$, and the MCSVM model with the highest prediction accuracy is retained for partition prediction during the online stage.

Algorithm 1: The generating process of eigenvalue matrix

```
Input initial RSS matrix H'
For RSSI_i^j in H' do:
    Compute w_i^j and max_i^index        // Based on Formulae (2) and (5)
End for
Generate W_{m×n}                         // Based on Formula (3)
For w_i in W_{m×n} do:
    For w_i^j in w_i do:
        If w_i^j ≤ w_i^{j+1} :           // Based on Formula (4)
            b_i^j = 0
        Else:
            b_i^j = 1
    End for
    B_i[n] = max_i^index
End for
```

4. Variable Weight Approximate Nearest Neighbor Complete Search Method.

4.1. **Traditional approximate nearest neighbor search.** Currently, the common approach for indoor search of input data in the online phase involves identifying the nearest neighbor target value of the input data based on the index of the sample data or fixed coarse clustering. This implies that numerous superfluous data points will be involved in the nearest neighbor target search during the online phase, significantly impacting the real-time and accuracy of the online phase location.

To address the aforementioned issues, the method of approximate nearest neighbor search [16] is employed to conduct the search for the nearest neighbor target value of the input data during the online phase.

Nevertheless, the conventional nearest neighbor search approach still requires exploration across multiple subclasses and employs an equal weight similarity measure in the traditional matching algorithm. As a result, achieving optimal target matching becomes more complex and does not lead to improved accuracy. Consequently, the conventional approximate nearest neighbor search method will no longer be suitable in instances of high accuracy.

4.2. **Variable weight approximate nearest neighbor complete search.** In the context of indoor positioning, it is required that the prediction accuracy can be guaranteed while optimizing the prediction efficiency. Consequently, the conventional approximate nearest neighbor approach in online search is enhanced through the integration of variable weight similarity measurement and historical data compensation, leading to the proposal of a variable weight approximate nearest neighbor complete search method. The

algorithm's prediction accuracy is enhanced by incorporating a weighting factor based on its original search efficiency.

4.2.1. *Variable weight similarity measure.* In conventional fingerprint location algorithms, the typical approach for assessing sample similarity involves computing the Euclidean distance between two samples and evaluating the degree of similarity based on the magnitude of the Euclidean distance.

As the number of dimensions increases, the Euclidean distance, which intuitively measures the straight line distance between samples, tends to become almost equal or nearly equal, rendering the results unreliable. The measurement of the Euclidean distance presents challenges in assessing the significance of various features within a single sample. This uniform treatment may disproportionately amplify features with substantial scales or differences, thereby potentially affecting the accuracy of indoor positioning.

In this paper, a method for calculating the special-shaped Euclidean distance is proposed to address the aforementioned issues. By considering the discreteness of various sample features in relation to the sample feature center as the weight, the equivalence of different sample features is eliminated, and the deliberate punishment exhibits characteristics of significant scale or difference, thereby enhancing the accuracy of approximate nearest neighbor search.

Upon obtaining the RSS sample $\boldsymbol{q}$ from a specific location area, the process begins with feature extraction, and quantization coding is used to reconstruct $\boldsymbol{q}$. This is followed by transforming into the eigenvalue vector $\boldsymbol{Q}$, and saving the weight vector $\boldsymbol{w_q}$, which represents the relative feature of each sample feature relative to its median.

Suppose that in the offline stage, through the optimal MCSVM model predicts that $\boldsymbol{Q}$ belongs to the sub-region $\boldsymbol{\Omega}_d = \begin{bmatrix}\boldsymbol{B}_i & \boldsymbol{B}_{i+1} & \cdots & \boldsymbol{B}_z\end{bmatrix}^T$, where $d \in \{1, 2, \ldots, K\}$, $(z-i) < m$, by utilizing the equation

$$D_c\left(\boldsymbol{q}, \boldsymbol{B}_c\right)^2 \approx D_c\left(\boldsymbol{Q}, \boldsymbol{B}_c\right)^2 = \left(\boldsymbol{w}_{\boldsymbol{q}}^T \cdot \left\|\boldsymbol{Q} - \boldsymbol{B}_c\right\|^2\right)^{\frac{1}{2}} \tag{8}$$

The Euclidean metric between $\boldsymbol{Q}$ and $\boldsymbol{B}_c$ can be computed, providing an approximate representation of the Euclidean metric between $\boldsymbol{Q}$ and $\boldsymbol{B}_c$. Here, $D_c(\boldsymbol{q}, \boldsymbol{B}_c)^2$ is the Euclidean metric of $\boldsymbol{q}$ and the cth sample $\boldsymbol{B}_c$ in the sub-region space $\boldsymbol{\Omega}_d$, $c \in \{i, i+1, \ldots, z\}$.

The actual physical coordinates of the corresponding samples are selected as the actual location of the RSS sample based on the minimum result of the special-shaped Euclidean measure calculated by Formula (8), and the algorithm flow is presented in Algorithm 2.

Algorithm 2: The process of nearest neighbor target search

Input the RSS sample somewhere in the area $\boldsymbol{q}$
Compute the eigenvector $\boldsymbol{Q}$ of $\boldsymbol{q}$ // Based on Formulae (2), (4), (5)
Save the weight vector $\boldsymbol{w_q}$ of $\boldsymbol{q}$
By predicting the sub-region $\boldsymbol{\Omega}_d$ that $\boldsymbol{Q}$ may belong to by MCSVM algorithm
For $\boldsymbol{B_c}$ in $\boldsymbol{\Omega_d}$ do:
    Compute $D_c(\boldsymbol{q}, \boldsymbol{B}_c)^2$ // Based on Formula (8)
End for
The nearest neighbor goal of $\boldsymbol{q} = D_c(\boldsymbol{q}, \boldsymbol{B}_c)^2$ minimum $\boldsymbol{B}_c$

4.2.2. *Historical data compensation.* In the indoor environment, numerous factors contribute to signal interference, potentially causing the RSS to deviate from the expected values. This deviation can lead to significant errors in the positioning results. While incorporating a variable weight similarity measure in the search process can enhance search accuracy, the presence of outliers may lead to misjudgment of partition by MCSVM.

Consequently, the approximate nearest neighbor search will lose its value under such circumstances. This paper aims to optimize the influence of outliers on the online search algorithm by implementing historical data compensation. Additionally, it seeks to confine the location system's search for nearest neighbor targets to a single sub-region, with the goal of reducing the complexity associated with optimal target matching in traditional methods.

Considering that the maximum step size of pedestrians is denoted as $\lambda_{\max}$ [17], the margin of error for the two predictions falls within the range of $[0, \lambda_{\max})$. The disparity between the historical and current positions serves as a discriminant index for assessing the necessity of historical data compensation. The method proposed is outlined as follows:

1) If it surpasses the margin of error:

a. Computing the mode of prediction results of the previous a times MCSVMS as the new sub-region number;

b. Employing the proposed online search method to re-search for the nearest neighbor target of the sample to be queried within the new sub-region.

2) Provided that it does not surpass the margin of error:

The proposed online search method is employed to directly search for the nearest neighbor target within the sub-region predicted by MCSVM.

5. Experimental Verification and Analysis.

5.1. Construction of experimental environment. The experimental setup is located in the L-shaped corridor area on the fourth floor of section C in the Internet of Things Engineering College. The origin is situated at the lower left corner, denoted as point A, with the east and north directions establishing the physical coordinate system for the X and Y axes, as illustrated in Figure 2. The experimental setup utilizes the HUAWEI nova 10 device running on HarmonyOS 3.0 and a low-power beacon equipped with the ATM2202 Bluetooth chip. A series of Bluetooth beacons are strategically positioned at

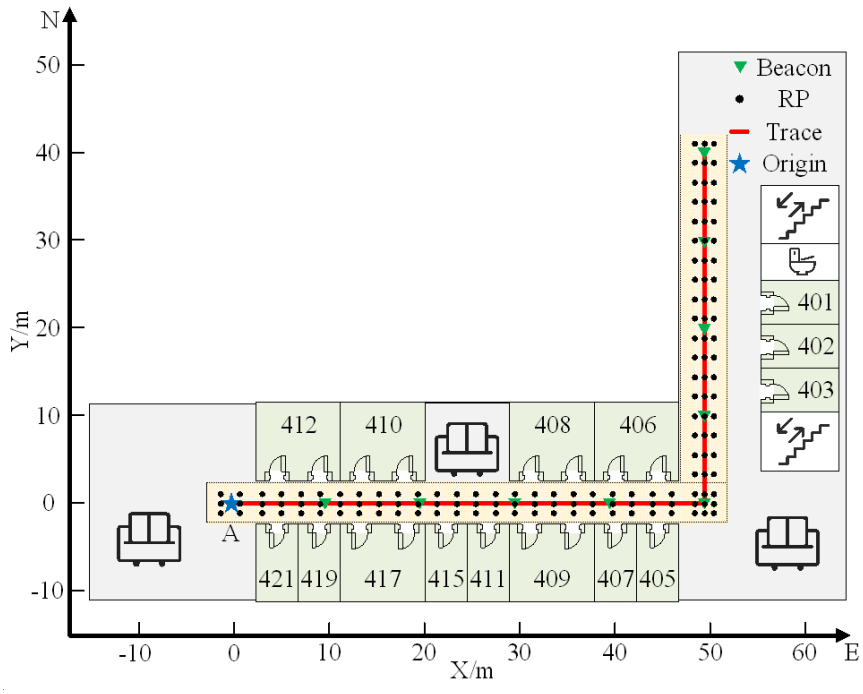

FIGURE 2. Experimental scenario

10-meter intervals within the designated area, totaling 10 beacons in all. Set the beacon's broadcast interval to 200 ms and the broadcast power to 0 dBm. The experimental data acquisition APP is developed using the Android Studio platform, with the data sampling frequency configured at 50 Hz.

A total of 501 reference points are chosen within the positioning area, with a 50 cm distance between each adjacent point. 50 sets of data are gathered at each reference point. After data cleaning, the mean data is computed and combined with the physical coordinates of the point to create fingerprints. This process is used to establish an offline fingerprint library $\mathbf{H}_{m\times(n+1)}$ (hereafter referred to as $\mathbf{H}$), which is further divided into the training set $\mathbf{H}_1$ and the test set $\mathbf{H}_2$. During the online phase, 167 reference points are selected at random. The application (APP) utilizes the device's Bluetooth module to conduct a scan of the Bluetooth beacon within its vicinity and acquire the signal strength. This signal strength is then utilized as the online phase test set $\mathbf{H}_3$ and stored locally on the device before being transmitted to the PC for analysis, where $\mathbf{H}_3 \cap \mathbf{H} = \varnothing$.

5.2. **Performance evaluation of feature extraction methods.**

5.2.1. *Feasibility analysis.* During the offline stage, the outcomes of location area zoning will have a direct impact on the positioning results during the online phase. The changing trends of RSS sample features in two distinct sub-regions are compared to validate the feasibility of the proposed feature extraction method in the offline phase. Randomly collect 15 RSS sample data points within sub-region 1 and sub-region 3, respectively. The sample feature change curve in the sub-region is plotted with the Bluetooth beacon number (sample feature number) on the x-axis and the RSS (sample eigenvalue) of the Bluetooth beacon on the y-axis, as depicted in Figure 3. The figure illustrates that the trends of various curves within the same region exhibit similarities, while the differences in the trends of curves across different regions are pronounced. Without loss of generality, the evolving pattern of various sample characteristics in relation to the median weight

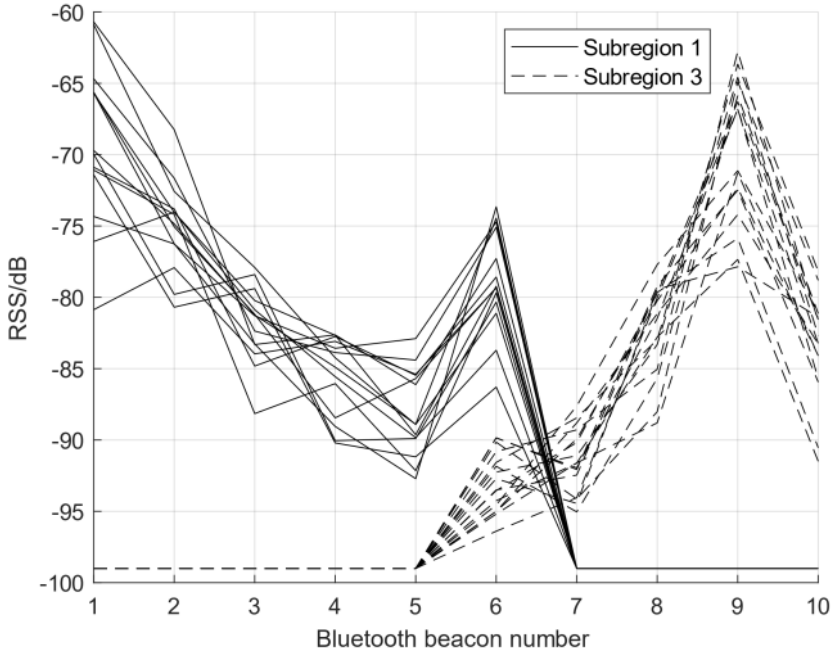

FIGURE 3. Change trend map of sample characteristics in sub-region

within the identical region mirrors that of the initial data. Consequently, the feature extraction method introduced in this paper can be applied to process the original fingerprint database.

5.2.2. *Reliability analysis.* To assess the efficacy of the feature extraction method in partitioning the location region, the prediction accuracy of the dataset is compared before and after feature extraction in the context of the MCSVM algorithm. The eigenvalue matrix is generated by extracting the feature of the original fingerprint database $\mathbf{H}$. The location area is divided into sub-regions of varying sizes by taking 59 data points as the partition interval within the range $[1, 60)$ at 1-meter intervals. The original fingerprint library $\mathbf{H}$ and eigenvalue matrix $\boldsymbol{B}$ were labeled according to different partition intervals, and subsequently divided into training sets $\left\{\mathbf{H}_1^1, \mathbf{H}_1^2, \mathbf{H}_1^3, \ldots, \mathbf{H}_1^{59}\right\}$, $\left\{\boldsymbol{B}_1^1, \boldsymbol{B}_1^2, \boldsymbol{B}_1^3, \ldots, \boldsymbol{B}_1^{59}\right\}$ and test sets $\left\{\mathbf{H}_2^1, \mathbf{H}_2^2, \mathbf{H}_2^3, \ldots, \mathbf{H}_2^{59}\right\}$, $\left\{\boldsymbol{B}_2^1, \boldsymbol{B}_2^2, \boldsymbol{B}_2^3, \ldots, \boldsymbol{B}_2^{59}\right\}$. The superscript of each collection denotes the partition interval. Subscript 1 denotes the training set, while subscript 2 denotes the test set. The MCSVM algorithm is employed to iteratively train the model 59 times using distinct training sets, and the resulting test set is utilized to assess the predictive accuracy of the model. The prediction accuracy curve of the dataset before and after feature extraction on the multi-classification SVM model under various partition intervals is depicted in Figure 4.

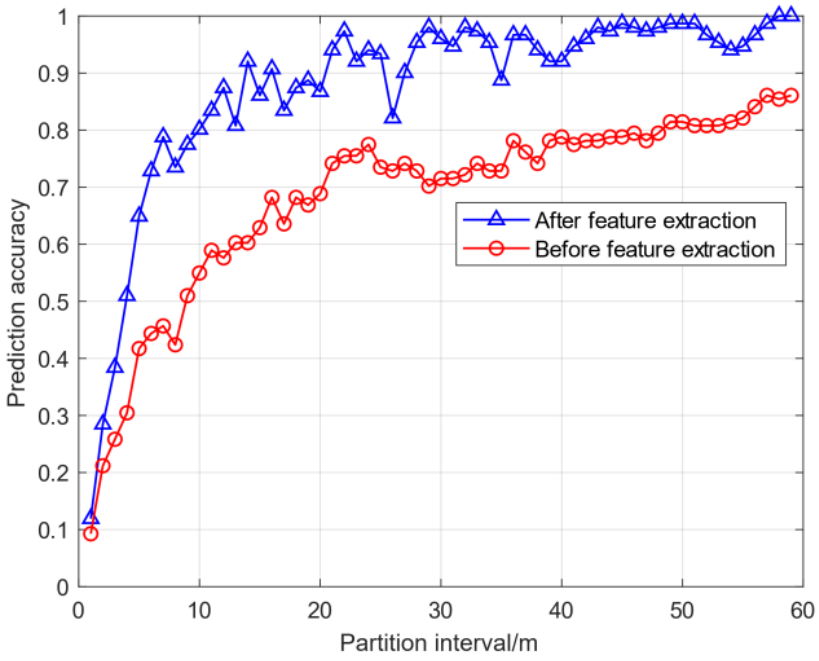

FIGURE 4. Prediction accuracy of MCSVM before and after feature extraction

Figure 4 illustrates that the prediction accuracy of the dataset before and after feature extraction on the MCSVM model exhibits an upward trend with the increase of partition interval. Moreover, when employing the identical algorithm model, the prediction accuracy of the dataset is higher after feature extraction using the MCSVM model compared to before feature extraction throughout the entire experiment. The average prediction accuracy following feature extraction is approximately 88%, whereas prior to feature extraction, it is around 68%. The eigenvalue matrix transformation method proposed in this study has been observed to significantly enhance the predictive capabilities of the MCSVM algorithm and increase the dependability of the positioning system.

5.3. **Optimal partition interval selection.** To guarantee the optimal positioning performance of the positioning system, it is essential to ascertain the optimal partition interval for the eigenvalue matrix and to assign labels to the eigenvalue matrix based on the optimal partition interval.

Simultaneously, $\mathbf{H}_3$ is utilized as the test set and the original fingerprint database $\mathbf{H}$ is employed as the training set to assess the adaptability of the algorithm. The eigenvalue matrix $\boldsymbol{S}$, tagged from feature extraction, is input into the MCSVM algorithm for model training. Subsequently, the test set $\mathbf{H}_3$ is predicted, and the curve depicting the relationship between partition interval and prediction accuracy is illustrated in Figure 5.

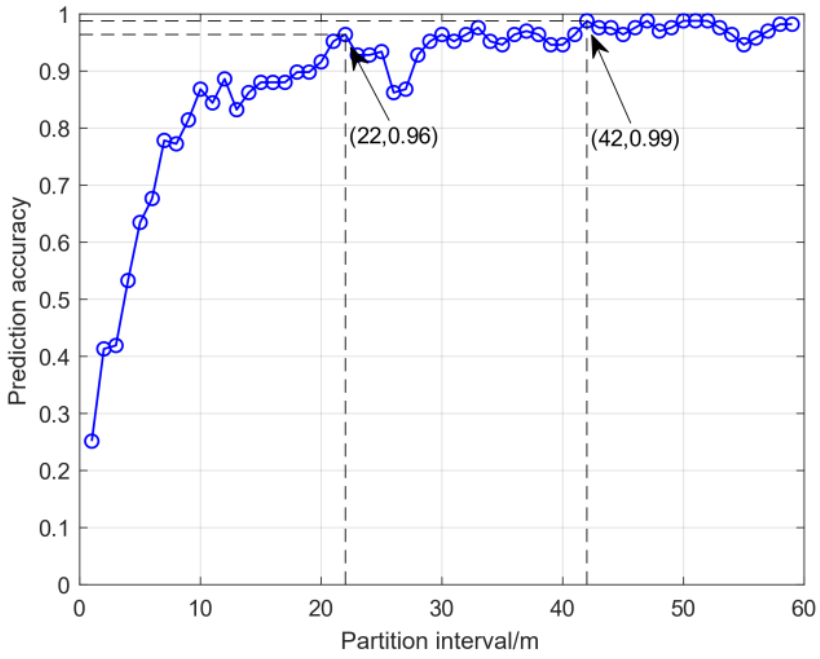

FIGURE 5. Prediction accuracy of MCSVM under different partition intervals

The chart illustrates that, according to the analysis in Section 5.2, the prediction accuracy of the MCSVM model on the test set X5 demonstrates an increasing trend as the partition interval increases. Upon using a partition interval of 22 meters, the prediction accuracy initially reaches its peak at 96%. The decline in the curve at 22-25 meters can be attributed to the instability of the RSS signal, which is caused by the open rest area within the positioning area at 20-30 meters. When the partition interval is 42 meters, the prediction accuracy reaches its second maximum of 99%.

Owing to the constraints of the physical space within the location area, setting the optimal partition interval at 42 meters would result in the division of the entire location area into only two sub-regions. While the prediction accuracy is notably high, the online phase necessitates the search for the optimal target within a large dataset, thereby hindering the guarantee of real-time positioning. As a result, the optimal partition interval is selected as 22 meters.

5.4. **Analysis of experimental results.**

5.4.1. *Positioning accuracy analysis.* In order to assess the optimization capability of the proposed method during the online stage, the algorithm is compared with the traditional

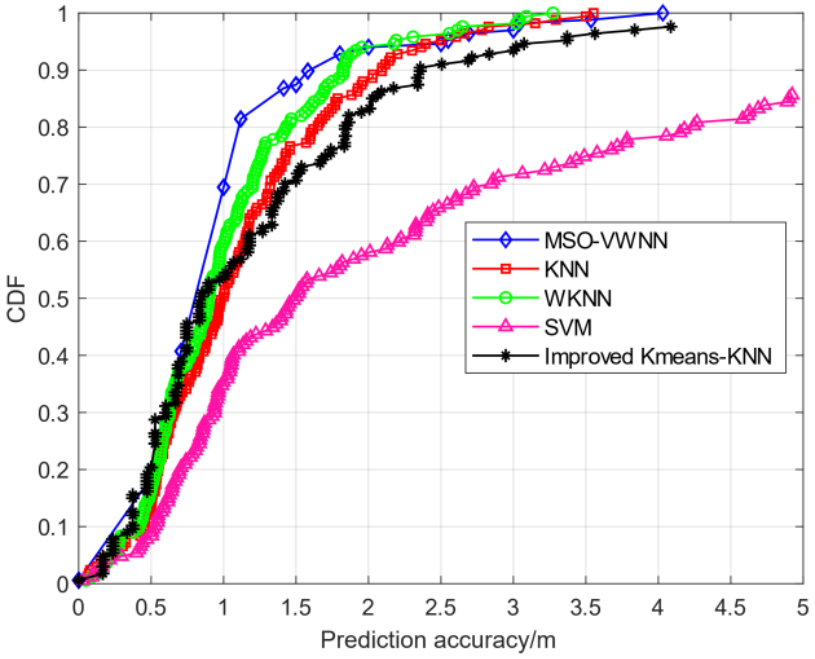

FIGURE 6. Cumulative probability distribution of positioning error

KNN algorithm [9], WKNN algorithm [10], SVM algorithm [12], and Kmeans-KNN algorithm [18] enhanced by agglomerative hierarchical clustering fusion. This analysis is conducted to evaluate the positioning accuracy and to generate the cumulative probability curve of positioning error on both the training set $\mathbf{H}$ and the test set $\mathbf{H}_3$, as depicted in Figure 6.

The figure illustrates that, with the exception of the SVM algorithm, the cumulative probability curve for positioning error within 0.75 m is nearly identical for the proposed algorithm, traditional KNN algorithm, WKNN algorithm, and improved Kmeans-KNN algorithm. The growth rate of the positioning error probability curve of the algorithm proposed in this paper exceeds that of other algorithms within the range of 0.75-1.5 meters. The probability of positioning error within 1 meter resulting from the algorithm proposed in this paper is 19.2% greater than that of the traditional KNN algorithm, 9.0% greater than that of the WKNN algorithm, and 15.0% greater than the improved Kmeans-KNN algorithm. The likelihood within a 1.5 m radius is 10.8% greater than that of the conventional KNN algorithm, 6.0% higher than the WKNN algorithm, and 16.8% higher than the enhanced Kmeans-KNN algorithm. The superiority of the proposed algorithm is attributed to the feature extraction of the original fingerprint database in the offline stage, which maximizes the effective information of the original data and fully expresses the data distribution of the location scene. Additionally, the online stage combines variable weight similarity measurement and historical data compensation, allowing for the full utilization of effective sample data information and the mitigation of the impact of irrelevant information on positioning accuracy. Consequently, the algorithm presented in this paper has the potential to significantly enhance the inadequate positioning accuracy of conventional positioning algorithms, thereby substantially improving the precision and resilience of the positioning system.

5.4.2. *System stability analysis.* To assess the stability of the proposed method's location, the maximum location error (MLE), average location error (ALE), standard deviation

TABLE 1. Error comparison of location algorithms

Units: m

Algorithm	MSO-VWNN	KNN	WKNN	SVM	Improved Kmeans-KNN
MLE	4.03	3.55	3.27	19.69	9.68
ALE	**0.80**	1.13	1.02	3.05	1.28
SD	0.72	0.70	**0.62**	3.80	1.23
RMSE	**1.08**	1.33	1.19	4.87	1.77

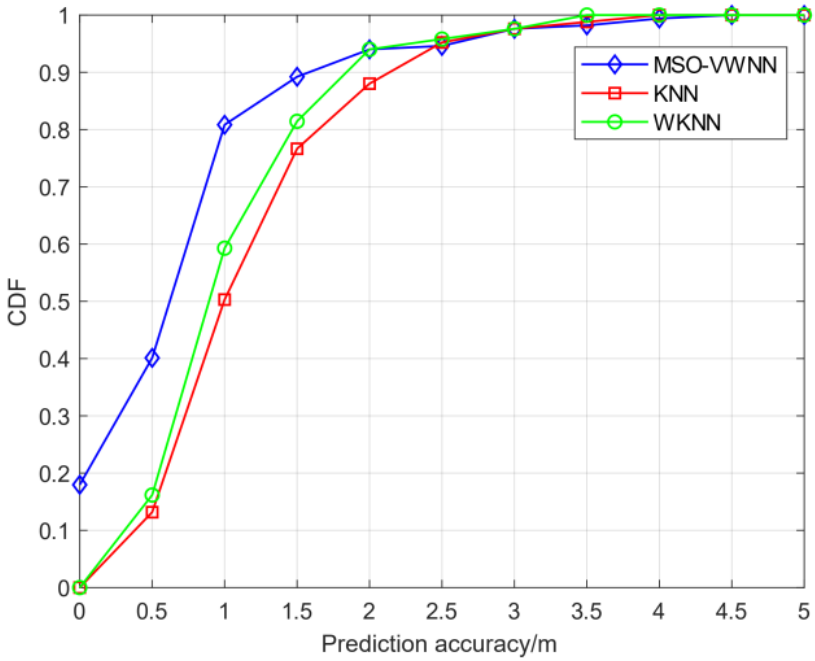

FIGURE 7. Cumulative probability distribution of error occurrence frequency

(SD), and root mean square error (RMSE) of the proposed algorithm are compared with those of the traditional KNN algorithm, WKNN algorithm, SVM algorithm, and the improved Kmeans-KNN algorithm in the test set, as presented in Table 1.

As shown in the table, the average error of the proposed algorithm is approximately 0.80 m, which is 0.22 m higher than that of the optimal algorithm compared to the other four algorithms. Meanwhile, the root mean square error of the proposed algorithm is about 10.8 m, which is 0.11 m higher than that of the optimal algorithm among the other four algorithms. This demonstrates that the algorithm proposed in this paper exhibits minimal deviation between the predicted value and the actual value during online positioning, resulting in excellent positioning performance.

To better demonstrate the benefit of the proposed algorithm's location stability, the distribution of location errors in the proposed algorithm is compared with those of the traditional KNN algorithm and WKNN algorithm in the test set $\mathbf{H}_3$. This comparison involves analyzing the frequency of different positioning errors in the three algorithms. The cumulative probability curve of different error occurrence frequencies is plotted with the positioning error on the x-axis and the cumulative probability of the occurrence frequency on the y-axis, as depicted in Figure 7.

As depicted in the figure, the frequency of occurrence of the algorithm proposed in this paper is 17.96% higher than that of the other two algorithms. Moreover, the frequency of

location errors within 0.5 meters is 26.95% higher than that of the KNN algorithm and 23.95% higher than that of the WKNN algorithm. Additionally, the frequency of location errors within 1 meter is 30.54% higher than that of the KNN algorithm and 21.56% higher than that of the WKNN algorithm. Moreover, the algorithm proposed in this paper exhibits a positioning error frequency of 89.22% within a range of 1.5 meters, indicating that the majority of positioning errors associated with the algorithm are concentrated within the $[0, 1.5]$ meter range. While Table 1 indicates that the proposed algorithm's overall positioning error is marginally more dispersed compared to the KNN and WKNN algorithms, the positioning error distribution is predominantly concentrated in positions with small errors. Consequently, overall, the algorithm suggested in this paper demonstrates superior location performance.

5.4.3. *Analysis of location efficiency.* To assess the efficacy of the proposed algorithm in optimizing online search efficiency and enhancing positioning accuracy, the average prediction time of the proposed algorithm is compared with that of the traditional KNN algorithm, WKNN algorithm, SVM algorithm, and the improved Kmeans-KNN algorithm. Five algorithms were employed to forecast the optimal target within the test set, and the average prediction time for each algorithm was computed, as presented in Table 2.

TABLE 2. Average prediction time of online phase algorithm

Units: ms

Algorithm	MSO-VWNN	KNN	WKNN	SVM	Improved Kmeans-KNN
Average prediction time	**0.31**	0.45	0.45	0.49	0.67

The data presented in the table indicates that the average time of the algorithm proposed in this paper is lower than that of other algorithms. This suggests that the algorithm not only enhances positioning accuracy but also effectively ensures positioning efficiency.

6. **Conclusion.** This paper proposes a novel feature extraction method in consideration of the impact of the distribution shape of the data set and the diversity of data content on the location results. This method ensures the preservation of the original data, alters the data distribution, maximizes data complexity, and efficiently conserves computational resources.

This study aims to address the complexity of determining the optimal target in fingerprint location and the limited accuracy of traditional methods. The paper proposes an approach that combines variable weight similarity measure and historical data compensation to enhance the traditional approximate nearest neighbor method. This approach not only facilitates rapid identification of the optimal target but also ensures the accuracy of target prediction, thereby effectively enhancing the positioning performance of the system.

It is important to acknowledge that the strength and characteristics of fingerprint signals in indoor environments will vary with changes in time, location, and environmental conditions. The discrete and unstable nature of this phenomenon presents challenges for the positioning system in accurately identifying and tracking fingerprints. Hence, the primary focus of research on fingerprint location will be the investigation of the impact of device power consumption, scene change, and environmental change on the failure of fingerprint mapping.

In the next stage, we will continue our study on the issue of fingerprint heterogeneity resulting from diverse devices and the challenge of fingerprint map failure due to environmental changes.

REFERENCES

[1] C. Y. Zhou, J. Y. Liu, M. Sheng et al., Exploiting fingerprint correlation for finger-print-based indoor localization: A deep learning based approach, *IEEE Transactions on Vehicular Technology*, vol.70, no.6, pp.5762-5774, 2021.

[2] A. M. S. Chong, B. C. Yeo and W. S. Lim, Automatic data acquisition system for Wi-Fi fingerprint-based indoor positioning system, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.231-252, 2022.

[3] S. G. Lee and C. Lee, Developing an improved fingerprint positioning radio map using the k-means clustering algorithm, *2020 International Conference on Information Networking (ICOIN)*, pp.761-765, 2020.

[4] J. Ren, K. Bao, G. Zhang et al., LANDMARC indoor positioning algorithm based on density-based spatial clustering of applications with noise-genetic algorithm-radial basis function neural network, *International Journal of Distributed Sensor Networks*, vol.16, no.2, 2020.

[5] J. Zhai, T. Li, F. Huang et al., An indoor positioning algorithm based on hierarchical clustering and adaptive weighted K-nearest neighbor combination, *Journal of Time and Frequency*, vol.43, no.4, pp.300-309, 2020.

[6] P. Fränti and S. Sieranoja, How much can K-means be improved by using better initialization and repeats?, *Pattern Recognition*, vol.93, pp.95-112, 2019.

[7] C. Li and Y. Mao, Improved indoor localization algorithm combining K-means clustering algorithm and Wasserstein generative adversarial network algorithm, *2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp.1-5, 2023.

[8] H. A. Abbas, N. W. Boskany, K. Z. Ghafoor et al., Wi-Fi based accurate indoor localization system using SVM and LSTM algorithms, *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pp.416-422, 2021.

[9] H. Ai, W. Zeng, J. Tao et al., Indoor location RF fingerprint data enhancement method based on diffusion model, *Journal of Communications*, http://kns.cnki.net/kcms/detail/11.2102.TN.20231025.0921.002.html, pp.1-13, 2023.

[10] S. Liu, R. de Lacerda and J. Fiorina, WKNN indoor Wi-Fi localization method using k-means clustering based radio mapping, *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pp.1-5, 2021.

[11] T. Xu, J. He, N. Zhu et al., Improved VWKNN fingerprint location algorithm based on discrete coefficients, *Journal of Beijing University of Aeronautics and Astronautics*, vol.48, no.7, pp.1242-1251, 2022.

[12] A. Affan, H. M. Asif and N. Tarhuni, VLC indoor positioning using RFR and SVM reduced features machine learning techniques, *2023 Wireless Telecommunications Symposium (WTS)*, pp.1-6, 2023.

[13] X. Tang, Q. Zhang, J. Wang et al., Research on wireless network indoor location system based on accurate time measurement, *Journal of Computer Science*, vol.45, no.3, pp.567-584, 2022.

[14] K. Liu, G. Lu and Y. Ma, Research on indoor location of passive tags based on azimuth optimization multi-information fingerprint, *Journal of Tianjin University (Natural Science and Engineering Technology Edition)*, vol.53, no.3, pp.221-228, 2020.

[15] A. Z. Broder, On the resemblance and containment of documents, *Proc. of Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp.21-29, 1997.

[16] L. Niu, Z. Xu, L. Zhao et al., Residual vector product quantization for approximate nearest neighbor search, *Expert Systems with Applications*, 232, 2023.

[17] X. Ma and T. Guo, A case study on the variation of 100m step length, step frequency and whole speed of men majoring in physical education, *Sports World (Academic Edition)*, no.4, pp.113, 115, 2018.

[18] C. Zhang, F. Zhang, Y. Liu et al., Optimization of bluetooth fingerprint indoor location algorithm based on fusion clustering, *Computer Simulation*, vol.37, no.7, pp.314-318, 2020.

Author Biography

Shunyuan Sun received the Ph.D. degree in Control Science and Engineering, from Jiangnan University, China, 2014. He is currently a fulltime associate professor at the School of Internet of Things Engineering, Jiangnan University, China. His research interests include sensors and instrumentation, industrial IOT applications, wireless sensor network routing and location technology. He has long been engaged in research and product development of process control and optimization, industrial sensor and instrumentation design, industrial Internet of Things applications, wireless sensor network systems and communication-related technologies. In the past 3 years, he has established long-term cooperative relations with more than 30 instrument and meter enterprises, and entrusted more than 30 projects. He has published over 30 papers in journals and conferences.

Jingyuan Yu received the B.S. degree in Electrical Engineering and Automation from Changzhou Institute of Technology, China, in 2021. He is currently pursuing his M.Sc. degree in Electronic Information at Jiangnan University. His main research interests include wireless sensor network.

Ningning Qin received the B.Sc. degree in Electronic Information Engineering from Jiangnan University, China, 2002; the Ph.D. degree in Light Industry Information Technology and Engineering, from Jiangnan University, China, 2008. She is currently a fulltime professor at the School of Internet of Things Engineering, Jiangnan University, China. Her research interests include network routing control, coverage control theory, platform system design of sensor network information utilization and analysis approximate reasoning. She has published over 50 papers in journals and conferences.