





Research Article

Light Gradient Boosting Machine-Based Link Quality Prediction for Wireless Sensor Networks

Linlan Liu ¹, Mingxiao Niu ¹, Chao Zhang ¹ and Jian Shu ²

¹School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China

²School of Software, Nanchang Hangkong University, Nanchang 330063, China

Correspondence should be addressed to Linlan Liu; 765693987@qq.com

Received 29 April 2022; Revised 18 July 2022; Accepted 19 July 2022; Published 5 August 2022

Academic Editor: Shaohua Wan

Copyright © 2022 Linlan Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Link quality prediction is a fundamental component of the wireless network protocols and is essential for routing protocols in wireless sensor networks (WSNs). Effective link quality prediction can select high-quality links for communication and improve the reliability of data transmission. In order to improve the accuracy of the link quality prediction model and reduce the model complexity, the link quality prediction model based on the light gradient boosting machine (LightGBM-LQP) is proposed in this paper. Specifically, agglomerative hierarchical clustering and manual division are combined to grade the link quality and obtain the labels of samples. Then, light gradient boosting machine (LightGBM) classification algorithm and Focal Loss are used to estimate the link quality grades. In order to reduce the impact of data imbalance, Borderline-SMOTE is employed to oversample the minority link quality samples. Finally, LightGBM-LQP predicts link quality grade at the next moment with historical link quality information. The experimental results on data collected from a real-world WSNs show that the proposed model has better prediction accuracy and shorter predicting time compared to related models.

1. Introduction

Wireless sensor networks (WSNs) are multihop self-organizing networks that are composed of a large number of inexpensive microsensor nodes [1]. WSNs are widely employed in military, intelligent logistics, industrial security monitoring, and other fields due to the comparatively low cost of sensor nodes and the functions of communication, perception, and computing [1–3].

Link quality prediction is the basis of topology control, routing, mobile management, and other upper layer protocols in WSNs, and it has a direct impact on communication performance and network scale [4]. Indeed, during the communication process, the wireless signal is affected by multipath interference, background noise, shadow fading, and other factors [5]. This is the main reason for wireless link quality fluctuations. There are many low-quality links in WSNs. When a data packet fails to be delivered, WSNs will resend it using the retransmission mechanism. Although this approach ensures that the data packet is successfully sent,

the increased delay reduces the network's overall efficiency. It is a fact that efficient upper layer protocols require an approach to avoid excessive retransmission on low-quality links [6–8]. To cope with this issue, the routing protocols of WSNs use various link quality prediction model (LQP) to select the best available paths considering higher-quality links, thereby preventing packet loss in advance [9]. The efficient LQP can serve as the foundation for upper layer routing to select the optimum connection, ensuring data transmission timeliness and stability for energy limited WSNs.

In order to improve the accuracy of the LQP, reduce the impact of data imbalance and decrease the complexity of the LQP; a link quality prediction model based on the light gradient boosting machine (LightGBM-LQP) is proposed. Specifically, the light gradient boosting machine (LightGBM) algorithm is used to estimate and predict the link quality because of its great training efficiency and high accuracy [10]. Borderline-SMOTE is used to adjust the distribution of the dataset to solve the problem of data imbalance.

The main contributions of this paper are as follows:

- (1) The link quality is graded by a combination of agglomerative hierarchical clustering and manually grading (AHCMG), and then, an effective grading is made according to the actual distribution of the samples
- (2) Borderline-SMOTE and Focal Loss functions are used to improve the accuracy of the link quality estimation model based on the light gradient boosting machine (LightGBM-LQE). By synthesizing new minority samples, Borderline-SMOTE is used to reduce interference caused by data imbalance. Focal Loss function is used in the estimation model training process to direct the model's attention on training error-prone link quality samples and improve the accuracy
- (3) A link quality prediction method based on LightGBM regression algorithm is proposed. LightGBM-LQP is designed to predict the link quality grade at the next moment based on the results of the LightGBM-LQE. The experimental results show that the proposed LQP has better prediction accuracy and shorter predicting time compared with other link quality prediction models in single-hop WSNs

The rest of this paper is organized as follows: Section 2 summarizes some related works in the field. Section 3 describes the problem description and the process of our model, including grading link quality, the construction of the LightGBM-LQE, and the presentation of the LightGBM-LQP. Section 4 provides the experimental results and discusses them, and Section 5 concludes this paper.

2. Related Work

This section mainly introduces the recent research of link quality prediction. The LQP in WSNs can be divided into LQP based on link characteristics, LQP based on statistics, and LQP based on machine learning.

2.1. LQP Based on Link Characteristics. LQP based on link characteristics mainly uses the physical layer parameters to predict the link quality. The main physical layer parameters include the link quality indicator (LQI), received-signal-strength indication (RSSI), and signal-to-noise ratio (SNR).

Audéoud and Heusse [11] pointed out that using only LQI as the input parameter of the prediction model is more effective than using only RSSI as the input parameter. They proposed a two-stage model: first, most of the links are quickly determined as available or unavailable by LQI, and then, the remaining links are tested to obtain their link quality. Instead of pursuing high prediction accuracy, the method focuses on quickly determining whether the link can be used. Fu et al. [12] proposed an abnormal link detection system RADIUS. In the case of rapid changes in RSSI, by determining the Bayesian threshold, combined with sliding time window data smoothing, distributed adaptation, and other technologies, as much as possible distinguishing

good links and bad links, experiments showed that the system can detect abnormal links and maintain stable errors.

Physical layer parameters are easy to obtain and can quickly reflect changes in link quality. They have the advantages of low cost and high agility and are often used for link quality prediction. Although LQP based on link characteristics provides a fast and inexpensive way to predict link quality, it is prone to overestimate link quality due to packet loss.

2.2. LQP Based on Link Statistics. LQP based on statistics mainly uses statistics and other methods to predict link quality. The commonly used parameters include packet reception rate (PRR), required number of packets (RNP), and expected transmission count (ETX).

Aiming at the problem of insufficient reliability of IEEE 802.15.4 link quality protocol in hospital environment, Akbar et al. [13] proposed a link quality prediction method based on fuzzy logic. First, the thresholds of LQI and RSSI of different link qualities in different hospital environments were obtained through experiments. Then, LQI, RSSI, and error rate (ER) are used as input parameters to predict link quality according to fuzzy logic rules.

LQP based on statistics mainly uses statistics and other methods to analyze the relationship between link layer parameters and link quality and build a mapping model to predict the link quality. However, the parameters of the data link layer need to be obtained by sending many probe packets, resulting in extra overhead and low real-time performance.

2.3. LQP Based on Machine Learning. LQP based on machine learning mainly excavates the potential relationship between input and output variables by machine learning method and predicts the link quality with the constructed model [14].

Feng et al. [15] proposed LQP based on extreme gradient boosting (XGBoost-LQP) to predict the link quality grade at the next moment. The correlation between hardware parameters and PRR was analyzed according to Pearson's correlation coefficient, and then, the input parameters of the model were determined. Experiment results showed that the proposed method makes better predictions in single-hop wireless sensor networks. Xue et al. [16] proposed LQP based on the random vector functional link network (RVFL-LQP) to predict link quality for WSNs in smart grid. According to the characteristic analysis of the wireless link, the original SNR sequence was decomposed into a time-varying sequence and a random sequence. Then, a time-varying link quality prediction model and a random link quality prediction model are constructed, respectively. They can predict the probability guarantee interval of SNR according to the output results of the two models. The experimental results showed that the method can effectively predict the probability guarantee interval of SNR and reflect the change of link quality and has high stability. Xu et al. [17] proposed a new model, which adopts recurrent neural network (RNN) to predict the LQI series, and then evaluated the link quality according to the fitting model of LQI and PRR. This method accurately mined the inner relationship among LQI series with the help of short-term memory

TABLE 1: Summary of related work on link quality prediction.

Type	Name	Input	Output	Strength	Weakness
LQP based on link characteristics	Two-stage model [11]	LQI	Classify link as reliable or weak	Quickly determining whether the link can be used.	The remaining links need more testing to be classified.
	RADIUS [12]	RSSI	Classify link as good or weak	It can adapt to dynamic environment changes.	The model accuracy is not high.
LQP based on statistics	FLS [13]	LQI, RSSI, and ER	Classify link as very low, low, medium, or high	It defines a general guideline and can be applied on other routing protocols.	The model accuracy is not high.
	XGBoost-LQP [14]	RSSI, LQI, and SNR	Classify link as bad, medium, or good	Data imbalance is addressed.	The model has high overhead.
LQP based on machine learning	RVFL-LQP [15]	SNR	Probability-guaranteed interval boundary of SNR	The dynamic stochastic features of link quality are described.	The results of the model need to further determine whether the link is available.
	RNN-LQP [16]	LQI	LQI	The temporal correlations of physical layer parameter series are considered.	The model has high time complexity.

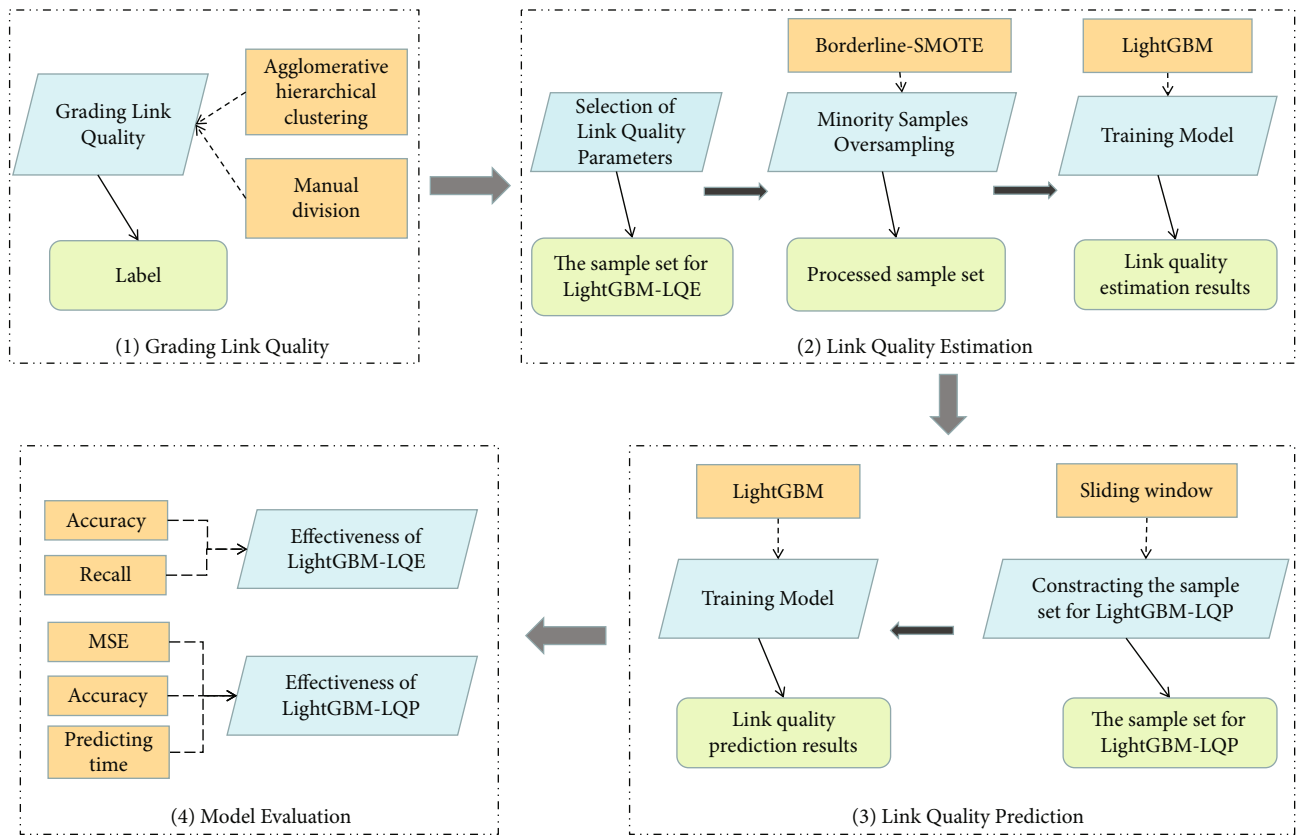


FIGURE 1: Block diagram of our methodology.

characteristics of RNN and effectively dealt with link fluctuations. Experiments showed that the proposed method is more suitable for low power wireless links with more fluctuations.

LQP based on machine learning can effectively mine the characteristics of training samples and deeply learn the

potential relationship between link quality parameters and link quality. Compared with other methods, it has higher accuracy and agility. It is important to note that most LQP models based on machine learning are offline models rather than online models. In other words, we train the model offline and then use the learned rules online. Therefore, the

Input: D (link quality data set), k_1 (the number of clustering), k_2 (the number of link quality grades).
Output: Link quality grades.
1: Take each sample in the link quality sample set as a cluster;
2: Calculate Euclidean distance for cluster merging by using $ED = D_i - D_j$;
3: When the number of clusters is reduced to k_1 , the agglomerative hierarchical clustering is changed to manual division;
4: According to the sample distribution among clusters, the distance between clusters, the link quality grade is manually graded into k_2 categories;
5: Rank link quality grade according to the average PRR of each cluster;
6: return: Link quality grades.

ALGORITHM 1: Grading link quality based on AHCMG algorithm.

Input: S' (Training sample set), L_{maj} (majority class label), L_{min} (minority class label).
Output: New training sample set S.
1: for every s_i in S' **do**
2: if the label of s_i in L_{maj}
3: Put s_i into majority class sample set C_{maj}
4: else
5: Put s_i into minority class sample set C_{min}
6: for every p_i in C_{min} **do**
7: Calculate its m nearest neighbors from the whole training set;
8: Calculate the number of samples belonging to the majority class in m neighbor samples, denoted as m' ($0 \leq m' \leq m$);
9: if $m/2 \leq m' < m$
10: Put p_i into boundary sample set $C_{boundary}$;
11: else
12: Put p_i into safe sample set C_{safe} ;
13: for every q_i in $C_{boundary}$ **do**
14: Synthesize the new minority samples by using (2);
15: Merge the synthesized minority sample with the original training sample to generate a new training sample set S;
16: return: new training sample set S.

ALGORITHM 2: Borderline-SMOTE algorithm for data imbalance.

time and resources consumed in the training of machine learning methods can not affect the use of LQP in WSNs. At present, LQP based on machine learning is the development trend of link quality prediction in wireless sensor networks.

Table 1 summarizes the related link quality prediction models, including the type, name, input, output, strength, and weakness of each model.

3. Modeling

In this section, we describe the link quality prediction problem in WSNs and introduce our methodology as well as related steps.

3.1. Problem Description. Give $P = \{P_1, P_2, \dots, P_t\}$ as the link quality parameters sequence, where $P_i (i = 1, 2, \dots, t)$ represents the link quality parameters set at time i , which is a subset of $\{RSSI, LQI, SNR, PRR, RNP, ETX\}$. Then, the link quality $LQ = \{LQ_1, LQ_2, \dots, LQ_t\}$ can be obtained by using $LQ = f(P)$, where $LQ_i (i = 1, 2, \dots, t)$ represents the link quality at time i and $f(\cdot)$ is the mapping function between link quality param-

eters and link quality. The link quality prediction in WSNs is to predict the link quality at the next moment based on the historical information of link quality within a period (sliding window w). Specifically, the link quality prediction problem can be defined as inputting a sequence $LQ_{i:i+w-1} = \{LQ_i, LQ_{i+1}, \dots, LQ_{i+w-1}\}$ to predict LQ_{i+w} .

3.2. Our Methodology. In this paper, LightGBM is employed to predict link quality. Notably, we use the link quality grades to represent link quality. Firstly, agglomerative hierarchical clustering and manual division are combined to grade the link quality, and the labels of samples are obtained. Then, we build LightGBM-LQE to estimate the current link quality grade. Finally, LightGBM-LQP predicts the next link quality grade by using historical link quality grades achieved by LightGBM-LQE. Figure 1 shows the block diagram of our methodology.

3.3. Grading Link Quality. In this paper, we use the link quality grades to represent link quality. At present, link quality grade classification methods are mainly divided into hard partition and soft partition. Hard partitioning is a method of

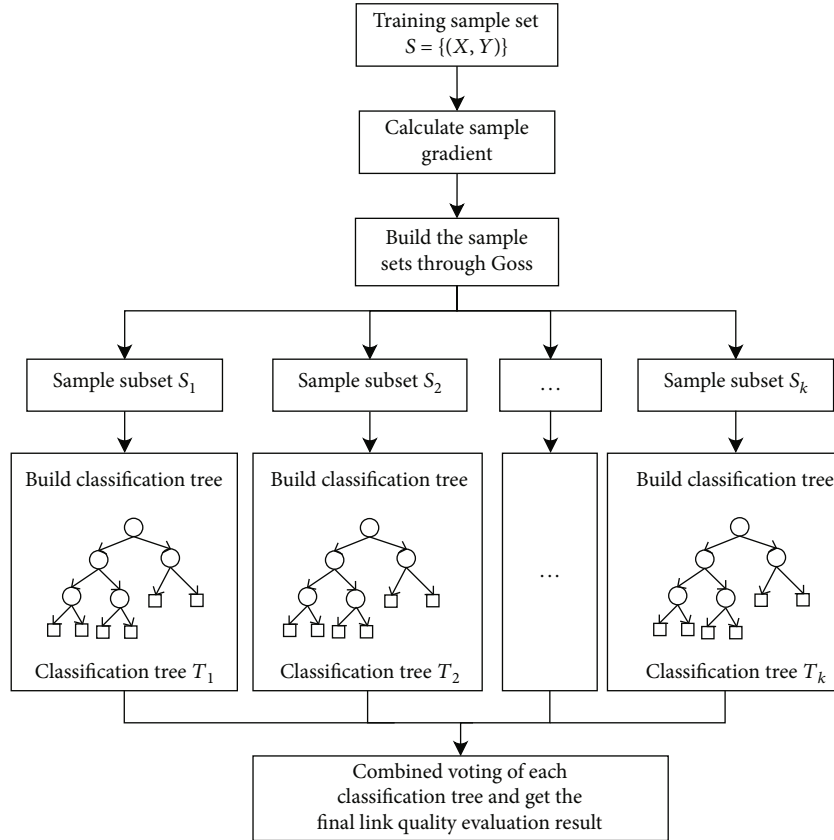


FIGURE 2: Link quality estimation process based on LightGBM.

TABLE 2: The selection of experimental parameters.

Parameters	Values
Transmission	31 dbm
Channel	26
Number of probe packets	30
Detection method	Active detection
Packet rate	10 pcs/s
Test cycle	8 s

grading link quality based on experimental results or prior information. Soft partition mainly adopts a clustering algorithm to grading link quality based on the link quality distribution. Indeed, the hard partition method is straightforward, but it is too subjective and coarse-grained. In this paper, AHCMG is used to grade link quality.

Agglomerative hierarchical clustering is a density-based clustering algorithm, which can solve nonspherical clusters [18]. The AHCMG method is used to prevent the problem of the link quality graded too finely. The specific steps are shown in Algorithm 1.

3.4. Link Quality Estimation

3.4.1. Selection of Link Quality Parameters. The physical layer parameters can be obtained quickly from the nodes

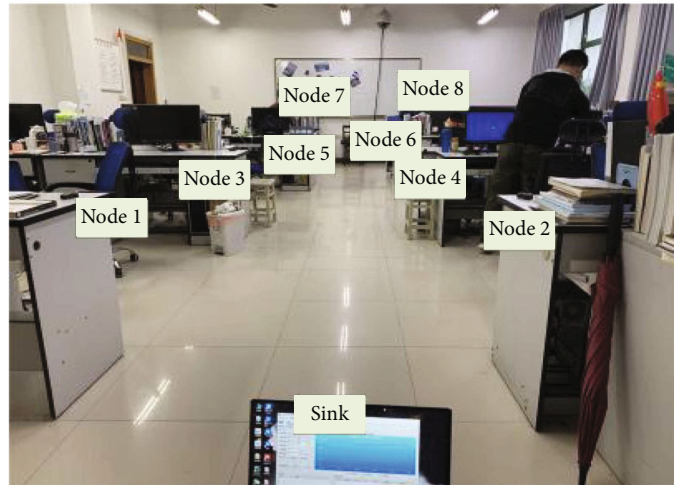
and are often taken as link quality estimation parameters. RSSI can determine whether a link is in the transition region quickly and accurately [19]. According to the mean and variance of LQI, it can partially accurately and stably portray the status information of link quality [20]. SNR cannot be used to accurately evaluate link quality on its own, but it can be used as an auxiliary parameter to improve the accuracy of the model [21].

A single physical layer parameter can only represent a specific link characteristic and cannot fully represent the link quality status. Therefore, we use multiple parameters. In this paper, the mean and coefficient variations (CV) of RSSI, LQI, and SNR are chosen as link quality parameters. The CV of RSSI, LQI, and SNR are calculated by

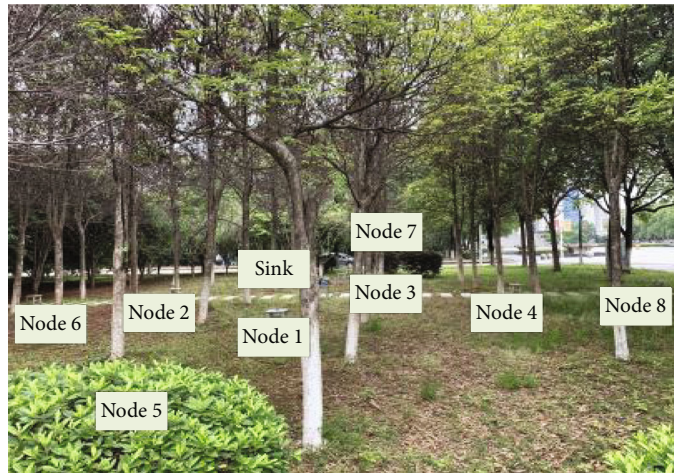
$$CV_i = \frac{\sigma(i)}{\mu(i)}, \quad i \in \{\text{RSSI, LQI, SNR}\}, \quad (1)$$

where $\sigma(i)$ is the standard deviation of link quality parameters in a period and $\mu(i)$ is the mean of link quality parameters in a period.

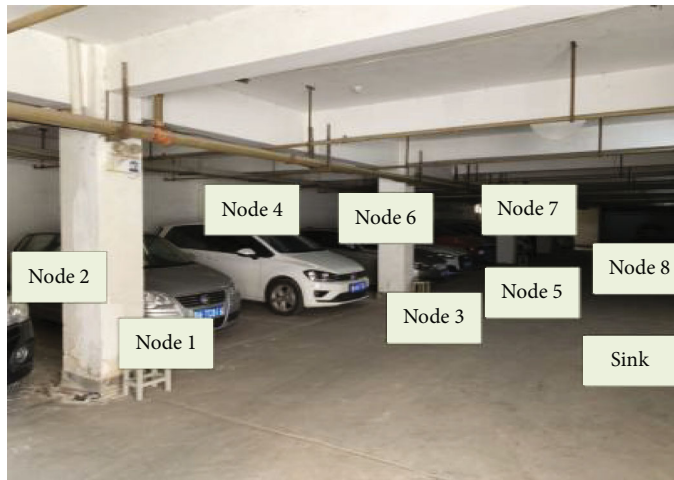
3.4.2. Minority Sample Oversampling. Research showed that the width of the links with intermediate quality is significant, ranging from 50% up to 80% of the transmission range [22], which means that the link quality usually shows an unbalanced condition. Therefore, the dataset we collected is unbalanced. Ali et al. [23] argued that in the classification



(a)



(b)



(c)

FIGURE 3: Experimental scenarios: (a) laboratory, (b) grove, and (c) parking.

of unbalanced samples, minority class boundary samples are more likely to be misclassified. To cope with this issue, Borderline-SMOTE is employed in this paper.

As an optimization of synthetic minority oversampling technique (SMOTE) [24], Borderline-SMOTE [25] is a minority samples oversampling method. Borderline-SMOTE

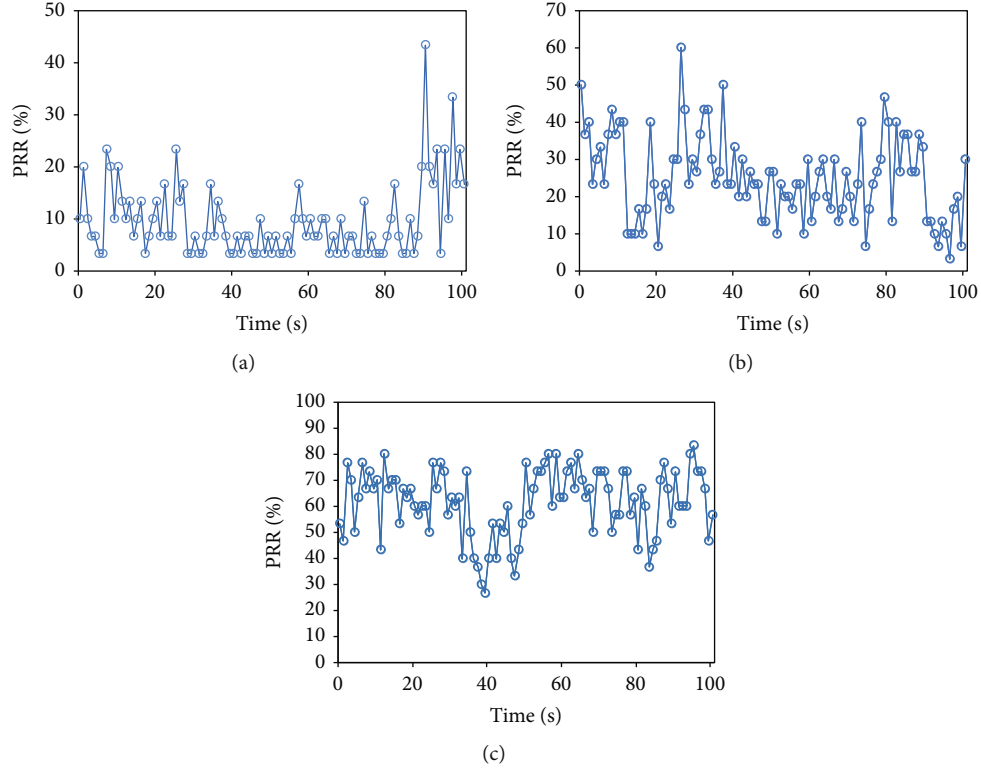


FIGURE 4: Time series diagram of the PRR in three scenarios: (a) laboratory, (b) grove, and (c) parking.

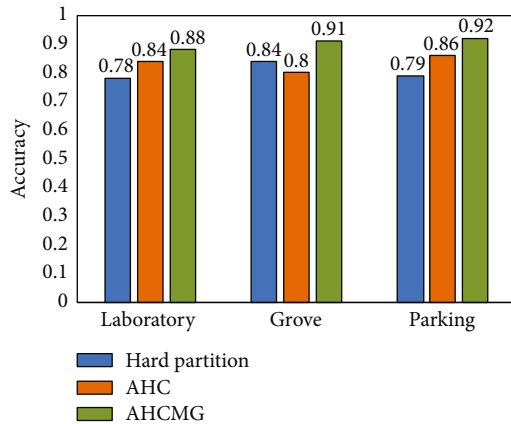


FIGURE 5: Comparison of estimation model results obtained by different link quality classification methods.

only uses minority samples on the boundary to synthesize new samples, so as to improve the class distribution of samples. The oversampling is calculated by

$$\text{synthetic}_j = q_i + r_j \times d_j \quad (j = 1, 2, \dots, s), \quad (2)$$

where q_i denotes the i -th link quality boundary samples, s denotes the number of link quality samples with q_i linear inter-

polation among neighbors, d_j denotes the distance between q_i and the neighbor j , and r_j is a random number between 0 and 1.

The specific steps are shown in Algorithm 2.

3.4.3. Construction of the LightGBM-LQE. LightGBM is a distributed gradient boosting algorithm based on the decision tree technique that is fast and high performance. It is an efficient implementation of GBDT [26] that can be used for sorting, classification, and regression in machine learning problems. LightGBM is a boosting algorithm that combines multiple weak learners to obtain strong learners with high performance. In this paper, classification and regression tree (CART) [27] is used as the base learner to construct the LightGBM-LQE. The specific process is shown in Figure 2.

LightGBM uses the gradient-based one-side sampling (GOSS) [10] algorithm to sample k times from the dataset and construct k sample subsets. The link quality samples in each subset are different, and we constructed classification trees for each subset separately. With LightGBM-LQE, the estimated link quality is achieved by combining the results of k classification trees.

GOSS is an algorithm that generates different sample subsets based on the gradient of the link quality samples. Note that samples with big gradients can participate in training almost every time by using GOSS, so the model pays more attention to the big gradient samples. According to GOSS algorithm characteristics, we use the Focal Loss [28] to replace the original loss function of LightGBM. Focal Loss can dynamically adjust the sample gradient according to the training results. If the sample is easy to be classified

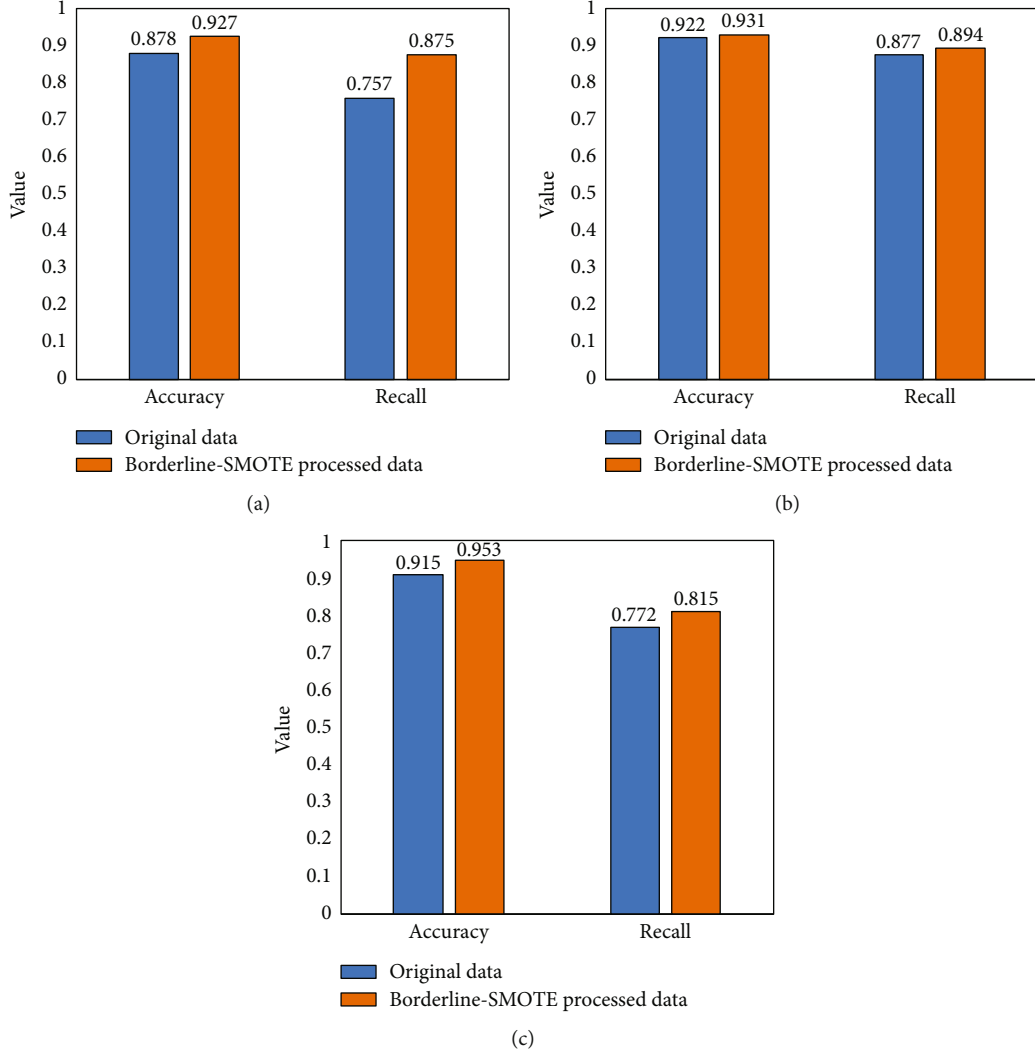


FIGURE 6: Accuracy and recall comparison of different data processing methods in the three scenarios: (a) laboratory, (b) grove, and (c) parking.

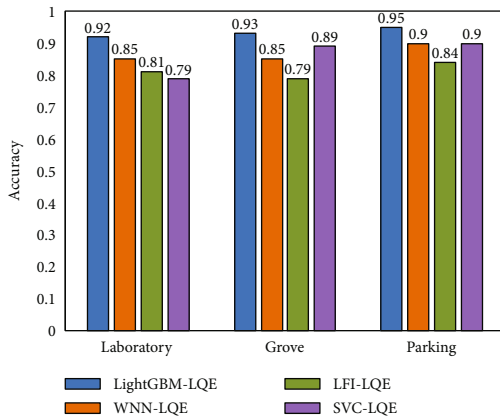


FIGURE 7: Comparison of accuracy of estimation models.

incorrectly, Focal Loss will increase the gradient of the sample so that it can participate in more training. Focal Loss is calculated

$$FL(L_i) = -\alpha_i(1 - p_i)^\gamma \log(L_i), \quad (3)$$

where L_i is the classification probability of link quality, α_i represents a parameter between 0 and 1, and γ is the focus parameter.

After k classification trees are obtained, the final output result is generated by voting, which is calculated

$$H(\chi) = \arg \max_Y \sum_{i=1}^K I(h_i(\chi) = Y), \quad (4)$$

where $H(\chi)$ represents the final link quality estimation grade, Y is the link quality grade, $h_i(\chi)$ is the classification result of the i -th classification tree, and $I(\cdot)$ is the indicative function.

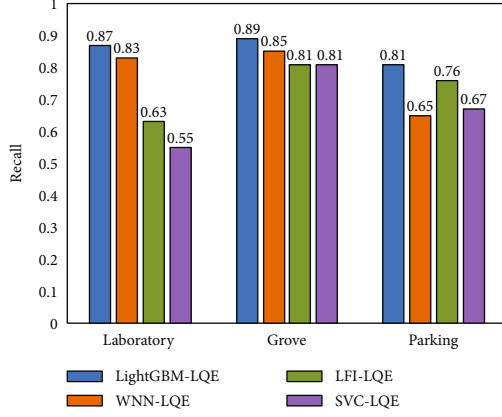


FIGURE 8: Comparison of recall of estimation models.

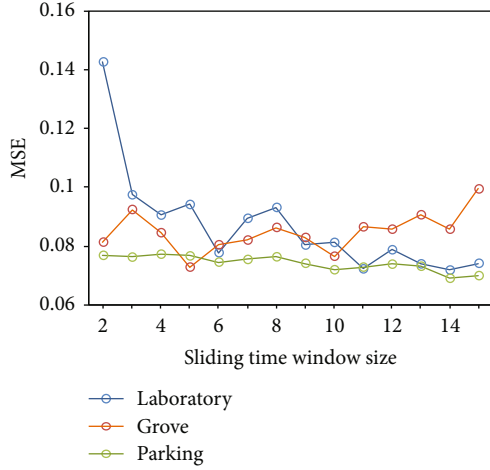


FIGURE 9: MSE of the LightGBM-LQP in different windows.

3.5. Link Quality Prediction. According to the characteristic that link quality shows strong temporal correlation in a short term, we take the link quality grade time series obtained by the LightGBM-LQE as input and use the sliding window method to determine the length of the input sequence. Then, multiple regression trees are trained to predict the link quality grade at the next moment.

The LightGBM regression algorithm is employed to construct LightGBM-LQP which predicts the link quality at the next moment. First, we calculate the gradient of each link quality sample. Then, we sort the samples according to the gradient. By retaining big gradient samples every time and randomly sampling small gradient samples, we train different regression trees through using different sample subsets. Each regression tree selects the largest gain node for splitting through the information gain formula during the training process. The information gain is calculated by

$$V_j(d) = \frac{1}{n} \left[\frac{(\sum_{x_i \in A} g_i + ((1-a)/b) \sum_{x_i \in B} g_i)^2}{n_l^j(d)} + \frac{(\sum_{x_i \in A} g_i + ((1-a)/b) \sum_{x_i \in B} g_i)^2}{n_r^j(d)} \right], \quad (5)$$

where n represents the number of link quality samples for each leaf node, a represents the proportion of big gradient samples, b represents the proportion of random sampling, g_i is the gradient of the i -th link quality sample, $n_l^j(d)$ is the total number of samples of the left child node, $n_r^j(d)$ is the total number of samples of the right child node, and j is the j -th node.

After M regression trees are built, the final prediction result of the model is calculated by

$$\widehat{Y}_i = \frac{\sum_{j=1}^k [f_j(x_i) * \omega_j]}{M}, \quad i = 1, 2, \dots, m, \quad (6)$$

where \widehat{Y}_i is the final prediction result, $f_j(x_i)$ is the prediction result of the j -th tree, ω_j is the weight of the j -th tree, and m is the number of test sets.

4. Experiment Setup and Analysis

TelosB nodes made by CrossBow are used to send and receive packets [29], and the WSN link quality testbed (LQT) developed by our lab is used to collect the link quality parameters. The experimental parameters of the LQT are shown in Table 2.

4.1. Model Evaluation. In our experiments, accuracy and recall are taken to evaluate the performance of LightGBM-LQE; the mean square error (MSE) and accuracy are applied to evaluate the performance of LightGBM-LQP.

Accuracy is the most common evaluation metric of classification models. The higher the accuracy, the better the performance of the classifier. The accuracy is calculated by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \quad (7)$$

where TP is the number of positive samples predicted to be positive, FP is the number of negative samples predicted to be positive, FN is the number of positive samples predicted to be negative, and TN is the number of negative samples predicted to be negative.

Recall is the percentage of predicted positive samples of the actual positive samples. The recall is calculated by

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (8)$$

Smaller MSE means higher prediction accuracy and vice versa. The MSE is calculated by

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (Y_i - \widehat{Y}_i)^2, \quad (9)$$

where N represents the total number of samples in the test set, Y_i is the true value, and \widehat{Y}_i is the predicted value.

4.2. Experimental Scenarios and Data Analysis. Considering the different interference factors, three scenarios with different representative interference are selected for experiments,

TABLE 3: MSE and accuracy of prediction models in the three scenarios.

	MSE			Accuracy		
	Laboratory	Grove	Parking	Laboratory	Grove	Parking
LightGBM-LQP	0.061	0.073	0.064	0.937	0.926	0.936
XGBoost-LQP	0.184	0.089	0.087	0.874	0.911	0.912
RVFL-LQP	0.079	0.084	0.071	0.930	0.916	0.929
RNN-LQP	0.093	0.086	0.077	0.911	0.914	0.923
SVR-LQP	0.085	0.086	0.074	0.915	0.914	0.926
GRU-LQP	0.090	0.084	0.072	0.918	0.916	0.928

TABLE 4: Predicting time of LQPs in the three scenarios.

	Predicting time (s)		
	Laboratory	Grove	Parking
LightGBM-LQP	0.031	0.013	0.031
XGBoost-LQP	0.343	0.087	0.281
RVFL-LQP	0.296	0.078	0.187
RNN-LQP	1.147	0.146	0.657
SVR-LQP	0.398	0.056	0.296
GRU-LQP	0.941	0.103	0.559

including laboratory scenario, grove scenario, and parking scenario, as shown in Figure 3. In each scenario, 1 sink node and 8 perception nodes are employed.

In the laboratory scenario, electronic communication devices such as Wi-Fi and Bluetooth may cause neighboring frequency interference and blocking interference, substantially impacting communication quality. In addition, human movement can disrupt communication between nodes. The sequence diagram of PRR in the laboratory scenario is shown in Figure 4(a). In general, the link quality of this node is abnormally unstable, with most PRR falling below 20%.

In the grove scenario, the occlusion of trees, light refraction of leaves, outdoor temperature, and air humidity can interfere with node communication. In addition, in this circumstance, the multipath effect is a major element influencing link quality. In Figure 4(b), we can find out that the PRR is fluctuant and at approximately 30%.

The interference in the parking scenario is mainly generated by the movement of vehicles and vehicle-mounted wireless devices. The characteristics of this kind of interference are the spectrum width, noise intensity decreases with the increase of spectrum, and the types of noise sources are very random. As shown in Figure 4(c), the link quality is relatively stable, with the exception of time 35-55 s. The link quality declines dramatically in time 35-55 s mainly because of vehicle passing.

4.3. Effectiveness of the Grading Link Quality Method. In this subsection, the AHCMG method is compared to the hard partition method and the agglomerative hierarchical clustering (AHC) method in three experimental scenarios to validate the effectiveness of the grading link quality method. Note that we calculate accuracy for the estimation model

under each grading method to compare the performance of different grading link quality method.

As can be seen in Figure 5, the AHCMG method has higher accuracy in all three experimental scenarios. In the laboratory scenario, the accuracy of using AHCMG method is 10% higher than the one of hard partition. In the grove scenario, the accuracy of using AHC is the lowest, which is only 80%. In the parking scenario, AHCMG method achieves the highest accuracy in the three scenarios. It is clear that grading link quality with AHCMG method is more efficient.

4.4. Effectiveness of Imbalanced Samples Handling. For the sake of improving the accuracy of the LightGBM-LQE on imbalanced data, Borderline-SMOTE is employed to over-sample the minority samples. In this subsection, recall and accuracy are used to evaluate the effectiveness of imbalanced samples handling.

Figure 6 shows the performance of LightGBM-LQE before and after oversampling. As we can see, in the laboratory scenario, the overall performance of the LightGBM-LQE is clearly improved, with accuracy increasing by 5% and recall increasing by about 12%. After Borderline-SMOTE processing, accuracy and recall increased by about 1% in the grove scenario. In the parking scenario, the accuracy and recall have increased by about 4%. We can conclude that oversampling the minority samples with Borderline-SMOTE can effectively reduce the impact of data imbalance and can improve the accuracy of the LightGBM-LQE as well.

4.5. Effectiveness of LightGBM-LQE. In order to verify the effectiveness of our estimation model, we compare the performance of LightGBM-LQE with the link quality estimation model based on wavelet neural network (WNN-LQE) [30], the model based on support vector classification (SVC-LQE) [31], and the model based on lightweight fluctuation (LFI-LQE) [32] and take accuracy and recall as the evaluation index. The comparison results are shown in Figures 7 and 8.

As shown in Figures 7 and 8, LightGBM-LQE has better accuracy and recall in all three experimental scenarios. In the laboratory scenario, it can be seen that LightGBM-LQE and WNN-LQE have high accuracy as well as high recall, while LFI-LQE and SVC-LQE have not bad accuracy with unsatisfactory recall. This may be because the laboratory scenario is

more complex, and LFI-LQE and SVC-LQE do not have good performance, especially in the minority samples.

In the grove scenario, all models have the highest recall. LFI-LQE, which has the lowest accuracy, also achieves a not bad recall. This could be because the interference received in the grove scenario is relatively simple and less. In the parking scenario, all estimation models have high accuracy, while the recall is unsatisfactory. In four models, LightGBM-LQE has the best performance. This could be because the types of noise sources are very random in the parking scenario. The comparison of accuracy and recall shows that the performance of LightGBM-LQE is superior to other estimation models.

4.6. Determination of Sliding Time Window Size. In this subsection, MSE is used to determine the sliding time window size. When different time window sizes are used, the MSE of the LightGBM-LQP can be computed. We use a sliding time window with a size ranging from 2 to 15 and choose the window size with the best prediction accuracy.

It can be seen from Figure 9 that the MSE of the LightGBM-LQP fluctuates with the change of the sliding time window size. Therefore, the size of the sliding time window has an important impact on the link quality prediction at the next moment. The prediction model has the best prediction accuracy in the parking scenario in which the sliding window size is 14. The best sliding window sizes in the laboratory and grove scenarios are 14 and 5, respectively.

4.7. The Evaluation Performance of LightGBM-LQP. To verify the performance of the LightGBM-LQP, we conduct comparative experiments with other LQPs using the same datasets in three scenarios. The compared models include XGBoost-LQP [15], RVFL-LQP [16], RNN-based model (RNN-LQP) [17], support vector regression-based model (SVR-LQP) [31], and gated recurrent unit-based model (GRU-LQP) [33].

To verify the prediction effect of the LightGBM-LQP, MSE and accuracy of LQPs in three scenarios are shown in Table 3.

From Table 3, the comparison results show that the LightGBM-LQP has the lowest MSE and the best accuracy in three scenarios. Compared with XGBoost-LQP, SVR-LQP, and RNN-LQP, LightGBM-LQP has a better performance in all three experimental scenarios, especially in the laboratory scenario. This may be because the laboratory scenario contains more types of interference and is more complex. RVFL-LQP is the second-best model. Compared with LightGBM-LQP, the MSE of RVFL-LQP is increased by 0.5%~0.6%, and the accuracy is decreased by 0.7%~1%. In three scenarios, both LightGBM-LQP and RVFL-LQP have the worst effect in the grove scenario, probably because the mutation of link quality is prone to occur in the grove scenario, while LightGBM-LQP and RVFL-LQP mask these mutations out. When the confidence level is 95%, the confidence interval of the accuracy of the proposed approach is [0.913, 0.961] in the laboratory scenario, [0.906, 0.946] in the grove scenario, and [0.916, 0.956] in the parking scenario.

4.8. The Time Complexity of LightGBM-LQP. Note that the model proposed in this paper is an offline model. In order to ensure the identical distribution of the training data and actual data, the data collected in real scenarios is used to train the model, and then, the trained model is deployed to the WSN nodes for prediction. The time complexity of LightGBM-LQP is $O(2 * M * \text{depth})$, where M represents the number of regression trees and depth represents the depth of trees. In this paper, M is 100, and the maximum depth of the tree is 6. Therefore, it is feasible to deploy the trained model to the node.

Table 4 shows the predicting time of LQPs for the same data to demonstrate the low complexity of LightGBM-LQP in comparison with other models. As shown, LightGBM-LQP has the shortest predicting time among the six LQPs. The predicting time of XGBoost-LQP, RVFL-LQP, and SVR-LQP is slightly longer than that of LightGBM-LQP, while the prediction time of RNN-LQP and GRU-LQP are much longer, which may be due to the fact that RNN-LQP and GRU-LQP are deep learning models with higher complexity.

5. Conclusion and Future Work

In this paper, a link quality prediction method LightGBM-LQP is proposed. Firstly, the link quality is graded by AHCMG to get the labels. Then, LightGBM classification algorithm is used to estimate the current link quality grade. In order to improve the estimation accuracy, Borderline-SMOTE is used to synthesize new samples by using the minority class samples on the boundary, so as to reduce the impact caused by data imbalance. Meanwhile, Focal Loss function is applied in the estimation model training process to make the model focus on training error-prone link quality samples. Finally, LightGBM-LQP predicts link quality grade at the next moment with historical link quality information. In three different experimental scenarios, compared with XGBoost-LQP, SVR-LQP, etc., LightGBM-LQP has the best performance. Therefore, the proposed LightGBM-LQP could be a very promising approach for link quality prediction.

LightGBM-LQP is an offline model that cannot respond to the changes of the external environment. In the future, we will focus on constructing an online model for link quality prediction that is both simple and efficient. Batch learning will be considered in the link quality prediction model to improve the responsiveness of the model.

Data Availability

The excel data used to support the findings of this study have been deposited in the github. The data can be obtained in the following link: <https://github.com/Azora-niu/Link-Quality-Estimation-Prediction-Data-for-WSN>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China under Grant 6196203 and Grant 62062050, in part by the Natural Science Foundation of Jiangxi Province under Grant 20202BABL202039, and in part by the Innovation Foundation for Postgraduate Student of Jiangxi Province under Grant YC2021-025.

References

- [1] Y. Chang, H. Tang, Y. Cheng, Q. Zhao, B. Li, and X. Yuan, "Dynamic hierarchical energy-efficient method based on combinatorial optimization for wireless sensor networks," *Sensors*, vol. 17, no. 7, p. 1665, 2017.
- [2] Y. Chang, X. Yuan, B. Li, D. Niyato, and N. Al-Dhahir, "A joint unsupervised learning and genetic algorithm approach for topology control in energy-efficient ultra-dense wireless sensor networks," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2370–2373, 2018.
- [3] T. J. Wang, X. Q. Li, and G. W. Bai, "Multi-target localization algorithm based on adaptive grid in wireless sensor network," *Journal of Communications*, vol. 40, no. 7, pp. 197–207, 2019.
- [4] L. Liu, H. Lv, J. Xu, and J. Shu, "A link quality estimation method based on improved weighted extreme learning machine," *IEEE Access*, vol. 9, pp. 11378–11392, 2021.
- [5] G. Cerar, H. Yetgin, M. Mohorčič, and C. Fortuna, "Machine learning for wireless link quality estimation: a survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 696–728, 2021.
- [6] Y. Huang, "Routing optimization of wireless sensor network based on energy and path constraints," *Journal of Xidian University*, vol. 47, no. 3, pp. 113–120, 2020.
- [7] C. Chen, L. Liu, S. Wan, X. Hui, and Q. Pei, "Data dissemination for Industry 4.0 applications in Internet of Vehicles based on short-term traffic prediction," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 1, pp. 1–18, 2022.
- [8] L. Zhao, C. Wang, K. Zhao, D. Tarchi, S. Wan, and N. Kumar, "INTERLINK: a digital twin-assisted storage strategy for satellite-terrestrial networks," *IEEE Transactions on Aerospace and Electronic Systems*, p. 1, 2022.
- [9] J. M. Reason and J. M. Rabaey, "A study of energy consumption and reliability in a multi-hop sensor network," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 6, pp. 26971–26992, 2018.
- [10] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [11] H. J. Audéoud and M. Heusse, "Quick and efficient link quality estimation in wireless sensors networks," in *2018 14th Annual Conference on Wireless on-demand Network Systems and Services (WONS)*, pp. 87–90, Isola 2000, France, 2018.
- [12] S. Fu, C. Y. Shih, Y. Jiang, M. Ceriotti, X. Huan, and P. J. Marrón, "RADIUS: a system for detecting anomalous link quality degradation in wireless sensor networks," 2017, <https://arxiv.org/abs/1701.00963>.
- [13] M. S. Akbar, H. Yu, and S. Cang, "Performance optimization of the IEEE 802.15. 4-based link quality protocols for WBASNS/IOTS in a hospital environment using fuzzy logic," *IEEE Sensors Journal*, vol. 19, no. 14, pp. 5865–5877, 2019.
- [14] D. P. Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: a survey," *Information Fusion*, vol. 49, pp. 1–25, 2019.
- [15] Y. Feng, L. Liu, and J. Shu, "A link quality prediction method for wireless sensor networks based on xgboost," *IEEE Access*, vol. 7, pp. 155229–155241, 2019.
- [16] X. Xue, W. Sun, J. Wang, Q. Li, G. Luo, and K. Yu, "RVFL-LQP: RVFL-based link quality prediction of wireless sensor networks in smart grid," *IEEE Access*, vol. 8, pp. 7829–7841, 2020.
- [17] M. Xu, W. Liu, J. Xu et al., "Recurrent neural network based link quality prediction for fluctuating low power wireless links," *Sensors*, vol. 22, no. 3, p. 1212, 2022.
- [18] L. Billard and E. Diday, "Agglomerative hierarchical clustering," in *Clustering Methodology for Symbolic Data*, pp. 261–316, John Wiley & Sons, Inc, 2019.
- [19] K. Srinivasan, P. Dutta, A. Tavakoli, and P. Levis, "Understanding the causes of packet delivery success and failure in dense wireless sensor networks," in *Proceedings of the 4th international conference on Embedded networked sensor systems*, pp. 419–420, Boulder, Colorado, USA, 2006.
- [20] D. Kim, H. Nam, and D. Kim, "Adaptive code dissemination based on link quality in wireless sensor networks," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 685–695, 2017.
- [21] C. J. Lowrance and A. P. Lauf, "Link quality estimation in ad hoc and mesh networks: a survey and future directions," *Wireless Personal Communications*, vol. 96, no. 1, pp. 475–508, 2017.
- [22] N. Baccour, A. Koubâa, L. Mottola et al., "Radio link quality estimation in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 8, no. 4, pp. 1–33, 2012.
- [23] A. Ali, S. M. Shamsuddin, and A. Ralescu, "Classification with class imbalance problem," *International Journal of Advances in Soft Computing and its Applications*, vol. 5, no. 3, 2013.
- [24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [25] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*, D. S. Huang, X. P. Zhang, and G. B. Huang, Eds., vol. 3644 of Lecture Notes in Computer Science, pp. 878–887, Springer, Berlin, Heidelberg, 2005.
- [26] Y. Zhang, X. Beudaert, J. Argandoña, S. Ratchev, and J. Munoa, "A CPPS based on GBDT for predicting failure events in milling," *The International Journal of Advanced Manufacturing Technology*, vol. 111, no. 1-2, pp. 341–357, 2020.
- [27] C. J. Drott, D. Norman, and D. Espes, "CART decreases islet blood flow, but has no effect on total pancreatic blood flow and glucose tolerance in anesthetized rats," *Peptides*, vol. 135, article 170431, 2021.
- [28] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, Venice, Italy, 2017.
- [29] A. D. Boursianis, M. S. Papadopoulou, A. Gotsis et al., "Smart irrigation system for precision agriculture—the AREThOU5A IoT platform," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17539–17547, 2021.
- [30] W. Sun, W. Lu, Q. Li, L. Chen, D. Mu, and X. Yuan, "WNN-LQE: wavelet-neural-network-based link quality estimation

- for smart grid WSNs,” *IEEE Access*, vol. 5, pp. 12788–12797, 2017.
- [31] J. Shu, S. Liu, L. Liu, L. Zhan, and G. Hu, “Research on link quality estimation mechanism for wireless sensor networks based on support vector machine,” *Chinese Journal of Electronics*, vol. 26, no. 2, pp. 377–384, 2017.
- [32] W. Liu, Y. Xia, R. Luo, and S. Hu, “Lightweight, fluctuation insensitive multi-parameter fusion link quality estimation for wireless sensor networks,” *IEEE Access*, vol. 8, pp. 28496–28511, 2020.
- [33] M. Abdel-Nasser, K. Mahmoud, O. A. Omer, M. Lehtonen, and D. Puig, “Link quality prediction in wireless community networks using deep recurrent neural networks,” *Alexandria Engineering Journal*, vol. 59, no. 5, pp. 3531–3543, 2020.