

UWB Wireless Positioning Method Based on LightGBM

Xuerong Cui^{1*}, Yuanxu Li², Juan Li², Bin Jiang¹, Shibao Li¹
and Jianhang Liu²

^{1*}College of Oceanography and Space Informatics, China
University of Petroleum (East China), Qingdao, 266580, China.

²College of Computer Science and Technology, China University
of Petroleum (East China), Qingdao, 266580, China.

*Corresponding author(s). E-mail(s): cuixuerong@163.com;

Contributing authors: 2205817578@qq.com;

lijuanlijuan@sina.com; jiangbin@upc.edu.cn;

lishibao@upc.edu.cn; liujianhang@upc.edu.cn;

Abstract

In the ultra-wideband indoor positioning scenario, the non-line of sight (NLOS) propagation may be caused by obstacles, which may lead to the deviation of ranging value and affect the positioning precision. Therefore, we propose a NLOS identification and error regression positioning algorithm based on light gradient boosting machine (LightGBM). Firstly, ReliefF algorithm combined with Spearman correlation coefficient is used to analyze the feature correlation, and eight channel features such as total channel impulse response power and standard deviation of noise are selected as NLOS identification features. Then, we adopt genetic algorithm to optimize the hyperparameters of LightGBM for NLOS identification. On this basis, the proposed error regression model based on convolutional neural network (CNN) combined with LightGBM is used to correct the ranging results, so as to achieve high-precision positioning. Through the verification on the public dataset, the NLOS identification accuracy reached 91.8%, and the positioning precision is improved by 45cm after correcting the ranging results.

Keywords: indoor positioning, UWB, NLOS identification, LightGBM, CNN, error mitigation

1 Introduction

With the rapid development of the Internet of Things and wireless communication technology, people's demand for location-based services (LBS) is constantly increasing, especially in positioning. Outdoors, the Global Navigation Satellite System (GNSS) provides high-precision positioning, which can meet most positioning requirements. However, the GNSS signal is too weak to penetrate walls, and the attenuation is severe indoors, so it is difficult to achieve a good positioning performance. In order to solve the indoor positioning problem in complex environment, the existing positioning technologies based on radio frequency identification (RFID), bluetooth (BT), wireless fidelity (WiFi) and other positioning technologies have achieved positioning precision at only meter level, which cannot meet people's demand for high-precision positioning services.

Compared with the above indoor wireless positioning technology, ultra-wideband (UWB) positioning technology has the characteristics of high time resolution and strong anti-multipath ability, and it can achieve higher positioning precision by using the method based on time of arrival (TOA) and time difference of arrival (TDOA)[1]. However, both TOA and TDOA algorithms need accurate time measurement values to calculate the distance values. Due to the complex and variable indoor environment, UWB signals may be blocked by obstacles in the transmission process, resulting in non line of sight (NLOS) errors. Compared with the line of sight (LOS) environment, it will produce longer propagation path and additional delay, and the positioning precision will be significantly reduced. How to identify NLOS signal and correct ranging error is a subject worthy of study.

Currently, NLOS identification is classified according to characteristic parameters, mainly based on ranging information and channel impulse response (CIR). Reference[2] proposes a method to identify different channels by statistical analysis of the ranging variance of the received signal, but this method needs more prior knowledge and is not easy to make subsequent error correction. Reference[3] combines the peak time with the rising time as a new channel feature, and makes a judgment based on the channel feature combined with the undetected peak number. This algorithm has a high recognition rate in specific scenarios and the threshold is properly selected. However, if the threshold selection is not appropriate or the peak time and rising time in some scenarios are significantly different from expectations, error is often introduced in recognition result. Reference[4] proposes a NLOS identification method which is independent of prior knowledge and little environmental dependence, but it is not universal and can only identify specific NLOS channels.

References[5, 6] put forward a NLOS channel identification method based on deep learning, which directly establishes the correspondence between CIR and whether it is NLOS channel, and omits the manual feature extraction process. However, this method has a large amount of computation and a long training time. Reference[7] adopts approximate maximum likelihood algorithm to directly correct NLOS error, but this method is only effective when the

error impact is small and there are many base stations. Reference[8] uses the combination of inertial sensor and UWB sensor to estimate the global attitude by using the complementary characteristics between sensors. However, this method relies on the second sensor, and the inertial sensor itself has accumulated errors, so the precision of error correction is limited. Especially with the increase of time, the cumulative errors will increase.

The above methods have the problems of low recognition rate or high recognition rate only in a certain environment, poor environmental applicability, large amount of computation, limited error correction precision and low positioning precision. In view of those deficiencies of the existing algorithm, this paper proposes a method based on ReliefF algorithm and Spearman correlation coefficient to select the characteristics of UWB channel quality indicators. For the selected channel quality index, a NLOS recognition algorithm based on genetic algorithm (GA) to optimize light gradient boosting machine (LightGBM) and a NLOS error regression model based on CNN combined with LightGBM is proposed. The accuracy of NLOS recognition is improved, and the ranging error under NLOS channel is corrected to improve the positioning precision.

In the second section, the system model is introduced, including correlation analysis, NLOS identification algorithm based on GA optimized LightGBM, and NLOS error regression model based on CNN combined with LightGBM. In the third section, the above models are experimentally analyzed to verify the performance advantages of the models proposed in this paper. Finally, the research content of the whole paper is summarized in the fourth section.

2 Methods

2.1 Correlation Analysis

In order to reduce the complexity of NLOS recognition and error regression model, the feature selection algorithm based on ReliefF algorithm[9] combined with Spearman correlation coefficient is used to extract feature subsets from the original feature set, remove redundant features and irrelevant features, reduce feature dimensions, and only features that are strongly correlated with predictors are retained.

The ReliefF algorithm is a feature weight algorithm dealing with multi-classification problems. The larger the feature weight, the stronger its ability to distinguish NLOS. The specific operation steps of the algorithm are shown in Algorithm 1:

In the weight update formula 1, $p(C)$ is the proportion of class C to the total samples in the original dataset, and $p(class(R))$ is the proportion of the category of randomly selected samples to the total samples. $\text{diff}(A, R_1, R_2)$ represents the distance between the sample R_1 and R_2 on the feature A . The specific calculation formula is as follows:

Algorithm 1 ReliefF algorithm

Require: training set S , sampling times m ,
 number of nearest neighbor samples k , feature weight W ,
 feature dimension n , feature weight threshold λ ;

Ensure: Set E with weight greater than threshold λ in all features;

- 1: Set E as an empty set, and set 0 to all feature weights $W=0$;
- 2: **for** $i = 1$ to m **do**
- 3: A sample set R is randomly selected from the training sample set S ,
 $R \in S$;
- 4: Find k nearest neighbor samples from the sample set of the same kind
 of R ;
- 5: Find k nearest neighbors from each distinct set of samples of R ;
- 6: **end for**
- 7: **for** $A = 1$ to n **do**
- 8: update feature weight:

$$W(A) = W(A) - \sum_{j=1}^k \frac{\text{diff}(A, R, H_j)}{mk} + \sum_{C \notin \text{class}(R)} \frac{\frac{p(C)}{1-p(\text{class}(R))} \sum_{j=1}^k \text{diff}(A, R, M_j(C))}{mk} \quad (1)$$

- 9: **end for**
- 10: **for** $i = 1$ to m **do**
- 11: **if** $W(A) > \lambda$ **then**
- 12: $E = E \cup A$;
- 13: **end if**
- 14: **end for**
- 15: return result E ;

$$\text{diff}(A, R_1, R_2) = \begin{cases} (|R_1[A] - R_2[A]|) / (\max(A) - \min(A)), & \text{If } A \text{ is continues} \\ 0, & \text{If } A \text{ is discrete and } R_1[A] = R_2[A] \\ 1, & \text{If } A \text{ is discrete and } R_1[A] \neq R_2[A] \end{cases} \quad (2)$$

Spearman correlation coefficient is used to measure the correlation between data features and predicted variable. By calculating the correlation between data sample features and model predicted values, it can reflect the importance of data features to predicted variables to a certain extent. The value range of Spearman correlation coefficient is $[-1, 1]$, where 1 represents perfect positive correlation and -1 represents complete negative correlation. The larger Spearman correlation coefficient is, the more important this feature is to predict variables.

2.2 NLOS Recognition Algorithm Based on GA Optimized LightGBM

In the NLOS recognition of UWB signals, the training process of traditional machine learning models such as support vector machine (SVM) and convolutional neural network (CNN) takes a long time, especially in the case of large data volumes, which can't guarantee real-time performance. This paper proposes to apply the LightGBM algorithm with fast model training speed and high accuracy to NLOS recognition of UWB signals.

2.2.1 LightGBM Algorithm

The LightGBM algorithm[10] is an integrated algorithm based on the gradient boosting framework launched by Microsoft Research Asia in 2017. Compared with the traditional gradient boosting decision tree (GBDT) algorithm, it has the characteristics of high performance dealing with large-scale data, higher accuracy, and faster model training. The basic principle of the algorithm is: for a set of sample data $X = \{(x_i, y_i)\}_{i=1}^N$, the goal of LightGBM is to find an approximate function $\hat{f}(x)$ for $f(x)$, which minimizes the expectation of the loss function $L(y, f(x))$.

$$\hat{f} = \arg \min_f E_{y,X} L(y, f(x)) \quad (3)$$

In which T regression trees $T_t(x)(t = 1, \dots, T)$ are used to approximate the final model:

$$f_T(x) = \sum_{t=1}^T T_t(x) \quad (4)$$

Regression tree $T_t(x)(t = 1, \dots, T)$ can be represented by $W_{q(x)}(q \in \{1, \dots, J\})$, J represents the number of leaves, q is the decision rule, and W is the weight of leaf nodes. Therefore, LightGBM can be trained in addition form in step t, and its loss is:

$$\Gamma_t = \sum_{i=1}^n L(y_i, f_{t-1}(x_i) + T_t(x_i)) \quad (5)$$

Compared with the traditional GDBT algorithm, LightGBM replaces the Pre-Sorted algorithm with the histogram algorithm, and constructs a histogram with a width of n while discretizing the continuous eigenvalues into n eigenvalues. When traversing the data, the discretized value is used as the index for cumulative statistics, and then the optimal point is traversed according to the discrete value of the histogram. In addition, LightGBM adopts the strategy of histogram difference acceleration, that is, the histogram of a leaf node can be obtained by the difference between the histogram of its parent node and the histogram of its sibling nodes, thereby accelerating the training process and reducing memory usage. The principle of the histogram algorithm is shown in Figure 1.

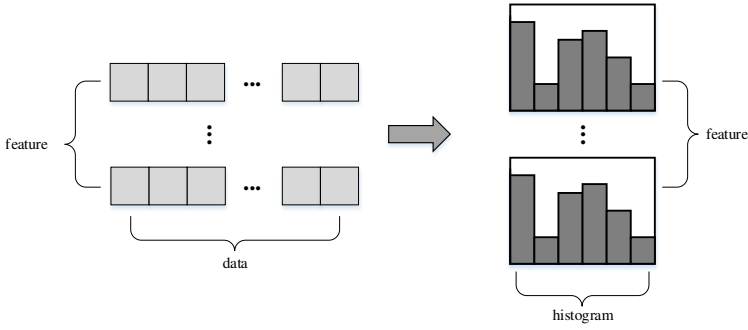


Fig. 1 Histogram diagram

At the same time, LightGBM also adopts a more efficient Leaf-wise growth strategy with depth limitation, which is realized by repeatedly splitting the leaf with the largest splitting gain from all the current leaf nodes each time. Therefore, under the same splitting times, compared with the Level-wise growth strategy adopted by the traditional GBDT algorithm, the Leaf-wise growth strategy has lower error and higher accuracy. Aiming at the shortcoming that the Leaf-wise growth strategy may grow a deeper decision tree, the LightGBM adds a maximum depth limit to the Leaf-wise growth strategy, which prevents overfitting and ensures high efficiency. The leaf growth strategy of Leaf-wise with depth restriction is shown in Figure 2.

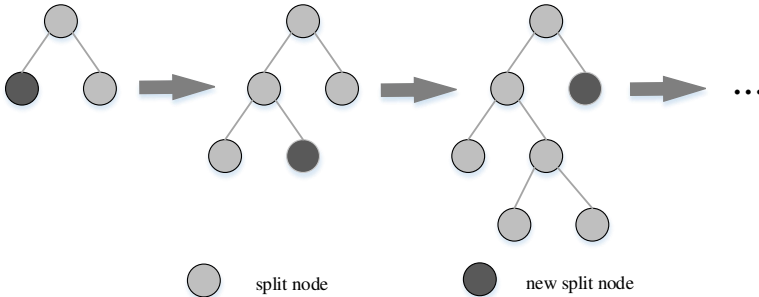


Fig. 2 Leaf-wise growth strategy chart with depth restriction

2.2.2 Genetic Algorithm

Genetic algorithm[11] is an adaptive random search heuristic algorithm. The algorithm searches for the optimal solution of the optimization problem by simulating the biological evolution process of Darwin's theory of biological evolution, and substitutes the processes of genetic replication and cross mutation in the biological evolution into the problem solving process. The algorithm has the advantages of high search efficiency and strong adaptability, and is widely used in various optimization problems. The basic steps of the algorithm are:

- (1) Set the evolution iteration counter $g = 0$, set the maximum evolutionary generation G , and randomly generate the initial population $P(0)$.
- (2) Calculate the fitness F_t of individuals in group $P(t)$.
- (3) According to the roulette algorithm, select excellent individuals and inherit them to the next generation.
- (4) Carry out mutation and crossover operations according to gene mutation rate p_v and cross reproduction rate p_c .
- (5) Generate a new individual $P(t+1)$, calculate its fitness value F_{t+1} , and sort according to the fitness value.
- (6) If $g \leq G$, then $g = g + 1$, go to step (2); if $g > G$, output the individual with the best fitness F in population P as the optimal solution, and terminate the calculation.

2.2.3 GA Optimized LightGBM Algorithm

In machine learning, different hyperparameter combinations of the model will directly affect the training speed and recognition accuracy of the model. Since LightGBM has many hyperparameters, in order to search for a combination of hyperparameters that make the model perform better, a optimization method is used.

Commonly used hyperparameter optimization methods include manual search method, grid search method and random search method. Manual search is a trial-and-error process by manually changing random hyperparameters set, which takes a long time and cannot guarantee finding the best combination of hyperparameters. The grid search method is a simple and violent exhaustive search method. When the number of hyperparameters is too large and the traversal granularity is not small enough, the optimal hyperparameter group may not be obtained. The random search method ignores forward information, resulting in extremely low search efficiency. In contrast, the genetic algorithm adopted in this paper is more intelligent. The fitness function gives the search process a clear goal, so that the search direction always develops towards the optimal result, and the probability of crossover and mutation brings randomness to the search process. , which is more conducive to seeking the global optimal point and obtaining the prediction model with the optimal performance.

This paper mainly optimizes four hyperparameters that have a great impact on model performance: `learning_rate`, `n_estimators`, `max_depth`, `min_child_weight`. Where `learning_rate` represents the learning rate, `n_estimators` represents the number of the tree, `max_depth` represents the depth of the largest tree, and `min_child_weight` represents the minimum leaf weight. The basic steps of the NLOS identification algorithm based on GA optimized LightGBM are shown in Figure 3.

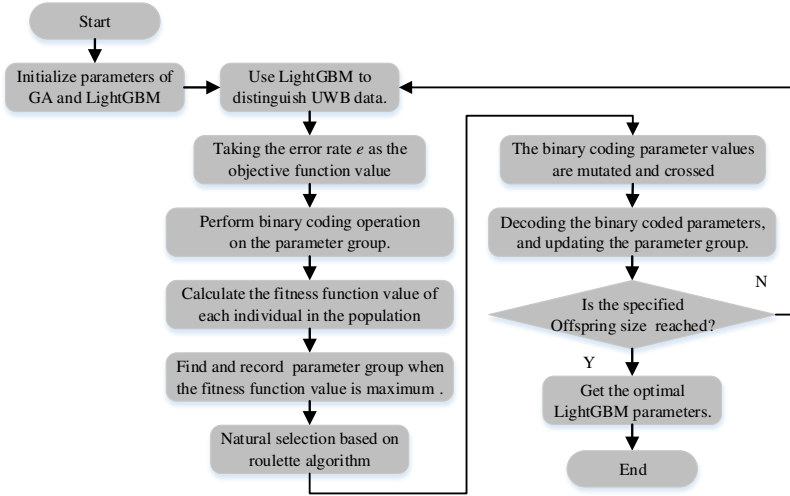


Fig. 3 GA Optimized LightGBM

2.3 NLOS Error Regression Model Based on CNN and LightGBM

2.3.1 CNN Principles

Convolutional Neural Network (CNN)[12] is a commonly used deep learning model. It uses CNN for regression compared with traditional regression methods such as linear regression and logistic regression, the advantage of CNN is that it can automatically extract features and abstract the original data layer by layer into the final features that can represent the output, thus realizing the mapping of high-level features to targets. Moreover, CNN can share weights, which further shortens training time and improves operating efficiency. CNN is a kind of feedforward neural network with convolution computation and depth structure, and its main feature is the convolution operation. CNN was initially mainly used for computer vision problems such as image classification and target detection, but with the deepening of research, CNN has also been gradually applied to speech recognition, natural language processing and other fields; and has achieved better classification and regression effects than traditional methods.

The classic CNN network model is shown in Figure 4. CNN is generally composed of the input layer, convolution layer, pooling layer, full connection layer and output layer. The input layer converts the dataset into the corresponding array form through preprocessing, and inputs it into the neural network. The convolution layer is the core of CNN and is composed of feature maps, each of which contains multiple neurons. The convolution kernel connects neurons with the input of the upper layer to extract local features. At the same time, the convolution layer uses activation function to map the input of

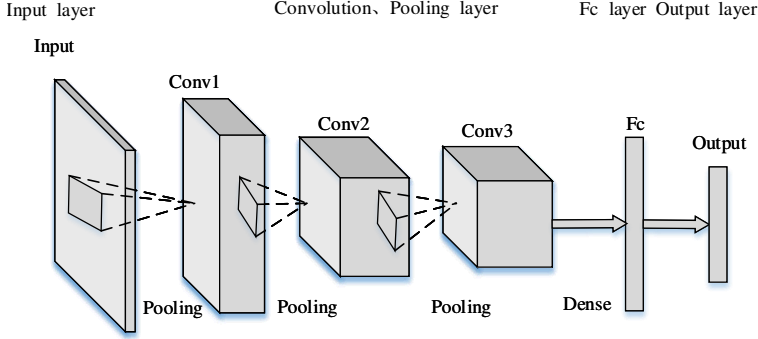


Fig. 4 CNN structure diagram

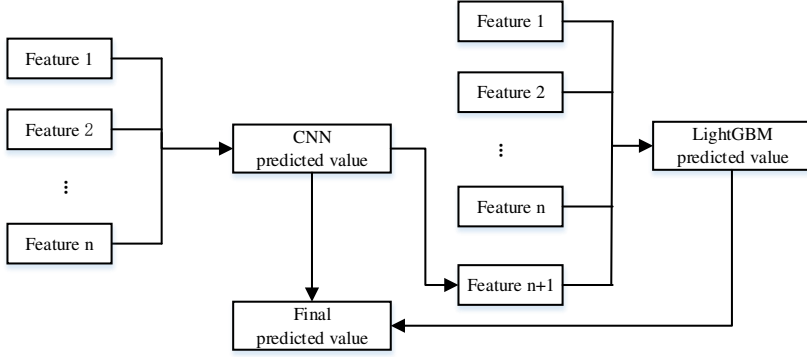
neurons to the output. Pooling layer is also called downsampling layer. Commonly used pooling layers include maximum pooling layer, random pooling layer and average pooling layer. Pooling layer mainly reduces the data dimension by reducing the resolution of feature graph, which can not only reduce the amount of computation, but also extract secondary features and prevent network training from overfitting. The fully connected layer is generally after the convolution layer and the pooling layer. The neurons in the fully connected layer are fully interconnected with the neurons in the previous layer, which can integrate the distinguishing features of each part in the convolution layer and the pooling layer, thus improving the performance of the whole neural network. The output layer outputs the classification result or regression result.

2.3.2 CNN Combined With LightGBM NLOS Error Regression Model

CNN is a neural network model, and LightGBM is a tree model. This paper combines them to obtain a CNN-LightGBM error regression model to improve the accuracy of the NLOS error regression value. The model combines the advantages of the neural network model and the tree model, considering that the tree model has high-level explanation, and using neural network model to make low-level decisions. Figure 5 shows the structure of the CNN-LightGBM model proposed in this paper.

The two models are trained using NLOS error dataset samples respectively. Because CNN can combine low-level local features into high-level features and automatically extract them, an initial predicted value can be obtained. Then, this predicted value is inputted into LightGBM as a new feature, thus obtaining LightGBM predicted value. After the prediction results of CNN and LightGBM are obtained, the prediction values of the above two models are weighted and combined by the error reciprocal method, and the prediction values of CNN-LightGBM error regression model are as follows:

$$y = w_1 y_{CNN} + w_2 y_{LightGBM} \quad (6)$$

**Fig. 5** CNN-LightGBM

y_{CNN} is the predicted value of CNN model, $y_{LightGBM}$ is the predicted value of LightGBM model, w_1 and w_2 are the weight coefficients corresponding to the predicted values of CNN and LightGBM respectively.

$$w_1 = \frac{\mu_2}{\mu_1 + \mu_2} \quad (7)$$

$$w_2 = \frac{\mu_1}{\mu_1 + \mu_2} \quad (8)$$

Among them, μ_1 and μ_2 are the prediction errors of CNN model and LightGBM model respectively. By giving a larger weight to the model with smaller error, it reduces the prediction error of the whole CNN-LightGBM model, thus improving the effect of NLOS error regression.

3 Results

3.1 Data Source

The NLOS identification part and NLOS error regression part of this paper adopts a public dataset, which is funded by EWINE (Elastic Wireless Networking Experiment) in the European Union's "Horizon 2020" program and established by Bregar et al [5]. The public dataset uses DW1000 to collect data, which can return data such as UWB ranging value, channel quality index and channel impulse response. The public dataset is divided into two parts, one is NLOS identification dataset for training NLOS identification model, and the other is NLOS error regression dataset for training NLOS error regression model and evaluating positioning performance.

The NLOS identification dataset is divided into seven small datasets in order to ensure the diversity of the environment, and the experiments are conducted in seven different indoor environments: office 1, office 2, small apartment, small workshop, kitchen, bedroom and boiler room. When collecting data, two UWB positioning base stations are used, one UWB node is used as

a tag and the other UWB node is used as an anchor. In seven different scenarios, 3000 LOS data samples and 3000 NLOS data samples are collected, with a total of 42000 pieces of data. In order to prevent the sequence of data sets from affecting NLOS recognition, the sample data are randomly scrambled. Each piece of data in NLOS identification dataset includes NLOS label, measurement distance, channel quality index, channel impulse response, etc.

The NLOS error regression dataset has measured the distance and actual distance between the tag node and anchor node. The experiment defines a grid with points spaced 1m apart in two office environments. One node is placed in a random position on the grid, and thereafter the second node is placed in a different position on the grid. For each grid point, 100 measurements are made. According to the office environment during measurement, the measurement results are divided into two different groups. One group is used as training data, and the other group is used as test data to evaluate the positioning performance. Each piece of data in NLOS error regression dataset includes the coordinates of tag node and anchor node, NLOS label, measurement distance, channel quality index and channel impulse response, etc.

3.2 NLOS Feature Selection

The feature selection algorithm based on ReliefF combined with Spearman correlation coefficient is to sum the weights calculated by ReliefF algorithm with Spearman correlation coefficient, so as to select features with strong distinguishing ability. In this paper, the first eight features with strong distinguishing ability are adopted, which not only considers the feature dimension, but also ensures the validity of the model. The eight selected features and their meanings are shown in Table 1.

Table 1 Features details

feature name	details
RANGE	measurement distance
FP_AMP1	first path amplitude part 1
FP_AMP2	first path amplitude part 2
FP_AMP3	first path amplitude part 3
STDEV_NOISE	standard deviation of noise
CIR_PWR	total channel impulse response power
MAX_NOISE	maximum value of noise
RXPACC	received RX preamble symbol

After the feature selection algorithm, the violin figure drawn by eight features is obtained, as shown in Figure 6.

The Violin plot is used to display the distribution state and probability density of multiple sets of data. It combines the characteristics of box plot and density plot, and is mainly used to display the distribution shape of data. It can be seen from the figure that the first eight features obtained by feature

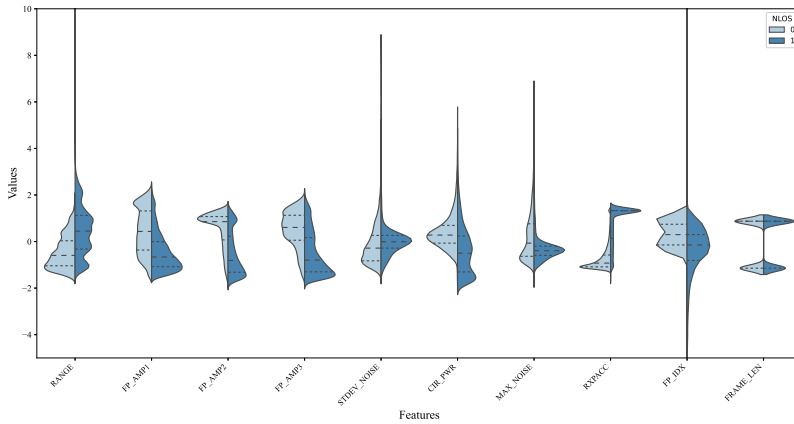


Fig. 6 Feature correlation diagram

selection algorithm have stronger correlation with whether they are NLOS data than the last two features.

3.3 Analysis of NLOS Recognition Results

3.3.1 Model Construction and Evaluation Indicators

Based on the parameter setting of reference[13] and a large number of experiments, the initial parameters of the genetic algorithm are determined in this paper. The specific settings are as follows: the population size is 500, the number of iterations is 100, the crossover rate is 0.6, and the variation rate is 0.01. At the same time, according to common processing method, 70% of the data set is used as training data and 30% as test data.

Using the previously described NLOS recognition algorithm based on GA to optimize LightGBM, the optimal hyperparameter combination of LightGBM is obtained through continuous iterative calculation, as shown in Table 2.

Table 2 Optimal hyperparameter combination of model

name	value
learning_rate	0.0406
n_estimators	1050
max_depth	8
min_child_weight	4

When evaluating the NLOS recognition model of UWB signal, this paper adopts four evaluation indicators that are widely used in classification prediction research to evaluate the performance of different models, including Accuracy, Recall, Precision and F1-measure. All indicators can be defined and

calculated according to True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Specifically, TP indicates that the UWB signal is actually NLOS, and the prediction result is NLOS; TN indicates that UWB signal is actually LOS, and the predicted result is LOS; FP represents that the actual LOS of UWB signal, and the prediction result is NLOS; FN indicates that UWB signal is actually NLOS, and the predicted result is LOS. The definitions and calculation formulas of each evaluation indicator are as follows.

Accuracy rate measures the overall classification precision of the classifier, specifically the ability of the classifier to correctly identify the NLOS and LOS of UWB signals. The calculation formula is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Recall rate, also known as Sensitivity, measures the classifier's ability to correctly identify UWB signals as NLOS. The calculation formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

The precision rate represents the ability of the classifier to correctly predict, in all samples, whose UWB signal prediction results are NLOS. The calculation formula is:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

F1 index evaluation standard is used to measure the comprehensive performance of the classifier, and higher scores are given to models with closer sensitivity and precision. The calculation formula is:

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (12)$$

3.3.2 Model Performance

In this paper, eight features selected by the feature selection algorithm based on ReliefF combined with Spearman correlation coefficient, such as measurement distance, first path amplitude part 1, first path amplitude part 2, first path amplitude part 3, standard deviation of noise, total channel impulse response power, maximum value of noise, and received RX preamble symbol, are used as the input of NLOS recognition model after normalization. The hyperparameters of LightGBM model are obtained by GA optimizing LightGBM, and the recognition effects of SVM[14], CNN[5], GBDT, LightGBM and GA-LightGBM are compared.

As shown in Table 3, this paper compares several NLOS recognition algorithms, in which CNN is used to train channel impulse response in the public dataset paper, and 152 columns of CIR indexed from the first path are

Table 3 Comparison of different NLOS recognition algorithms

algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
SVM	85.1	90.3	78.4	84.1
CNN	87.4	85.9	89.4	87.6
GBDT	89.4	91.3	86.9	89.3
LightGBM	90.6	92.5	88.5	90.5
GA-LightGBM	91.8	93.2	90.2	91.7

taken as input, and the recognition rate is 87.4%. The LightGBM algorithm uses unoptimized hyperparameters. GA-LightGBM algorithm hyperparameters combination is optimized by genetic algorithm for LightGBM. The experimental results show that LightGBM and GA-LightGBM algorithms both have recognition rates reaching over 90%. GA-LightGBM algorithm is based on LightGBM, and the hyperparameters are optimized by genetic algorithm. It is superior to other NLOS recognition algorithms in many indicators, and the evaluation indicators are relatively balanced.

3.4 Analysis of Error Regression Results

3.4.1 Model Construction and Evaluation Indicators

The input of NLOS error regression model is 8 columns of data selected by feature selection algorithm based on ReliefF combined with Spearman correlation coefficient. In order to carry out two-dimensional convolution operations, the NLOS signal data must be expanded first by increasing the dimension.

In this paper, the NLOS error regression model adopts a CNN structure with four convolution layers, and the activation function of the convolution layer is Relu. The Relu function can give neurons sparse activation, and increase the nonlinear relationship among the layers of the neural network, thus making the convergence speed of the CNN model more stable. The model structure and initial parameter settings of CNN are obtained by analyzing a large number of experimental results. Conv1 convolution kernel is 3×3 , with 32 channels, Conv2 convolution kernel is 2×2 , with 64 channels, Conv3 convolution kernel is 2×2 , with 32 channels, Conv4 convolution kernel is 2×2 , with 16 channels. After each convolution layer, there is a BatchNormalization layer (BN), which solves the problem of uneven distribution of feature data in the hidden layer among samples, thus improving the training speed and accuracy. After that, it is the maximum pool layer with a pool kernel size of 2×2 . The maximum pool layer uses downsampling method to reduce the parameter quantity, improve the operation speed and prevent overfitting to some extent. After that comes the full connection layer. The optimizer of the CNN model is Adam, which can automatically adjust the learning rate. The objective function is MSE.

The CNN model separately trains NLOS error dataset samples, and selects the model with better regression effect to save. The prediction result obtained

by the better model is used as a new feature, and together with the initial features are used as the input of LightGBM for training, and the LightGBM model with better regression effect is selected to obtain the prediction value. The prediction values of the above two models are weighted and combined by the error reciprocal method to obtain the prediction value of CNN-LightGBM error regression model.

When evaluating the NLOS error regression model of UWB signal, this paper uses three widely used evaluation indicators to evaluate the performance of different models, including mean absolute error (MAE), mean square error (MSE) and determination coefficient (R^2). Their meanings and calculation formulas are as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (13)$$

MAE is the average value of the absolute values of the difference between the real value and the predicted value of each sample, which reflects the aggregation of the true residuals predicted by the regression model. y_i is the true value of the sample and $\hat{y}_i$ is the predicted value of the regression model.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (14)$$

MSE is the average of the square of the difference between the predicted value and the real value of each sample. The smaller the value, the better the fitting effect of the regression model.

$$R^2 = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2} \quad (15)$$

R^2 is used to measure the explanation degree of each independent variable to the change of dependent variable through numerical quantification of the fitting degree of the regression model. Its value ranges from negative infinity to 1. When R^2 is negative, it indicates that the regression model has a very poor fitting effect. When R^2 is close to 1, it means that the regression model has a better fitting effect. is the average of the sample true values.

3.4.2 Model Performance

This paper adopts the error regression model based on CNN combined with LightGBM to compare the regression effects of LinearRegress, AdaBoost, CNN, LightGBM and CNN-LightGBM.

Table 4 Comparison of different NLOS error regression models

regression model	MAE(m)	MSE(m)	$R^2(\%)$
LinearRegress	0.2402	0.1031	0.4701
AdaBoost	0.1956	0.0616	0.5198
CNN	0.1261	0.0366	0.7582
LightGBM	0.0591	0.0141	0.9197
CNN-LightGBM	0.0465	0.0071	0.9614

As shown in Table 4, this paper compares various NLOS error regression model. CNN-LightGBM is superior to the NLOS error regression model constructed by CNN or LightGBM alone in several evaluation indicators. the MAE is reduced to 0.0465, MSE drops to 0.0071, and R^2 reached to 0.9614.

3.5 Positioning Error Analysis

There are two positioning methods after NLOS recognition of UWB signals: the first method is to discard NLOS data directly and only use the remaining LOS data for location solution; The second is to establish a regression model between NLOS data and ranging error according to the channel quality index, input NLOS data into the NLOS error regression model to obtain the error value, subtract the error value from the ranging value to get the corrected ranging value, and then use the corrected ranging value for positioning solution[15].

For the node position estimation of a distance-based positioning system, without prior knowledge of environmental geometry, several anchor nodes with known positions are required. For two-dimensional positioning, three base stations are needed, while for three-dimensional positioning, there are at least four base stations within the range of one tag node. For the sake of simplicity and operation speed, the Least Squares (LS) and Weight Least Squares (WLS) are selected in this paper. The derivation process of the algorithm is not specifically described here. The WLS algorithm is used in the second positioning method. Because the NLOS error regression model can obtain the regression error, the regression error is used as the error weighting coefficient of the weighting matrix in WLS algorithm to correct the distance information and improve the positioning precision of the algorithm. The localization effect of the algorithm using the public dataset is shown in Figure 7.

As shown in Figure 7, ls_los_nlos is the relationship between the number of base stations and the error obtained by directly using LS algorithm for positioning without NLOS identification of all data. The ls_los_GA-LightGBM uses GA to optimize LightGBM for NLOS identification, and then removes the NLOS data. When the number of remaining LOS data is greater than or equal to 3, which is the minimum number of base stations for two-dimensional positioning, it shows the relationship between the error obtained by the positioning solution and the number of base stations. Compared with the positioning

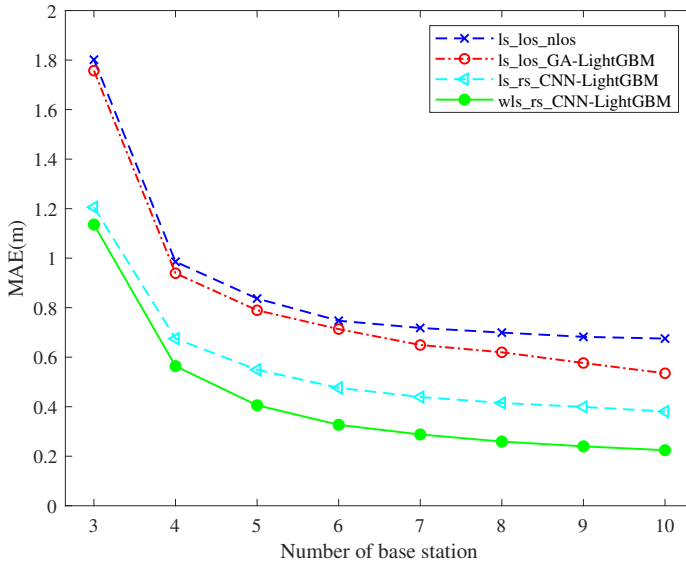


Fig. 7 Relationship between average error and number of base stations

results obtained without any processing, the error is reduced, and the positioning precision is constantly improved with the increase of the number of base stations, which verifies the effectiveness of NLOS identification algorithm based on GA-optimized LightGBM proposed in this paper.

The ls_rs.CNN-LightGBM refers to the relationship between the number of base stations and the error obtained by error regression of NLOS data by CNN combined with LightGBM model, and positioning solution by LS algorithm is performed after correcting the ranging error. Compared with the positioning results obtained without any processing, the error is reduced by about 60cm when the number of base stations is 3, and by about 29cm when the number of base stations is 10, which verifies the effectiveness of NLOS error regression model based on CNN and LightGBM proposed in this paper. The wls_rs.CNN-LightGBM refers to the error regression of NLOS data by CNN combined with LightGBM model, and the relationship between the error and the number of base stations by WLS algorithm after correcting the ranging error. Compared with the positioning results obtained by LS algorithm, the error is reduced by about 7cm when the number of base stations is 3, and, the error is reduced by about 16cm when the number of base stations is 10. Compared with the positioning results obtained without any processing, the error is reduced by about 67cm when the number of base stations is 3, and by about 45cm when the number of base stations is 10, which verifies the effectiveness of WLS algorithm in positioning after NLOS error regression model processing.

After NLOS identification, the NLOS data is directly eliminated and only the remaining LOS data is used for positioning. Compared with the positioning

without any processing, the positioning precision is improved, but there is still a big error. Then the proposed NLOS error regression model corrects the ranging results, giving higher positioning precision and lower error.

4 Conclusions

UWB signals may be affected by obstacles when propagating indoors, resulting in non-line-of-sight propagation and introduces deviations in time measurement values, which in turn affects the precision of distance measurement values. This paper proposes a NLOS identification algorithm based on GA-optimized LightGBM to distinguish whether it is NLOS data through feature selection of UWB channel quality indicators. At the same time, in order to reduce the influence of NLOS error on localization, a NLOS error regression model based on CNN combined with LightGBM is proposed. It is verified in the public dataset that the two algorithms can effectively distinguish the NLOS data and correct the ranging error, so as to obtain higher positioning precision.

Declarations

- **Funding:** This work was supported by the National Natural Science Foundation of China (No. 52171341, 61902431, and 61972417), the science and technology project of Qingdao west coast new area under grant No. 2020-84, and the scientific foundation of Shandong province under grant No. ZR2020MF005.
- **Data Availability Statement:** The data that support the findings of this study are available from the first and corresponding author, upon reasonable request.
- **Conflict Of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- [1] von Tschirschnitz, M., Wagner, M., Pahl, M.-O., Carle, G.: A generalized tdoa/toa model for tof positioning. In: 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–8 (2019). <https://doi.org/10.1109/IPIN.2019.8911742>
- [2] Yu, K., Dutkiewicz, E.: Nlos identification and mitigation for mobile tracking. *IEEE Transactions on Aerospace and Electronic Systems* **49**(3), 1438–1452 (2013). <https://doi.org/10.1109/TAES.2013.6557997>
- [3] Hai, L., Yunlong, X., Yueyuan, L., Xingyue, Y., Wei., Z.: Research on nlos recognition method based on cir characteristic parameters. *China Measurement & Test* **47**(09), 20–25 (2021)

- [4] Yu, K., Wen, K., Li, Y., Zhang, S., Zhang, K.: A novel nlos mitigation algorithm for uwb localization in harsh indoor environments. *IEEE Transactions on Vehicular Technology* **68**(1), 686–699 (2019). <https://doi.org/10.1109/TVT.2018.2883810>
- [5] Bregar, K., Mohorčič, M.: Improving indoor localization using convolutional neural networks on computationally restricted devices. *IEEE Access* **6**, 17429–17441 (2018). <https://doi.org/10.1109/ACCESS.2018.2817800>
- [6] Bregar, K., Mohorčič, M.: Nlos identification in wlans using deep lstm with cnn features. *Sensors (Basel)* **18**(11) (2018). <https://doi.org/10.3390/s18114057>
- [7] Gazzah, L., Najjar, L., Besbes, H.: Improved hybrid aml algorithm without identification of los/nlos nodes. 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), 1–5 (2015). <https://doi.org/10.1109/VTCSpring.2015.7145874>
- [8] Nyqvist, H.E., Skoglund, M.A., Hendeby, G., Gustafsson, F.: Pose estimation using monocular vision and inertial sensors aided with ultra wide band. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–10 (2015). <https://doi.org/10.1109/IPIN.2015.7346940>
- [9] Hancer, E., Xue, B., Zhang, M.: Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems* **140**, 103–119 (2018). <https://doi.org/10.1016/j.knosys.2017.10.028>
- [10] Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems 30 (NIP 2017)* (2017)
- [11] Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications* **80**(5), 8091–8126 (2021). <https://doi.org/10.1007/s11042-020-10139-6>
- [12] Jeong, E.R., Joung, J., Jung, S., Chung, S.: Cnn-based tx–rx distance estimation for uwb system localisation. *Electronics Letters* **55**(17) (2019)
- [13] Xuemei, L., Yitong, X.: Improved xgboost model based on genetic algorithm for hypertension recipe recognition. *COMPUTER SCIENCE* **45**(S1), 476–481 (2018)
- [14] Huang, C., Molisch, A.F., He, R., Wang, R., Tang, P., Ai, B., Zhong, Z.: Machine learning-enabled los/nlos identification for mimo systems in dynamic environments. *IEEE Transactions on Wireless Communications*

19(6), 3643–3657 (2020). <https://doi.org/10.1109/TWC.2020.2967726>

- [15] Ding, L., Zhu, X., Zhou, T., Wang, Y., Jie, Y., Su, Y.: Research on uwb-based indoor ranging positioning technolog and a method to improve accuracy. In: 2018 IEEE Region Ten Symposium (Tensymp), pp. 69–73 (2018). <https://doi.org/10.1109/TENCONSpring.2018.8692064>