IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2023, accepted 8 June 2023, date of publication 16 June 2023, date of current version 27 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3286860

RESEARCH ARTICLE

Traffic Processing Model of Big Data Base Station Based on Hybrid Improved CNN Algorithm and K-Centroids Clustering Algorithm

XIEFEI HE[1], TAO YU[2], YANG SHEN[1], AND SHI'AN WANG[1]
[1]College of Information Engineering, Guangzhou Institute of Technology, Guangzhou 510075, China
[2]School of Information, Guangdong Polytechnic of Science and Trade, Guangzhou 511500, China

Corresponding author: Xiefei He (hexiefei_2021@163.com)

This work was supported in part by the Characteristic Innovation (Natural Science) Project of Colleges and Universities in Guangdong Province through the Project "Design and Implementation of Big Data Privacy Security Protection System Based on CDH (Cloudera's Distribution Including Apache Hadoop) Platform" (No. 2021KTSCX265) and in part by the Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau "Research and Application of Big Data Integration Framework for Pipeline Type NC (Numerical Control) Machining Equipment" (No. 202235310).

ABSTRACT Wireless communication network (WCN) is very important for providing convenient mobile network communication services. Random Phase Multiple Access (RPMA) WCN under heterogeneous network architecture is widely used in wireless network construction worldwide due to its low power consumption and high cell density. However, this kind of WCN cannot meet the application scenario of high communication quality. Therefore, this research builds an RPMA communication quality prediction model for big data wireless base stations, which combines the convolutional neural network (CNN) algorithm and the lifting regression tree algorithm. It can be used to find the elements that have obvious influence on the communication quality. The reason for choosing the convolutional neural network algorithm is that its nonlinear feature relationship search and processing ability is excellent, and its computational complexity is relatively small. However, it is a black box algorithm and cannot obtain the importance coefficients of each feature, which is not conducive to subsequent analysis. Therefore, it is also necessary to select a relatively simple modified regression tree algorithm to participate in the calculation. The model constructed by integrating convolutional neural network and lifting regression tree algorithm is the RPMA wireless big data base station communication quality prediction model. A base station planning and deployment model based on weighted K-centroids algorithm is designed to obtain a better base station deployment scheme. In the CNN-DT model, the importance coefficients of T_B_diff, P_La and P_Lo are the largest and significantly larger than those of other features, which are 0.352, 0.289 and 0.264 respectively. The weighted K-centroids clustering algorithm designed in this study has the best overall downlink reception signal (RSSI) value distribution. For RSSI bucket "- 140∼- 130", the number of test points of WK centroids model, K-means model, GMC model, Mean Shift model and spectral clustering model accounted for 1.95%, 6.25%, 4.25%, 8.22% and 7.13% respectively. The model constructed in this study based on CNN and improved regression tree algorithm can accurately predict the communication quality of wireless big data base stations. This study's main contribution is that it can be used in conjunction with the base station planning and deployment model based on weighted K-centroids algorithm to improve the accuracy and effectiveness of location selection for the RPMA wireless communication network base station.

INDEX TERMS CNN, promote the regression tree, K-centroids, big data base station, WCN.

I. INTRODUCTION

The Internet of Things (IoT) is currently developing quickly, and in a short time, there will be 100 billion IoT

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

terminals connected, with a gigabit data transmission rate. The traditional short-range wireless technology and cellular network technology can no longer meet the needs of the IoT service [1]. For this reason, a new communication mode, i.e. low-power wide-area network, is gradually popularized. This communication mode can effectively make up for the shortcomings of existing cellular network and short-range wireless technology, and meet the diversified application requirements of the IoT [2]. The heterogeneous communication system developed based on low-power WAN has the advantages of small physical size of base stations, low transmission power and low construction cost. The deployment of small base stations can be increased according to business requirements [3], [4]. For example, in high-traffic areas such as office buildings and supermarkets, the problem of signal attenuation caused by obstacles such as building walls can be avoided, thus compensating for the blind spots in the coverage of macro stations. In addition, in this case, heterogeneous communication systems can flexibly increase the network capacity of macro stations, reduce the load on macro stations, and compensate for the shortcomings of difficult deployment and high installation costs of macro base stations. Due to the widespread use of wireless networks, the coverage area of the network is gradually increasing. In a few application scenarios with high communication quality requirements, this communication network cannot meet the application requirements well. The normal operation of the communication network depends on a large number of base stations, and the distribution and construction of base stations will also significantly affect the operation of the communication network. Therefore, it is necessary to develop a more scientific and reasonable base station layout plan based on the characteristics of wireless communication systems. Despite the fact that earlier researchers have carried out academic research specifically addressing this type of issue, the methods offered either lack adequate automation or there are still substantial discrepancies between signal quality prediction results and actual quality levels, which is also the real reason for conducting this research. The main innovation of this study lies in the selection of CNN algorithm with strong nonlinear feature processing ability and regression tree algorithm to build an intelligent model that can predict the communication quality of large wireless data base stations, which can overcome the subjective impact of relying solely on manual experience to judge signal quality. Another innovation is the development of a base station location model for wireless large data base station communication based on improved K-centroids clustering. This model can automatically comprehensively consider various factors that affect the location of big data base stations, and make comprehensive optimization site selection decisions. Not only is the decision-making speed extremely fast, but also the evaluation method of each decision can be completely consistent. The theoretical contribution of this study is to provide a different RPMA wireless big data base station communication quality intelligent prediction method and an automated base station deployment scheme that is different from manual methods.

II. RELATED WORKS

To as accurately estimate the communication quality of WCN as feasible and to achieve better wireless network base station planning schemes, industry experts and academics have conducted a significant amount of study. Liu C and others found that different users in cellular wireless networks have different traffic status and communication quality. To find out the factors that affect user traffic status and communication quality, the author's team designed a wireless network communication quality prediction model based on alternating direction multiplication. The prediction model designed in this study outputs the largest feature importance coefficient of the user's mobile phone hardware device and location. These two factors have the greatest impact on WCN quality [5]. There is a communication delay in the IoT communication network supporting unmanned vehicles. In response to this problem, Tranter T G et al. developed a communication quality demand prediction model for the IoT communication network for unmanned vehicles based on an integrated machine learning algorithm. The main factor affecting the communication quality of IoT for unmanned vehicles is the driving speed. Therefore, the study proposes that limiting the driving speed of unmanned vehicles will appropriately increase the density of communication network base stations in the high-speed driving section of unmanned vehicles [6]. Yin F and others found that it is difficult to plan the base station of the medical IoT. Poor planning of the base station will significantly increase the maintenance cost of the subsequent IoT. Therefore, this study designed a medical Internet base station planning model based on an improved K-means algorithm, which can significantly improve the quality of location planning of medical Internet base stations [7]. Chen Q and others found that there are many underground problems of communication quality in WCN of mining area based on RPMA technology. Therefore, the author team collected a large number of data on the use of wireless Internet users in mines, and used statistical methods to analyze the hidden characteristics of the data. The communication quality of users deep in the ground and between higher mountains will be significantly reduced. The problem of insufficient communication quality can be solved by arranging increasingly dense base stations in these locations [8]. The research team of Hga B found that some low quality WCNs are often affected by the transmission environment noise. Therefore, the design is in a wireless communication signal filtering technology based on duty cycle scheduling. This filtering method significantly reduces the communication noise in the low-quality WCN and reduces the probability of the delay problem in the user communication process [9]. Wang T and others found that the space dynamic communication base station has the characteristics of high mobility, flexible deployment and wide coverage. It has a wide range of needs in the fields of emergency communications, earthquake relief and island communications. To solve the problem of the limited number of users caused by the limited bandwidth of the space base station, an airship laser communication terminal applied to the space dynamic base station is designed. This terminal has a capture probability of better than 98%, which satisfies the requirements for bidirectional transmission of high-speed space signal laser carriers between spacecraft. The tracking accuracy is better than 6-15 Rad, and the communication rate is 2.5 Gbps. The typical capture time is 5 seconds [10]. Sharma R and others believed that when the number of users accessing wireless sensor networks is large, it is necessary to take measures to save the energy consumption of network operation. Therefore, the author team developed a base station location planning algorithm for wireless sensor networks based on the improved K-centroids clustering algorithm. The energy consumption of the wireless sensor network structure designed by this algorithm is 16.89% lower than that of the manual planning scheme under the condition of large user access [11].

In conclusion, a number of studies have been conducted to identify the variables that affect the communication quality in the big data image network in an effort to improve the planning quality of the base station in the big data base station. Various targeted application models are designed. However, these research results rarely involve intelligent planning and location of RPMA communication network base stations in low-power WAN. This type of wireless network base station is widely used in the current society, and it is necessary to conduct research, which is also one of the starting points of this study. The main contribution of this study is the successful establishment of a model based on convolutional neural networks and the lifting regression tree algorithm, which can be used to predict the communication quality of big data wireless base stations. At the same time, the study proposed a base station planning and deployment model based on the weighted K-centroids algorithm. These achievements are expected to be applied to the site selection work of RPMA wireless communication networks, reducing the energy consumed by engineers in these tedious and repetitive tasks.

III. BIG DATA BASE STATION TRAFFIC PROCESSING MODEL INTEGRATING CNN AND K-CENTROIDS CLUSTERING ALGORITHM

A. STRUCTURAL DESIGN OF BIG DATA BASE STATION TRAFFIC PROCESSING MODEL BASED ON HYBRID ALGORITHM

RPMA in low-power WAN is an advanced wireless communication technology. Compared with the traditional low-power wide-area network technology, RPMA technology has greater performance advantages. However, due to the high density of signal base stations and uneven service distribution, the corresponding network planning is difficult [12], [13]. Therefore, an RPMA large data base station signal quality prediction and base station deployment model with multiple algorithms is now developed.

The application scenario of this study is shown in Figure 1. In this graph, the RPMA network has a star topology and connects multiple terminals to the nearby RPMA base station by wireless means. The function of the base station is to receive the data uploaded by the terminal and transmit it to their respective backhaul connections.

Considering the form of RPMA network, a network planning method integrating multiple machine learning algorithms is proposed. The calculation process is shown in Figure 2. The planning system must use network base station information, terminal test pilot data, terminal geographic location data, etc. in order to take into account the operational objectives of the RPMA large data base station network [14]. The system must first clean the collected data, remove the features that contain default values and a large number of duplicate data, and extract various features that affect the signal coverage quality. Then, the selected features are fed into the prediction model to train the best signal coverage quality prediction model [15], [16]. Specifically, the analysis model shown in Figure 2 is mainly composed of the signal quality prediction model and the base station deployment planning model. It is necessary to carry out the next BTS adjustment according to the BTS signal and space coverage after the last deployment [17]. To estimate the signal impact of the base station change and to serve as a foundation for subsequent base station adjustments, it is important to create a signal quality prediction model.

Figure 2 shows that the target of the coverage terminal has a significant impact on the coverage strength of the wireless network, so it is necessary to adjust the base station location according to the coverage. In general, insufficient signal strength is the main reason for poor regional coverage, and there are three reasons for this. The first is the base station. The antenna height, antenna gain, antenna azimuth and transmitting power of the base station will affect the signal strength. The second is the transmission path. Shadow fading and obstructions will cause path loss. Finally, there is interference. The overlapping part of the signal coverage area of several adjacent base stations will have co-frequency interference. Some buildings and prominent terrain on the surface may also cause multiple disturbances. To predict the signal quality of the base station, it needs to consider the mapping relationship between these three factors and the signal quality.

B. BASE STATION SIGNAL QUALITY PREDICTION MODEL BASED ON HYBRID CNN AND GRU ALGORITHM

Predicting signal quality is fundamentally a regression task that may be handled by conventional machine learning techniques. The decision tree algorithm is chosen in this instance to take part in the prediction. However, the original data integration and extraction ability of the decision tree algorithm is very weak, unable to find the implicit

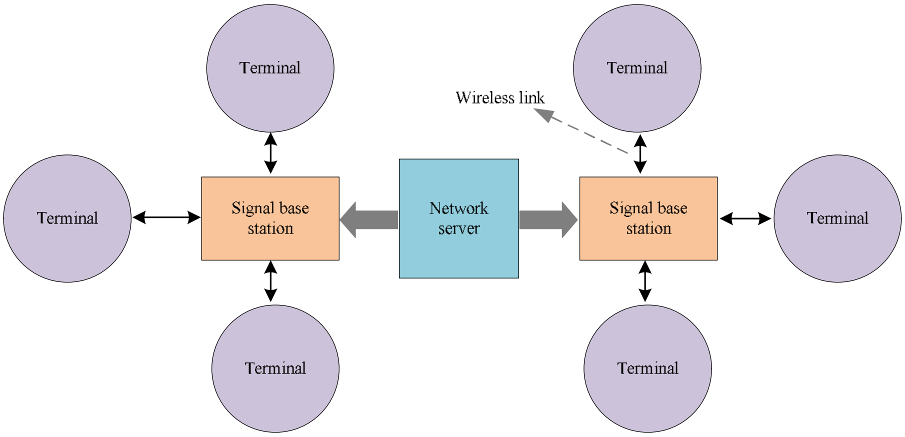


FIGURE 1. Typical RPMA network topology.

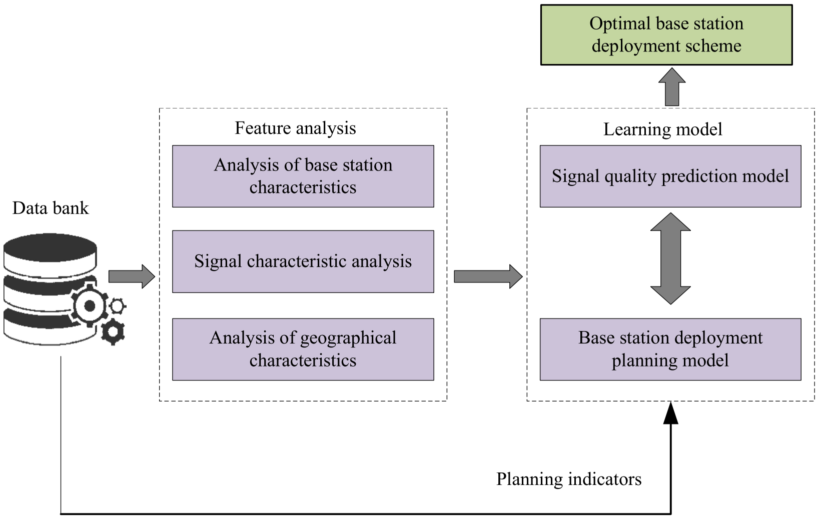

Optimal base station deployment scheme
Feature analysis
Analysis of base station characteristics
Signal characteristic analysis
Analysis of geographical characteristics
Data bank
Learning model
Signal quality prediction model
Base station deployment planning model
Planning indicators

FIGURE 2. Calculation flow of RPMA big data base station network planning system.

relationship between input features, thus unable to achieve good prediction results. Therefore, the CNN algorithm and decision tree algorithm are combined to build a base station signal quality prediction model. Figure 3 depicts the model's organisational structure. The base station signal data that is then entered into the prediction model must first go through numerous pre-processing steps. The CNN model is then fed the feature association between high latitude and synthesis. After preprocessing, the original picture and the output data from the CNN method are then fed into the regression decision tree algorithm. After training the optimal decision tree model, the prediction result and the corresponding input feature importance coefficient are output.

The characteristics of the input data required by the model are shown in Table 1. Since the RPMA uplink received signal strength is generally close to the reception sensitivity, it is appropriate to select the downlink new received signal strength received by the terminal as an indicator to evaluate the signal quality. The process of establishing the prediction model is to find the mapping function f between the data set characteristics and the signal quality evaluation indicators, as shown in equation (1).

$$y_k = f(x_k) \tag{1}$$

x_k and y_k represent the characteristic variable and the predicted signal quality value of the model output in formula (1).

The current convolution layer neurons of CNN module are calculated according to formula (2).

$$X_j^l = f_n\left(\sum_{i\in M_j} X_i^{l-1} \cdot K_{ij}^l + b_j^l\right) \tag{2}$$

where X_j^l represents the output of the j neuron in the l layer, K_{ij}^l and b_j^l are the weight coefficient and bias coefficient of the corresponding neuron. F is the mapping function. Extracting the key features is also required to lessen computational complexity. As a result, the feature map must be compressed,

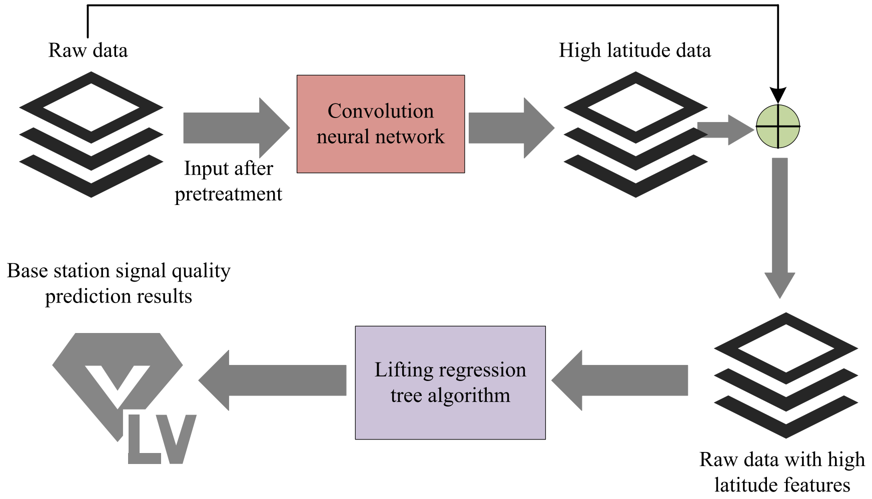

Raw data
High latitude data
Input after pretreatment
Convolution neural network
Base station signal quality prediction results
Lifting regression tree algorithm
Raw data with high latitude features
LV

FIGURE 3. Structure of base station signal quality prediction model based on hybrid decision tree and CNN algorithm.

TABLE 1. Input data attributes of signal quality prediction model (part).

Number	Attribute type	Attribute English name	Attribute Chinese name	Explain
#01	Base station side attributes	B_La	Base station longitude	Describe the base station location
#0		B_Lo	Base station latitude	
#0		B_H	Base station height	Base station basic properties
#0		B_po	Transmit power	
#0		A_H	Antenna hanging height	
#0		B_name	Base station name	
#0		Area number	Deployment area number	
#0		Last connection time	End connection time	Represents the latest communication status
#0		Last connection address	End connection address	
#0		UL_Per	Uplink bit error rate	Describe recent communication quality
		UL_SNR	Uplink SNR	
		UL_SF	Uplink spread factor	
	Test point side attribute	DL_RSSI	Downlink accepted signal strength indicator	Describe the communication quality of the test point
		DL_SF	Average spread spectrum factor of downlink transmission	
		P_La	Longitude of test point	Describe the test point location
		P_Lo	Test point latitude	

and the pooling layer's downsampling is determined using formula (3).

$$X_j^l = f\left(\beta_j^l down\left(X_i^{l-1}\right) + b_j^l\right) \tag{3}$$

$down(\cdot)$ is to calculate the maximum and average value of the characteristic value of the current layer in formula (3). Here, the loss function in the form of two-norm is used to construct the convolution neural network. The calculation method is shown in formula (4).

$$loss = \sum_{n=1}^{N} \|Y_n - \Gamma(X_n)\|^2 \tag{4}$$

$\Gamma(X_n)$ represents the prediction result of the neural network, and X_n is the input data in formula (4). To prevent the neural network from over-fitting, it is also necessary to add regularization items. The commonly used regularization methods are one-norm regularization L_{re1} and two-norm regularization. The calculation methods are shown in formula (5) and (6) respectively.

$$L_{re1} = L + \lambda \|\theta\| \tag{5}$$

$$L_{re2} = L + \lambda \|\theta\|^2 \tag{6}$$

In formulas (5) and (6), θ is the parameter to be optimized, λ is the weight attenuation coefficient and it is also the loss function. The two-norm regularisation term can suppress the parameter size while knowing that the parameter is not zero, and can also prevent the model from being too sparse. Therefore, the two-norm regularisation term is selected to design the neural network. Finally, the training method of the CNN is designed. The objective function is λ, and the optimal network parameter found is θ^*, so that equation (7) exists.

$$\theta^* = \arg\min_{\theta} L \tag{7}$$

In the study, the gradient descent method is used to optimize the network. In the calculation process, the derivative of λ

is required to obtain the gradient. Then the gradient is used to update the model until the model completes convergence, as shown in formula (8).

$$\theta_{j+1} = \theta_j + \frac{lr \cdot \partial L}{\partial \theta_j} \tag{8}$$

θ_j is the model parameter when iterating to j-th, $\frac{\partial L}{\partial \theta_j}$ is the calculated gradient, and lr is the learning rate of the neural network in formula (8). The characteristic data and original data output by CNN model will be input into the lifting regression tree algorithm. In this model, the base result $f_M(x)$ is generated in the form of addition, as shown in formula (9).

$$f_M(x) = \sum_{m=1}^{M} T\left(x; \gamma_m\right) \tag{9}$$

$T\left(x; \gamma_m\right)$ represents m decision tree models in formula (9). According to formula (10), the integrated tree model contains a total of m decision trees.

$$T\left(x; \gamma_m\right) = \sum_{j=1}^{J} c_j I, \quad \left(x \in R_j\right) \tag{10}$$

In formula (10), γ_m represents the internal parameters of m models. R_j is the region division of each tree model on the input variable set. The corresponding region constant is c_j. J is the number of leaf nodes in the decision tree. The regression integration algorithm uses the square error to calculate the loss function. After the training of the integrated regression algorithm is completed, the corresponding correlation measure $J_n(T)$ of the n-th input variable is calculated using equation (11).

$$J_n(T) = \sum_{i=1}^{L-1} \hat{j}_n^2 I(v_t = n) \tag{11}$$

T represents the decision tree with L leaf nodes in formula (11). $\hat{j}_n^2 I(v_t = n)$ represents the corresponding error improvement term when the current model takes X_n as the split variable on the t non-leaf node. Therefore, the loss improvement average of the decision tree can be obtained, as shown in formula (12).

$$J_n^2 = \frac{1}{M} \sum_{m=1}^{M} J_n^2(T_m) \tag{12}$$

At this point, the base station signal quality prediction model based on CNN and regression tree algorithm has been built, and the operation steps of the model are as follows. Step 1 is to pre-process the input data of the model; Step 2 is to input the processed data into the CNN module to extract high-latitude and comprehensive feature relationships; Step 3 is to input the output data of the CNN algorithm and the pre-processed original image into the regression decision tree algorithm to train the optimal regression decision tree. The final step 4 is to calculate the output prediction results and corresponding input feature importance coefficients based on the optimal decision tree module.

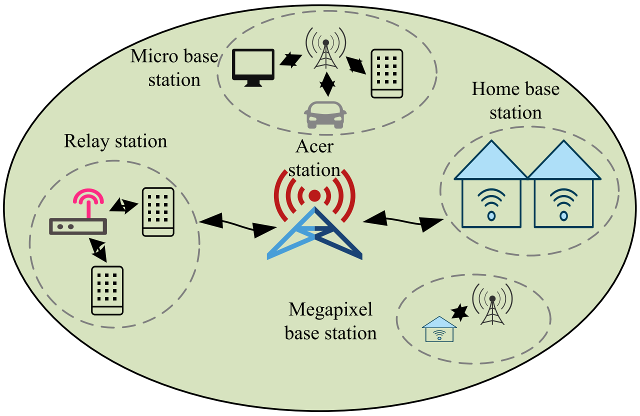


FIGURE 4. Typical architecture of heterogeneous network.

C. DESIGN OF BASE STATION DEPLOYMENT MODEL BASED ON WEIGHTED K-CENTROIDS CLUSTERING ALGORITHM

The deployment of heterogeneous network architecture has lower power consumption and higher cell density, which can well meet the communication needs of large base users. Its typical structure is shown in Figure 4. However, this type of base station distribution technique will cause a lot of interference with one another, making later network operation challenging and raising the cost of the base station network's construction. A base station deployment model based on the weighted K-centroids clustering method has now been developed to address these issues. The main reason for choosing the K-centroid clustering algorithm is that it has good clustering performance in large-scale, unstructured distributed clustering analysis and has been widely applied. Chen et al. used a variety of non K-centroids algorithms to cluster large-scale unstructured data, and found that the clustering results were inferior to the clustering effect of K-centroids clustering algorithm [18].

In the typical K-means clustering algorithm, it is considered that each data point is important to determine location cluster center equivalently. However, cluster center's stability determined according to this form of equal weight is insufficient, and the initial cluster center with poor quality may be generated. In order to measure the effect of data points on the position of the base station, weight is now introduced into this study rather than processing each point with equal weight. As a result, the K-centroids clustering algorithm is created and is based on the weighted concept. In the experiment, it is assumed that the algorithm contains n interrupt data points, which constitute set $P = \{p_1, p_2, \ldots, p_n\}$ and base station location set $B = \{b_1, b_2, \ldots, b_k\}$. The base station deployment planning model designed in this study is calculated based on the existing base station location. The current base station location and number can initialize clustering parameters, that is, the initialization center of the cluster and cluster numbers to be clustered. The normalized distance is used to calculate the subordinate

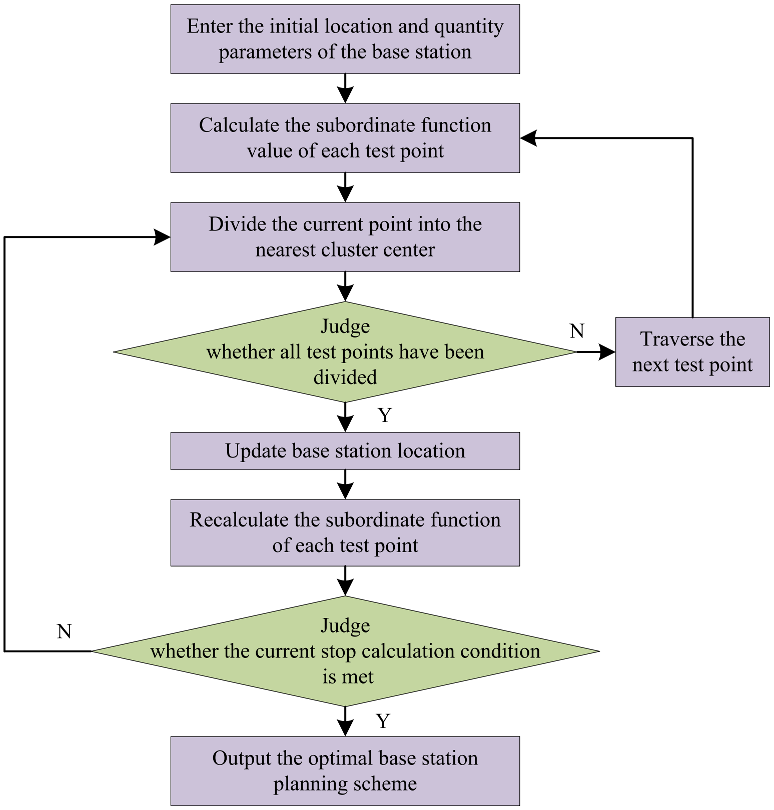
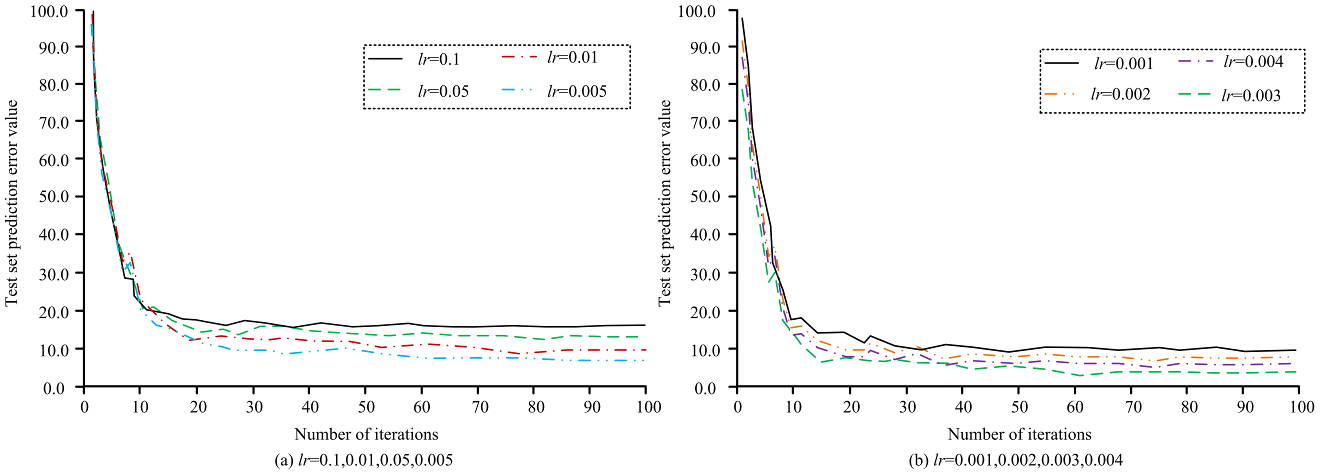

Enter the initial location and quantity parameters of the base station
Calculate the subordinate function value of each test point
Divide the current point into the nearest cluster center
Judge whether all test points have been divided
N
Traverse the next test point
Y
Update base station location
Recalculate the subordinate function of each test point
Judge whether the current stop calculation condition is met
N
Y
Output the optimal base station planning scheme

FIGURE 5. Base station deployment model based on weighted K-centroids clustering algorithm.



(a) lr=0.1,0.01,0.05,0.005

(b) lr=0.001,0.002,0.003,0.004

FIGURE 6. Prediction error statistics of test set under different learning rate parameters of prediction model.

function $f\left(b_j \left| p_i\right.\right)$, as shown in equation (13).

$$f\left(b_j \left| p_i\right.\right) = \frac{\left\| p_i - b_j \right\|^2}{\sum_{j=1}^{k} \left\| p_i - b_j \right\|^2} \tag{13}$$

K-centroids clustering algorithm will divide data point p_i into the nearest base station location b_j according to the calculated $f\left(b_j \left| p_i\right.\right)$ value. After assigning the data points, the algorithm also needs to iterate the location of the cluster centre. The reference indicators of the iteration are mainly the distance influence degree and the coverage weight. From the perspective of distance indicators, the base station should cover terminals farther away as much as possible, and terminals closer to the base station will usually be better. On the contrary, terminals far away from the base station

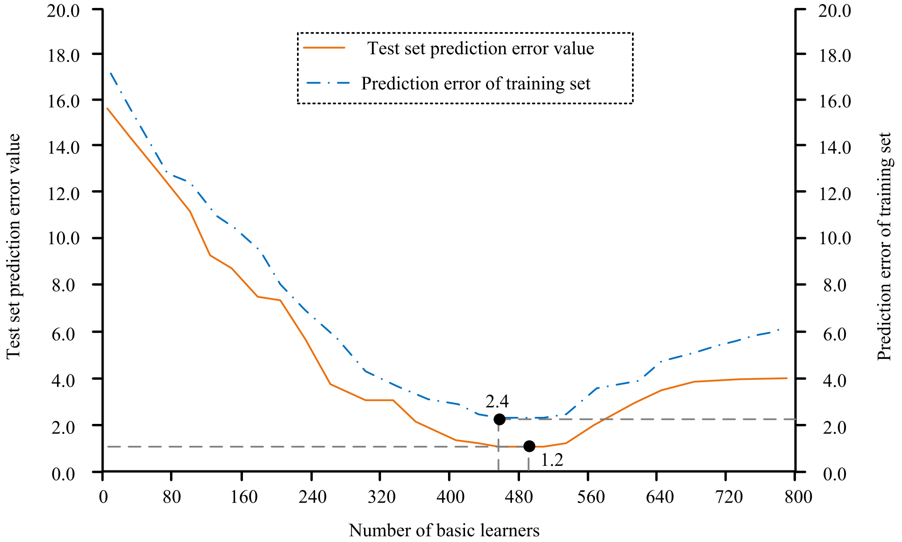


FIGURE 7. Training results of prediction model under different basic learners.

TABLE 2. Setting conditions involved in the experiment.

Number	Full name	referred to as	Set value	Explain
1	Partition ratio between test set and training set	/	3:7	/
2	Convolutional neural network learning rate	lr	0.003	Used to control the learning speed and effectiveness of CNN
3	Number of basic learners	n_{base}	504	The number of basic decision trees that make up the integrated regression tree algorithm
4	Total layers of CNN	k	4	/
5	Number of neurons in the first hidden layer	n_{cell1}	256	The number of neurons in the second layer of the CNN algorithm
6	Number of neurons in the second hidden layer	n_{cell2}	512	The number of neurons in the third layer of the CNN algorithm

may suffer from signal decay due to building blocking signal transmission and other reasons. In other words, the impact on communication increases with the distance between the base station and the terminal. As a result, the subordinate function is used in this study to describe the effect of distance. For the coverage weight, the purpose of optimizing base station's location is to hope that the terminals in the base station can receive good signals. Therefore, the algorithm needs to focus on the terminals with low coverage quality and greater impact weight is given. According to the above contents, each data point will generate the corresponding weight value $w(p_i)$, and the band formula of the cluster center position can be obtained by combining $f\left(b_j\,|p_i\right)$ and $w(p_i)$, see formula (14).

$$b_j = \frac{\sum_{i=1}^{n} f\left(b_j\,|p_i\right) w\left(p_i\right) p_i}{\sum_{i=1}^{n} f\left(b_j\,|p_i\right) w\left(p_i\right)} \tag{14}$$

The total objective function of K-centroids clustering algorithm is calculated according to formula (15).

$$\min \sum_{i\in\{i|y_i\leq\bar{y}\}} (y_i-\bar{y})^2 \tag{15}$$

y_i and $\bar{y}$ are the calculated RSSI value of a data point and the coverage threshold required to theoretically meet the coverage requirements in formula (15). $i \in \{i\,|y_i \leq \bar{y}\}$ represents the point where the communication signal quality in the region is lower than the threshold. The objective function of the algorithm use the sum of least squares error to halt iteration while planning. As of now, the weighted K-centroids clustering method-based base station deployment model has been created. The calculation process is shown in Figure 5. From this Fig.5, the calculation steps of the algorithm are as follows: the first step of the algorithm is to calculate the membership function values of all test points based on input data; Step 2 is to partition the current address data to the nearest cluster center; The third step is to determine whether all test points have been partitioned. If not, to traverse the next address data and return to the first step. Otherwise, the positions of all current base stations are updated. The next step is to recalculate the dependent functions of each test point and determine whether the current stop calculation condition is met. If the condition is not met, return to step 2. Otherwise, output the best base station address planning solution.

TABLE 3. Comparison of computing efficiency of each base station deployment algorithm.

Number of test points to be calculated	Average parsing time/ms					Maximum parsing time/ms				
	WK-centroids	K-means	GMC	Mean Shift	Spectral Clustering	WK-centroids	K-means	GMC	Mean Shift	Spectral Clustering
10	15	51	14	56	62	20	59	17	64	78
100	130	311	104	324	348	169	396	283	412	541
1000	1214	2748	965	2895	2946	1887	3348	2215	3412	3756
10000	12450	26522	9551	26941	27451	17620	30524	27420	35886	29547
47535	42596	112566	40620	113251	125668	85562	165652	152109	156430	178089

IV. FUNCTION TEST OF BIG DATA BASE STATION TRAFFIC PROCESSING MODEL

A. EXPERIMENTAL SCHEME AND PARAMETER DESIGN OF BASE STATION COMMUNICATION QUALITY PREDICTION MODEL

The experiment verifies the application effect of the big data base station communication quality prediction model and the base station planning and deployment model based on multiple machine learning algorithms designed in this study. The actual RPMA base station communication data of a certain region in China is selected for the performance test. The data set includes 38 RPMA base stations and 158,450 test point data. The test set and training set are divided according to the ratio of 3:7. In the research, MATLAB simulation calculation package is used to realize the designed algorithm model. Then, the super-parameters of the large data base station communication quality prediction model are debugged. The key super-parameters include the algorithm learning rate and the number of basic learners. When selecting different learning rate parameters, the prediction error value of the prediction model on the test set is calculated, as shown in Figure 6. Due to the large number of times the learning rate is debugged, the combination plot is used to display all the debugging schemes. The horizontal axis in Figure 6 represents the number of iterations in training, and the vertical axis represents the prediction error value of the prediction algorithm in the test set during training. Different colours and linear curves represent different learning rate schemes. From Figure 6, as the iterations increases, the prediction error value of the prediction model under each scenario on the test set shows a trend of overall fluctuation and decline. However, from the standpoint of learning rate, the prediction error of the test set of the algorithm in the later stages of training falls lower and lower when the learning rate starts to decline from 0.1. When the learning rate is 0.003, the prediction error of the test set after the convergence of the algorithm reaches the minimum value. Specifically, when the number of iterations is 100, the prediction error of the prediction model with a learning rate of 0.003 in the test set is 2.69, which is lower than all other schemes. Because the prediction error is greater than the former when the learning rate continues to drop. Therefore, the learning rate parameter of the prediction model should be 0.003.

Different numbers of basic learners are selected for training. The statistical results are shown in Figure 7. The horizontal axis in Figure 7 shows the number of basic learners in different prediction models. The left vertical axis and the right vertical axis represent the prediction error values on two sentences after training each prediction model. From Figure 7, as basic learners number of the integrated prediction model increases, the prediction error of the model in the test set and training set shows a trend of decreasing first and then increasing. When the number of basic learners is 504, the prediction error value of the model in the test set is the smallest, which is 1.2. But when the number of basic learners is 453, the prediction error value of the model in the training set is the smallest, which is 2.4. The performance of the model depends on the calculation results of the test set, so it is most appropriate to set the number of basic learners of the prediction model to 504.

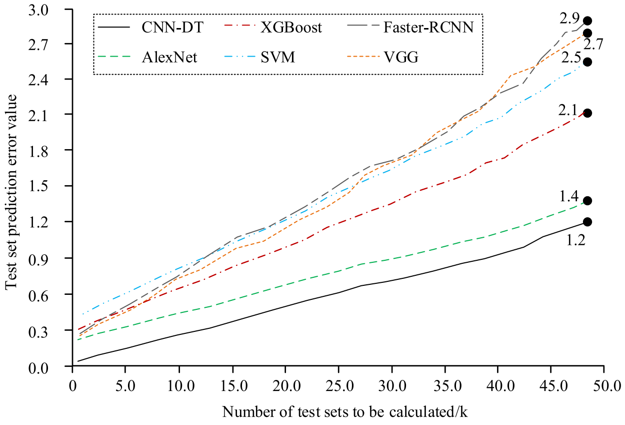


FIGURE 8. Comparison of prediction errors of each prediction model on the test set.

According to the analysis results in Figures 6 and 7, the learning rate parameter of the prediction model designed in this study is set to 0.003, and the number of basic learners is set to 504. Based on industry experience, the parameters of the CNN module are set to a random initialization method with initial values, consisting of four layers: input layer, first hidden layer, second hidden layer and output layer. The number of neurons in the first and second hidden layers is 256 and 512 respectively, and the output layer outputs the signal quality prediction value according to the input data size. In addition, due to the complexity of the research problem and for the purpose of simplifying the experimental conditions, the following

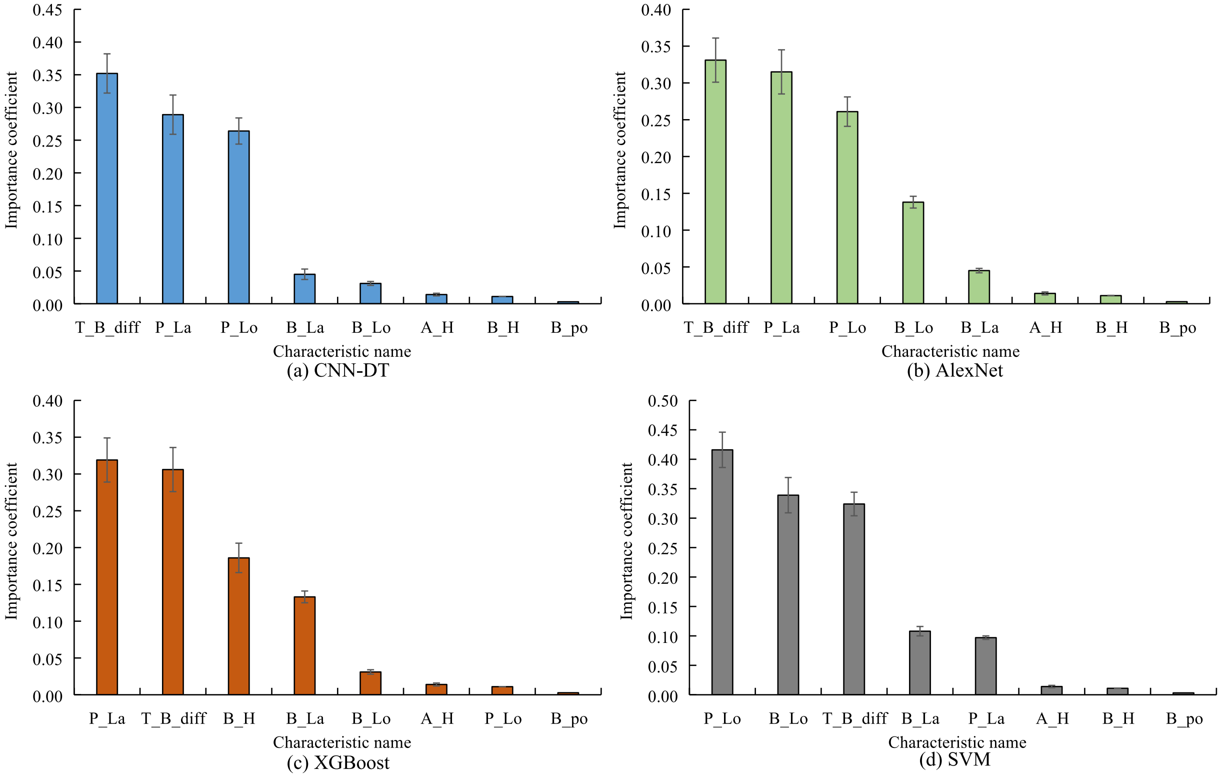


FIGURE 9. Ranking of importance coefficients of input data characteristics of CNN-DT model.

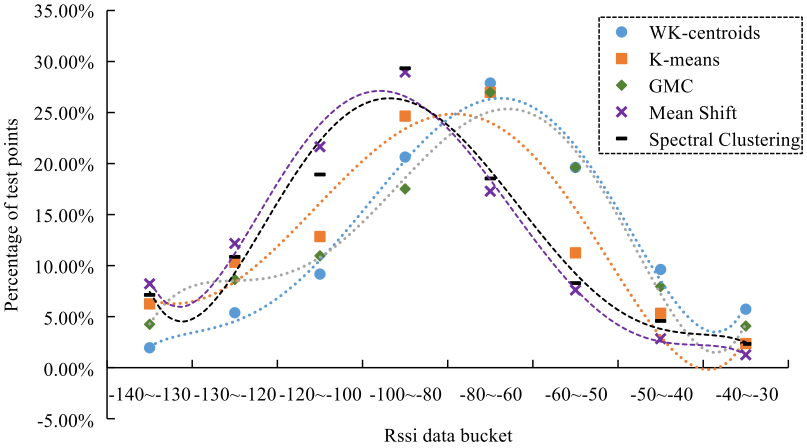


FIGURE 10. Test point RSSI barrel splitting results of each model.

reasonable assumptions are proposed. The first assumption is that the actual geographical area corresponding to the data can install signal base stations; The second signal data is collected under normal operation of the base station; During the collection period of the third signal data, there was no significant difference in the distribution of communication terminal types among each base station. Summarize the setting conditions involved in the experiment and obtain Table 2.

B. PERFORMANCE TEST OF BASE STATION COMMUNICATION QUALITY PREDICTION MODEL

AlexNet, VGG, and Faster RCNN neural networks, XGBoost, and SVM algorithm are selected to construct a comparative prediction model, and the prediction error of each model on the test set after training is analyzed. In Figure 8, the horizontal axis represents the number of test set samples to be predicted, and different line styles represent different prediction algorithms. "CNN-DT" is the prediction model designed for this study that combines CNN and grassroots decision tree algorithm. Figure 8 shows that the total prediction error values of each prediction model show a monotonic linear growth trend as the number of samples in the test set to be forecasted rises. However, in general, prediction models based on XGBoost and SVM algorithms have a larger growth slope. Moreover, the prediction error value of the communication quality data of the test point of the prediction model designed in this study is always lower than that of all comparison models. For example, when the testset consists of all test sets, the total prediction errors of CNN-DT, AlexNet, VGG, Faster RCNN, XGBoost, and SVM algorithms are 1.2, 1.4, 2.7, 2.9, 2.1, and 2.5, respectively.

The communication quality prediction model designed in this study has the highest prediction quality. Due to the worst prediction model quality constructed by VGG and Faster RCNN, these two models will not be displayed in subsequent comparisons. Figure 9 illustrates the analysis of the significance of the model's input attributes. In Figure 9, the horizontal axis represents the input feature name. "T_B_diff", "P_La" and "P_Lo" are the distance between the test point and the base station, and the longitude and latitude of the test point. The vertical axis represents the importance coefficient value of each feature, and different subgraphs represent different prediction algorithms. From Figure 9, only AlexNet and CNN-DT model have the most similar results of feature importance. In the latter, the importance coefficients of T_B_diff, P_La and P_Lo are the largest and significantly higher than those of other characteristics, 0.352, 0.289 and 0.264 respectively.

C. PERFORMANCE TEST OF BASE STATION DEPLOYMENT MODEL

The common classical K-means, Gaussian mixture clustering (GMC), Mean Shift clustering, and spectral clustering algorithm are selected to build a comparative deployment model. RSSI is used as an indicator to evaluate the effectiveness of the model deployment. To calculate the RSSI bucket splitting results of each model's test points, as shown in Figure 10. The two horizontal axes in Figure 10 represent the RSSI indicator data buckets for each test point, while the vertical axis represents the percentage number of test points for each data bucket. Icons with different colors represent different deployment algorithms, and dashed lines with different colors represent the fitting curve of RSSI ratio data for different deployment algorithms. "In this study, a base station deployment model was developed based on the weighted K-centroids algorithm. Analysing Figure 10, the number of test points for each deployment model is relatively small when the RSSI bucket value is small or large. However, the weighted K-centroids clustering algorithm designed in this study has the best overall RSSI value distribution. The smaller the RSSI value of the test point, the worse the signal coverage. For the RSSI bucket "- 140∼- 130", the number of test points of WK centroids model, K-means, GMC, Mean Shift, and spectral clustering model accounted for 1.95%, 6.25%, 4.25%, 8.22%, and 7.13% respectively.

Finally, the computational efficiency of each base station deployment algorithm is analyzed and Table 3 is obtained. The WK-centroids model developed in this study has average and maximum deployment times that fall between those of the K-means model and the GMC, and the relationship between the quantity of test points to be deployed and deployment time is not strictly linear. The more test points to be deployed, the smaller the single point deployment time of each model. The spectral clustering deployment model has the lowest computing efficiency and poor stability, because the maximum computing time has the largest increase compared with the average deployment time under the same conditions. When the test points to be deployed are all test sets (i.e. 47535), the average and maximum deployment time of WK centroids model, K-means, GMC, Mean Shift and spectral clustering model are 42596ms, 112566ms, 40620ms, 113251ms, 125668ms and 85562ms, 165652ms, 152109ms, 156430ms and 178089ms respectively.

V. CONCLUSION

There are some issues with RPMAWCN's heterogeneous network architecture, including low signal quality in some places and a challenging planning process for base stations. This study created a big data base station communication quality prediction model by fusing the lifting regression tree method with CNN to address these issues. And a weighted K-centroids algorithm-based base station planning and deployment model is created. The prediction error value of communication quality data at test points of the prediction model designed in this study is always lower than that of all the comparison models. When the test set consists of all test sets, the total prediction errors of CNN-DT, AlexNet, VGG, Faster RCNN, XGBoost, and SVM algorithms are 1.2, 1.4, 2.7, 2.9, 2.1, and 2.5, respectively. Only the AlexNet model and CNN-DT model have similar feature importance results. While in the latter, T_ B_ diff, P_ La, P_ The importance coefficients of Lo are the highest and significantly higher than other features, with values of 0.352, 0.289, and 0.264, respectively. The weighted K-centroids clustering algorithm developed in this study has the best overall numerical distribution of RSSI. For the RSSI bucket "- 140∼- 130", the number of test points of WK centroids model, K-means, GMC, mean shift and spectral clustering model accounted for 1.95%, 6.25%, 4.25%, 8.22% and 7.13%, respectively. The computational efficiency of the base station planning model based on the WK centroids algorithm is between K-means and GMC. When the gesture images to be deployed are all test sets, the average and maximum deployment times of these five models in sequence are 42596ms, 112566ms, 40620ms, 113251ms, 125668ms and 85562ms, 165652ms, 152109ms, 156430ms and 178089ms respectively. The biggest contribution and contribution of this research to the wireless communication industry lies in its provision of a new automated work mode, enabling base station planning and site selection personnel in the industry to break away from the heavy frontline site selection work and focus on more valuable and creative aspects, improving the efficiency of human capital utilization.

REFERENCES

[1] D. L. Msongaleli and K. Kucuk, "Optimal resource utilisation algorithm for visible light communication-based vehicular ad-hoc networks," *IET Intell. Transp. Syst.*, vol. 14, no. 2, pp. 65–72, Feb. 2020.

[2] Y. Lee, T. Ahn, C. Lee, and S. Kim, "A Novel path planning algorithm for truck platooning using V2V communication," *Sensors*, vol. 20, no. 24, pp. 7022–7047, 2020.

[3] D. K. Anguraj, K. Thirugnanasambandam, R. S. Raghav, S. V. Sudha, and D. Saravanan, "Enriched cluster head selection using augmented bifold cuckoo search algorithm for edge-based Internet of Medical Things," *Int. J. Commun. Syst.*, vol. 34, no. 9, Jun. 2021, Art. no. e4817.

[4] S. Pattnaik and P. K. Sahu, "Assimilation of fuzzy clustering approach and EHO-greedy algorithm for efficient routing in WSN," *Int. J. Commun. Syst.*, vol. 33, no. 1, pp. 47–68, 2020.

[5] C. Liu, T. Wu, Z. Li, and B. Wang, "Individual traffic prediction in cellular networks based on tensor completion," *Int. J. Commun. Syst.*, vol. 34, no. 16, Nov. 2021, Art. no. e4952.

[6] T. G. Tranter, R. Timms, and P. R. Shearing, "Communication-prediction of thermal issues for larger format 4680 cylindrical cells and their mitigation with enhanced current collection," *J. Electrochem. Soc.*, vol. 167, no. 16, pp. 7–12, 2020.

[7] F. Yin, P. Xiao, and Z. Li, "ASC performance prediction for medical IoT communication networks," *J. Healthcare Eng.*, vol. 2021, pp. 1–7, May 2021.

[8] Q. Chen, W. Wang, F. R. Yu, M. Tao, and Z. Zhang, "Content caching oriented popularity prediction: A weighted clustering approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 623–636, Jan. 2021.

[9] H. Gao, F. Han, B. Jiang, H. Dong, and G. Li, "Recursive filtering for time-varying systems under duty cycle scheduling based on collaborative prediction," *J. Franklin Inst.*, vol. 357, no. 17, pp. 13189–13204, Nov. 2020.

[10] T. Wang, X. Zhao, C. Lv, J. Wang, Y. Song, X. Yu, C. Zhou, and N. An, "Blimp-borne laser communication technology based on space dynamic base station," *IEEE Photon. J.*, vol. 13, no. 5, pp. 1–7, Oct. 2021.

[11] R. Sharma, N. Mittal, and B. S. Sohi, "Flower pollination algorithm-based energy-efficient stable clustering approach for WSNs," *Int. J. Commun. Syst.*, vol. 33, no. 7, May 2020, Art. no. e4337.

[12] O. Abbasi, H. Yanikomeroglu, A. Ebrahimi, and N. M. Yamchi, "Trajectory design and power allocation for drone-assisted NR-V2X network with dynamic NOMA/OMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7153–7168, Nov. 2020.

[13] G. E. Figueras-Benítez and R. E. Badra, "Genetic algorithm for biobjective optimization of indoor LTE femtocell deployment," *Int. J. Commun. Syst.*, vol. 33, Aug. 2020, Art. no. e4564.

[14] Y. Yao, Y. Song, H. Ge, Y. Huang, and D. Zhang, "A communication-aware and predictive list scheduling algorithm for network-on-chip based heterogeneous muti-processor system-on-chip," *Microelectron. J.*, vol. 121, Mar. 2022, Art. no. 105367.

[15] Y. Guo and H. Tang, "An improved consensus algorithm for MAS with directed topology and binary-valued communication," *Neurocomputing*, vol. 468, pp. 407–415, Jan. 2022.

[16] B. Liu and Z. Ding, "A consensus-based decentralized training algorithm for deep neural networks with communication compression," *Neurocomputing*, vol. 440, pp. 287–296, Jun. 2021.

[17] Y. Andoh, S. Ichikawa, T. Sakashita, N. Yoshii, and S. Okazaki, "Algorithm to minimize MPI communications in the parallelized fast multipole method combined with molecular dynamics calculations," *J. Comput. Chem.*, vol. 42, no. 15, pp. 1073–1087, Jun. 2021.

[18] G. Chen, L. Wang, M. Alam, and M. Elhoseny, "Intelligent group prediction algorithm of GPS trajectory based on vehicle communication," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 3987–3996, Jul. 2021.

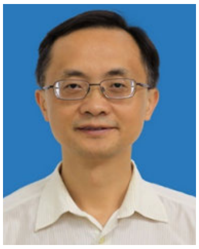
XIEFEI HE was born in Tangshan, Hebei, China, in 1981. She received the B.S. degree from the Hunan University of Science and Technology, in 2005, and the M.Eng. degree in software engineering from the Huazhong University of Science and Technology, in 2012.

Since 2005, she has been a Teacher with the Information Engineering College, Guangzhou Institute of Technology, where she was an Associate Professor of computer software, in 2019. She has authored one book, more than 20 articles, 25 software copyrights, and two utility model invention patents. Her current research interests include big data analysis, database technology and software engineering, virtual simulation, and artificial intelligence.

TAO YU was born in Shaoyang, Hunan, China, in 1978. He received the B.S. and M.S. degrees in industrial automation from Central South University, in 2000, and the M.S. degree from the Guangdong University of Technology, in 2003.

From 2003 to 2009, he was an engineer of embedded systems in a mobile phone company. Since 2010, he has been a Lecturer with the Information Engineering Department, Guangdong Polytechnic of Science and Trade. He is the author of three books and more than 16 articles. His current research interests include embedded system application to industrial internet, data acquisition from intelligent sensors, and wearable network design based on Bluetooth and Zigbee technology.

YANG SHEN was born in Hefei, Anhui, China, in 1978. He received the Ph.D. degrees in software engineering from the South China University of Technology, in 2019. He received the title of Information System Project Manager (Senior), in 2012. Since 2020, he has been an Assistant Professor with the Information Engineering College, Guangzhou Institute of Technology. He is the author of more than 20 articles. His current research interests include big data analysis and software engineering.



SHI'AN WANG was born in Anshun, Guizhou, China, in 1971. He received the B.S. degree in applied mathematics from Guizhou University, Guizhou, in 1993, and the M.S. degree in fluid mechanics from the Nanjing University of Aeronautics and Astronautics, Nanjing, in 2000.

From 1993 to 1997, he was a Software Designer with the Guizhou Guihang Aircraft Design and Research Institute, where he was the Software Project Director, from 2000 to 2005. From 2005 to 2011, he was the Director of the Software Teaching and Research Section, Computer Department, Guangdong Songshan Vocational and Technical College. Since 2012, he has been a Professor with the Information Engineering College, Guangzhou Institute of Technology. He is the author of one book, more than 20 articles, and more than ten inventions. His current research interests include computer application technology, software technology, virtual reality technology, augmented reality technology, and mixed reality technology. He was the Vice Chairman of the Education Simulation Professional Committee of the China Educational Technology Association and the Director of the Expert Committee of the Guangdong MR Education Technology Innovation Alliance.