J Inf Process Syst, Vol.20, No.1, pp.131~147, February 2024
https://doi.org/10.3745/JIPS.03.0193

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Dynamic Caching Routing Strategy for LEO Satellite Nodes Based on Gradient Boosting Regression Tree

Yang Yang[1], Shengbo Hu[1,2,*], and Guiju Lu[3]

Abstract

A routing strategy based on traffic prediction and dynamic cache allocation for satellite nodes is proposed to address the issues of high propagation delay and overall delay of inter-satellite and satellite-to-ground links in low Earth orbit (LEO) satellite systems. The spatial and temporal correlations of satellite network traffic were analyzed, and the relevant traffic through the target satellite was extracted as raw input for traffic prediction. An improved gradient boosting regression tree algorithm was used for traffic prediction. Based on the traffic prediction results, a dynamic cache allocation routing strategy is proposed. The satellite nodes periodically monitor the traffic load on inter-satellite links (ISLs) and dynamically allocate cache resources for each ISL with neighboring nodes. Simulation results demonstrate that the proposed routing strategy effectively reduces packet loss rate and average end-to-end delay and improves the distribution of services across the entire network.

Keywords

Gradient Boosting Regression Tree Cache Allocation, Low-Earth Orbit Satellite, Inter-Satellite Links, Spatial and Temporal Correlation, Traffic Prediction

1. Introduction

Satellite communication has the advantages of abundant radio frequency (RF) resources, large coverage area, and low ground interference, and it plays an important role in communication and television broadcasting in remote areas [1]. The 6th generation mobile communication network will integrate ground and satellite communication and build an integrated air-ground–space network to achieve true full ground–air coverage in a wider space. The integrated air–ground–space network connects users, aircraft, and various communication platforms on the ground, at sea, in the air, and in deep space through inter-satellite links (ISLs) and satellite–ground links, realizing a high-capacity information network that accurately acquires information, processes it quickly, and efficiently transmits it [2]. As the backbone of this integrated network, satellite networks have advantages over traditional ground networks, such as global coverage, simple access, support for various services, and on-demand bandwidth allocation.

They also play an increasingly important role in global communication, navigation, and positioning; environmental and disaster monitoring; and military applications [3]. Although China's ground communication network is developing rapidly, it is difficult to cover special environments such as deserts, forests,

Manuscript received July 11, 2023; first revision November 6, 2023; accepted November 30, 2023.
* **Corresponding Author:** Shengbo Hu (hsb@gznu.edu.cn)
[1] School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China (yyt614026597@163.com, hsb@gznu.edu.cn)
[2] National Space Science Center, the Chinese Academy of Sciences (CAS), Beijing, China
[3] School of Mechanical and Electrical Engineering, Guizhou Normal University, Guiyang, China (LuGuiJu202109@163.com)

and oceans. Owing to their seamless global coverage capabilities, satellite networks can provide continuous communication connections in areas where ground networks cannot provide coverage and can provide reliable redundant connections in the event of wireless infrastructure damage in areas already covered by terrestrial networks.

In the construction of satellite communication systems, low-Earth orbit (LEO) satellites have significant advantages over geostationary Earth orbit (GEO) satellites. The round-trip delay of LEO satellites is less than 100 ms, while that of GEO satellites is approximately 600 ms. Thanks to the relatively short propagation distance of LEO satellites, the signal propagation loss is smaller, which is particularly beneficial for the low-power design of small devices, such as Internet of Things (IoT) terminals [4]. In addition, because GEO satellites are synchronized with the Earth, the position between the terminals and GEO satellites is relatively stationary, and the satellite–ground communication links are severely restricted by terrain. When there are obstacles in the line of sight between the terminal and satellite, the communication quality drops sharply. By contrast, LEO satellites and terminals are in relative motion, and a communication window exists even if the terminal is near an obstacle [5]. Owing to their late start, satellite systems in China are under significant frequency resource constraints and often need to be compatible with applications such as broadband Internet access and IoT services. Therefore, the coexistence of low- and high-speed services and delay-sensitive and -insensitive services is common. To ensure timely data transmission, ISLs are required to enable inter-satellite data routing. Therefore, designing reasonable routing strategies based on the connectivity provided by ISLs has become a key issue for enabling satellite systems to efficiently transmit data in a timely manner.

This paper first proposes a traffic prediction method for LEO satellite systems, followed by a dynamic cache allocation routing (DCAR) strategy based on traffic prediction. Currently, ground network traffic prediction models, such as time-series analysis, nonlinear analysis, and artificial intelligence, are widely used. For example, support vector machines (SVMs) [6] have been applied in actual ground network traffic prediction, and the prediction results are applicable to traffic allocation strategies. However, compared with ground networks, satellite networks have limited available resources and time-varying topologies. Thus, traditional ground network traffic prediction algorithms inevitably face challenges in terms of prediction accuracy and operational efficiency if directly applied to satellite networks. Combination models based on neural networks [7-9] have achieved good results in prediction due to their strong fault tolerance, fast parallel computing, and powerful learning ability. However, parameter selection can only rely on experience, and the algorithm has a high time complexity, long training time, and slow convergence speed. Therefore, this paper proposes a gradient boosting regression tree (GBRT)-based method. First, the time–space correlation of satellite network traffic is determined through correlation coefficients to determine time–space-related traffic. Then, the GBRT is used for training to obtain smooth parameters. Relatively accurate prediction values were obtained through testing. Second, a DCAR strategy based on traffic prediction is introduced, which can be adaptively adjusted during network operation. This strategy dynamically adjusts the cache length of ISLs based on traffic prediction results and selects the optimal path with the least end-to-end delay and highest cache hit rate.

2. Related Work

The following factors must be considered when designing routing strategies for satellite networks: first, satellite nodes have limited cache and processing capabilities, and routing strategies should not be too

complex. Second, the state of ISLs changes frequently changes. Whether it is link interruption caused by high bit error rate (BER), link interruption and reconnection caused by relative movement between satellites, or node failure, the network topology will be in an unstable state. Routing strategies must have adaptive rerouting capabilities under changing topologies while considering the signaling and processing overhead required for rebuilding routes [10].

In a collaborative satellite–ground network, both the satellite and ground can cache content to provide more efficient service when requested, reducing the communication load of repeated content transmission and round-trip links. Wu et al. [11] proposed a dual-layer satellite–ground caching scenario in which the ground used an information-centric network (ICN) to decouple data from the location where the service was actually provided through the joint optimization of satellite–ground content caching placement, reducing the uplink and downlink bandwidth consumption. An ICN decouples data using a location-independent naming method, which can solve the inconvenience caused by the mobility of the data source and improve the efficiency of content search and transmission by collecting information on the same file stored on different devices. There have also been studies involving multi-satellite distributed caching strategies. Liu et al. [12] proposed the selection of a LEO satellite constellation network, considered the scenario of multiple satellites jointly serving multiple ground terminals, and solved the cache placement problem using the stable matching algorithm. Zhong et al. [13] designed a multi-layer caching network for user–base station communication based on a ground social network, and remote scheduling of uncached content was accomplished using satellites.

In research on routing strategies for LEO networks, the initial routing strategy was connection-oriented. With the development of TCP/IP and networking, connectionless routing has attracted more attention [14]. Many satellite network routing strategies focus on masking the dynamic nature of LEO networks and determining the optimal path with minimal end-to-end delay, such as data packet routing based on virtual nodes and shortest path routing based on virtual topology [15]. To achieve a more even utilization of ISLs, the idea of alternate backup links was adopted to use source-based and all-node-based alternate link forwarding. To further consider the effect of traffic on routing strategies, routing selection was optimized based on the traffic flow category, thereby satisfying the requirements of different traffic classes [16]. However, both these strategies require the entire network topology state to be updated in advance every time the routing path is calculated, leading to high signaling overhead. In addition, obtaining an accurate network topology state in a time-varying LEO network is difficult. Inspired by ground ad hoc routing strategies, researchers proposed location-assisted on-demand routing (LAOR) [17]. The goal of LAOR is to reduce the end-to-end delay with minimal signaling overhead. To improve the signaling overhead of the LAOR strategy under a high business request frequency, an agent-based load balancing routing (ALBR) strategy, which defined two types of agents, mobile and fixed agents, was proposed [18]. Each satellite node deploys both types of agents during its operating period. However, when the mobile agent explores the available path between the source and destination nodes, it introduces additional traffic load and signaling overhead into the network. Furthermore, although LAOR and ALBR can collect network state information, due to the long propagation delay between satellite nodes, neither strategy can accurately obtain the actual network state.

Li et al. [19] proposed a state-aware and load-balanced (SALB) routing strategy for LEO networks to address load imbalance and node congestion. Their strategy employed queue occupancy grading based on link state estimation and delay weight adjustment, taking into account different scenarios, including load changes and node failures. The routing overhead was significantly reduced through route table

resetting and dynamic updates based on the shortest path tree algorithm. Li et al. [20] proposed a semi-distributed load balancing routing (SDLBR) algorithm that considers node congestion and link states. Based on the current congestion information, priority was set to respond to congestion, the correct forwarding direction was selected, and routing decisions were made based on real-time node status through a semi-distributed forwarding mechanism.

From the above analysis, it can be seen that existing routing strategies generally select paths based on the local traffic of the current and neighboring satellite nodes, and the scope of signaling interaction is limited. Although this reduces the signaling overhead, the lack of consideration of global state makes it difficult to ensure the global optimality of the routing strategy results. In addition, the cache length of ISLs is fixed and is not dynamically adjusted based on the traffic state.

3. System Modeling

The satellite-to-ground terminal communication process in an LEO satellite communication system is illustrated in Fig. 1. First, the satellite broadcasts to terminals within a certain area on the ground. The terminal initiates a registration request to the satellite using a public channel. After receiving a response from the satellite, the terminal feeds back the registration completion information through a dedicated channel. A communication link is then established through the application and response of the terminal's location information. Subsequently, based on a series of interactive information, the satellite and terminal select an appropriate transmission mode for information transmission and release the communication link after the transmission is completed.

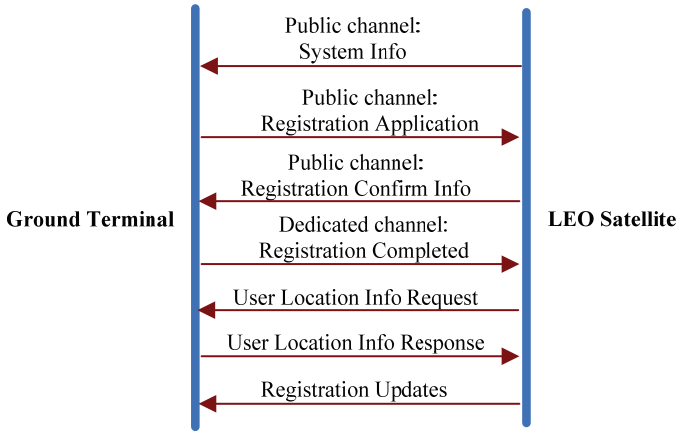

Fig. 1. LEO satellite communication process.

Significant differences exist between LEO satellite mobile and terrestrial cellular communication systems. Handover in the ground cellular communication system is caused by a user moving out of a location zone. In an LEO satellite mobile communication system, handover is not only caused by the user's movement but also by the high-speed movement of the satellite itself. Specifically, the high-speed movement of the LEO satellite and certain regions of the Earth are the main factors affecting handover.

Therefore, users will frequently initiate handovers during the communication process. The LEO

satellite mobile communication system performs interstellar routing, which is jointly completed by ground stations and satellites. Designing routing algorithms, reducing interaction signaling between the ground and satellites, and reducing communication delays are the keys to interstellar link design; the call access control function is mainly completed on satellites. Due to the limited processing capacity of onboard equipment, strategies related to call access control must be simple and practical.

Fig. 2 shows a system diagram of LEO satellite-to-ground terminal communication. Due to the proximity of LEO satellites to the ground, the Earth cannot be treated as a point.

The communication distance d between the satellite and ground terminal varies with the motion of the satellite and is calculated as follows [21]:

$$d = \sqrt{R_e^2 \sin\alpha + 2H_s R_e + H_s^2} - R_e \sin\alpha, \tag{1}$$

where R_e denotes radius of the Earth, H_s represents altitude of the satellite orbit, and α is the elevation angle between the satellite and ground terminal. It can be observed that the minimum value of d is achieved when α=90°, and the maximum value is achieved when α=0°. It should be noted that at very low values of α, due to the obstruction of tall buildings or mountains, line-of-sight propagation between the terminal and satellite is not possible, and multipath transmission dominates the transmission, further deteriorating the performance of the communication link.

In the communication process, the spatio-temporal correlation of the satellite network traffic is very important. We must update the output traffic of each inter-satellite link based on the actual traffic situation and perform reasonable cache allocation.

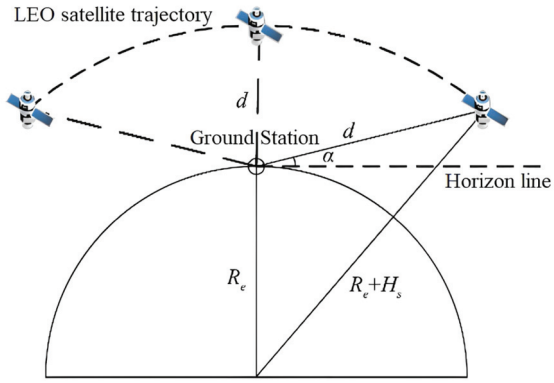

Fig. 2. LEO satellite communication system.

4. Dynamic Cache Allocation based on GBRT

The proposed method first determines the spatio-temporal correlation of the LEO satellite network traffic using the correlation coefficient to identify the correlated traffic, and then GBRT is employed to train and smooth the parameters and obtain relatively accurate prediction values through testing. Finally, the predicted traffic of each ISL is updated based on the actual traffic conditions, and cache allocation is adjusted based on the updated traffic ratio.

4.1 Spatio-temporal Correlation Traffic

The spatio-temporal correlation of the satellite network traffic is analyzed using the time-delayed Pearson correlation coefficient [22]. Assuming that the traffic load for the *i*-th satellite is represented as a time series, we have the following:

$$X_i = \{x_i(t-n+1), x_i(t-n+2), \dots, x_i(t-1), x_i(t)\}, \tag{2}$$

where $x_i(t)$ denotes the latest traffic observation value for the *i*-th satellite. The time-delayed Pearson correlation between the two traffic time series X_i and X_j with a delay of d_l is defined as follows:

$$\rho_{ij}(d_l) = \frac{\sum_{m=t-n+1}^{t}[x_i(m) - x_i'][x_j(m-d_l) - x_j']}{\sqrt{\sum_{m=t-n+1}^{t}[x_i(m) - x_i']^2 \sum_{m=t-n+1}^{t}[x_j(m-d_l) - x_j']^2}}, \tag{3}$$

where $\rho_{ij}(d_l)$ is the Pearson correlation coefficient for X_j in advance of X_i by d_l time units, which measures the degree of correlation between the current traffic of the *i*-th satellite and the historical traffic of the *j*-th satellite. n is the length of the traffic time series used for comparison, and x_i' and x_j' are the means of X_i and X_j, respectively. When $\rho_{ij}(d_l)$ approaches +1 or -1, there is a strong spatio-temporal correlation between the two time series, whereas when $\rho_{ij}(d_l)$ is approximately 0, there is no spatio-temporal correlation between the two time series.

Using spatio-temporally-related traffic as predictive input eliminates subjectivity and randomness in the selection of input variables while simultaneously providing a more accurate description of traffic patterns, eliminating unnecessary interference, and ultimately enhancing training speed and prediction accuracy.

4.2 Traffic Prediction based on GBRT

The concept of gradient boosting originated from Bentejac et al. [23] and can be regarded as an optimization algorithm based on an error function. GBRT is a machine learning technique for solving classification and regression problems, generating a strong predictive model in the form of an ensemble of weak predictive models, such as decision trees. The core of GBRT lies in computing a basic model in each iteration, with the next computation aimed at reducing the residual of the previous model and building a new model in the direction of the residual reduction gradient. Therefore, by continuously adjusting and optimizing the weights of weak classifiers to form a strong classifier, the loss function can be minimized and optimized. To improve the prediction accuracy of the LEO network, a traffic prediction model was designed based on GBRT. Fig. 3 shows a flowchart of the prediction model.

Each basic model is a regression tree (RT) used to correct the residual generated in the previous iteration of calculation. The residual is the difference between the predicted and observed values; the smaller its value, the better the predictive performance of the model. To improve the performance of the model and reduce the residual value, improvements are made in two aspects: the loss function and input variables. The Huber loss function is introduced to reduce the residual loss of the samples. This is a loss function for robust regression [24], where variable a is expressed as a residual $a = y - f(x)$ as follows:

$$L(y, f(x)) = \begin{cases} \frac{1}{2}(y, f(x))^2, & |y, f(x)| \le \delta \\ \delta\,|y, f(x)| - \frac{1}{2}(\delta)^2, & |y, f(x)| > \delta \end{cases} \tag{4}$$

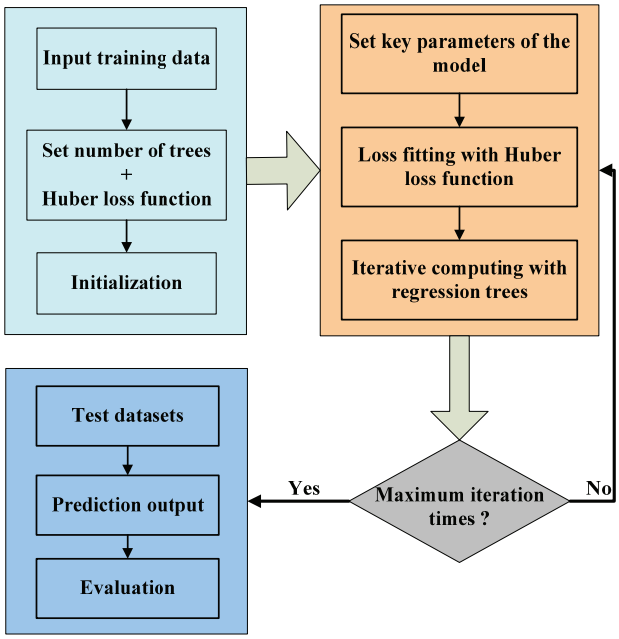

Fig. 3. Flowchart of the prediction model.

The spatio-temporally correlated traffic in the LEO system exhibits temporal and spatial characteristics. In the determined area, traffic with a high correlation coefficient with the target satellite is selected as the independent variable, and the traffic of the target satellite is selected as the dependent variable to establish a GBRT-based prediction model with spatio-temporally correlated traffic as the input variable.

Consider a training dataset $(x, y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x is the input variable and y is the corresponding output variable. The leaves of each RT are denoted by J_m, so the input space is divided into J_m disjoint regions $(R_{1m}, R_{2m}, \ldots, R_{jm})$, which can be expressed as follows:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} I(x), \tag{5}$$

where b_{jm} is a constant value for region b_{jm}, and $I(x)$ is an indicator function that takes the value 1 if $x \in R_{jm}$ and 0 otherwise.

The objective function of the prediction model is defined as follows:

$$\tilde{F}(x) = \arg\min_{f(x)} \sum_{i=1}^{N} L\big(y, f(x)\big), \tag{6}$$

where $\tilde{F}(\mathrm{x})$ minimizes the expected value of a given loss function $L\big(y, f(x)\big)$ over the dataset.

Let the number of trees be M. The Huber loss function is introduced for fitting, and the model is initialized as a constant function:

$$f_0 = \arg\min_{f(x)} \sum_{i=1}^{N} L_\delta(y, f(x)). \tag{7}$$

The model iterates in the direction of gradient descent to reduce the residuals. The GBRT model fits the residual as an RT $h_m(x)$. To correct the residual of the previous RT, a new RT is added at each step, and the model is updated as follows:

$$f_m = f_{m-1}(x) + \rho_m h_m(x), \tag{8}$$

where ρ_m denotes the step size of gradient descent. The GBRT constructs a new model in the direction of gradient descent and continuously updates the model by minimizing the expected value of the loss function so that the model eventually stabilizes.

The model has three key parameters: number of trees M, learning efficiency μ, and depth of trees D. In general, as the number of trees increases, training errors can be reduced. However, if the number of trees is too large, the generalization ability of the model will be significantly weakened, and the predictive performance of the model will decrease due to overfitting. Therefore, it is necessary to optimize the number of model trees to minimize prediction errors; overfitting can also be avoided by controlling the number of trees. In addition, the gradient boosting algorithm introduces a shrinkage factor, which gradually approaches the optimal result in small steps to avoid overfitting caused by approaching the result too quickly. The model is represented as follows:

$$f_m = f_{m-1}(x) + \mu * \rho_m h_m(x), \tag{9}$$

where the parameter $0 < \mu \leq 1$ is the weight reduction coefficient of each weak classifier. GBRT with a low learning rate can effectively improve the generalization ability of the model; however, a very low learning rate requires more trees to make the model converge, which may reduce the prediction accuracy. The depth of the trees, i.e., the complexity of the basic model trees, is also a factor that affects prediction accuracy. The proposed model improves the prediction accuracy by continuously optimizing the combination of M, μ, and D.

4.3 Cache Allocation based on Traffic Prediction

The basic idea of the DCAR strategy is to perform cache allocation based on the predicted traffic results during the network initialization phase according to the traffic ratio of each satellite link. During the system operation phase, the output traffic of each satellite link is updated according to the actual traffic situation, and the cache allocation is adjusted based on the updated traffic ratio.

4.3.1 Initialization

During the initialization phase, each satellite node is in a light-load state and congestion will not occur. Therefore, each satellite node calculates the routing path based on the traffic prediction results when forwarding data packets. Taking satellite node V_i as an example, the business traffic from V_i to other nodes V_j is composed of two parts: ground node traffic and traffic from neighboring satellite nodes flowing in, represented by J_{ij}. Therefore, the traffic J_{ij} from V_i to all other V_j can be obtained, and this traffic will be forwarded to neighboring nodes through four satellite links of V_i, with a forwarding traffic

size of J_{ij}^k, where $k = 1,2$ represent the upper and lower ISLs in the same orbit plane, respectively, and $k = 3,4$ represent the left and right ISLs between different orbit planes, respectively. After obtaining J_{ij}^k, the total traffic J_i^k on the four satellite links of V_i can be represented as follows:

$$J_i^k = \sum_{j=1, j\neq 1}^{MN} J_{ij}^k. \tag{10}$$

If the total cache resources of V_i are defined as Q, then the cache resources allocated to the four satellite links Q_k can be represented as follows:

$$Q_k = Q\frac{J_i^k}{\sum_{k=1}^4 J_i^k}. \tag{11}$$

It can be seen that $\sum Q_k = Q$, which means that the total cache resources allocated to the four satellite links Q_k sum up to the total cache resources Q.

4.3.2 Dynamic cache allocation

During the system operation phase, V_i periodically monitors the traffic output of the four satellite links, with a monitoring cycle of t_δ. Assuming that in the last monitoring cycle, i.e., $(t - t_\delta, t)$, the output traffic of the ISL is $\hat{J}_i^k$ and the average output traffic in the time interval $(0, t - t_\delta)$, is $\hat{J}_i^k$, then the output traffic can be updated as follows:

$$\hat{J}_i^k = (1 - \lambda)\hat{J}_i^k + \lambda\hat{J}_i^k, \tag{12}$$

where $0 < \lambda < 1$ denotes the forgetting factor. The output traffic should be a long-term average value, and a sharp drop should be avoided. Therefore, when λ is too small, the change in output traffic will be very slow, making it difficult to represent the long-term average value and affecting the efficiency of cache resource allocation. When λ is too large, the change in output traffic will follow short-term traffic fluctuations, making it difficult to achieve accurate traffic prediction.

In addition to traffic monitoring, V_i periodically monitors the queue usage of the four satellite links, with a monitoring cycle of t_δ. The cache queue occupancy rate of each satellite link obtained in each monitoring cycle is defined as Q_k^r, and two threshold values are set: Q_k^a and Q_k^b. When $Q_k^r \leq Q_k^a$, the state of the satellite link is defined as idle. When $Q_k^a \leq Q_k^r \leq Q_k^b$, the state of the satellite link is defined as normal. When $Q_k^r > Q_k^b$, the state of the satellite link is defined as congested. To avoid packet loss in the cache queue, the value of Q_k^b must satisfy the following:

$$Q_k^b Q_k + \frac{(\hat{J}_i^k - B)(t_\delta - t_d)}{\tilde{P}} \leq Q_k, \tag{13}$$

where B is the capacity of the ISL, t_d is the one-hop propagation delay of the ISL, and $\tilde{P}$ is the data packet size. Therefore, Q_k^b can be obtained as follows:

$$Q_k^b = 1 - \frac{(\hat{J}_i^k - B)(t_\delta - t_d)}{Q_k\tilde{P}}. \tag{14}$$

The value of Q_k^b should ensure sufficient time to recalculate the routing path before reentering the congested state. If the time consumed to calculate the routing path is t_{RP}, then the value of Q_k^b should satisfy the following:

$$Q_k^b Q_k + \frac{(\hat{J}_i^k - B)t_{RP}}{\tilde{P}} \leq Q_k^b Q_k . \tag{15}$$

It can be seen that Q_k^a and Q_k^b can be calculated once $\hat{J}_i^k$ is determined. In the proposed method, the satellite node V_i dynamically adjusts the cache allocation based on the status of the four cache queues in each monitoring cycle t_δ. Taking link k as an example, if its cache queue occupancy rate in the current monitoring cycle is Q_k^r and the total length of the queue is Q_k, then if the link is in the idle state, the length of its cache queue will be adjusted to $\tilde{Q}_k = Q_k - (Q_k - Q_k^r Q_k)/2$, and the cache resources released by the link will be $Q_k' = Q_k - \tilde{Q}_k = (Q_k - Q_k^r Q_k)/2$; if the link is in the normal or congested state, the released cache resources will be free, i.e., $Q_k' = 0$. By traversing all satellite links, the total cache resources released are $Q_{total} = \sum Q_k'$, and V_i allocates Q_{total} according to the cache resource demand ratio for each congested link. Thus, a dynamic cache allocation strategy is realized based on traffic prediction.

5. Experiment and Analysis

An Iridium satellite constellation was constructed using STK, as shown in Fig. 4, and the proposed routing strategy was simulated using the OPNET simulation software [25]. OPNET has rich statistical parameter collection and analysis functions. It can directly collect commonly used statistical performance parameters at each network level and has a variety of statistical parameter collection and processing methods. Special network parameters can be collected through underlying network programming, and simulation reports can be prepared and output. The simulation adopted the network structure of the Iridium system, with an LEO height of 670 km, 6 orbit planes, and 11 satellites per orbit plane. Each satellite established ISLs with two neighboring satellites in the same orbit plane (upper and lower) and two neighboring satellites in adjacent orbit planes (left and right), except for the satellites on the opposite sides of the orbit planes. The capacity of the ISLs was set to 50 Mbit/s, and that of the satellite-to-ground links was set to 100 Mbit/s. The one-hop propagation delay t_d of the ISLs was set to 20 ms. In the satellite network, data were transmitted based on IP, and the data packet size was set to 10 KB. The monitoring period t_δ was set to 30 ms, and the calculation time t_{RP} for routing strategy was set to 5 ms. The total cache queue length Q of the satellite was set to 600 data packets. The simulation used the orbit agent strategy [26], and each agent collected the link status and queuing delay information every 50 seconds, i.e., each satellite node refreshed its routing table every 50 seconds. A total of 165 terminals were used in the simulation. Each terminal's traffic flow was set to follow an on-off flow with a Pareto distribution, with a shape parameter of 1.2 and an average activation and waiting time of 200 ms. The simulation lasted for 48 hours and was repeated 10 times, with the average of the 10 runs taken as the final simulation result.

5.1 Evaluation Metrics

Mean absolute percentage error (MAPE) and mean absolute error (MAE) were used to evaluate the performance of the traffic prediction model, with smaller values indicating more accurate predictions:

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y(i)-y'(i)}{y(i)}\right|, \tag{16}$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y(i)-y'(i)|, \tag{17}$$

where $y(i)$ denotes the actual traffic value, $y'(i)$ is the predicted traffic value, and m is the number of test samples.

In addition, we compared the performances of different routing strategies in terms of packet loss rate and average end-to-end delay.

5.2 Results of Traffic Prediction

At time t, the LEO_2_3 satellite was selected as the target satellite X_1 for traffic prediction. The load of X_1 at time t is denoted as $X_1(t)$. By observing the time series of X_1's traffic load during this period, we found that the traffic was relatively high during the daytime and low in the early morning, indicating a periodic variation in satellite network traffic. We selected the area covered by the nine Iridium satellites as the study area. In this area, if the angle between the line connecting a satellite and the ground point directly below X_1 and the line connecting X_1 and its own ground point was less than or equal to 60°, we collected the satellite network traffic of that satellite over a certain period as the condition for analyzing the spatio-temporal correlation of the satellite network traffic. We first collected the traffic time series of X_1 in the last 400 minutes and obtained measurements every 60 minutes.

Let S be the satellite sequence number and C be the correlation coefficient with the target satellite. Table 1 lists the Pearson correlation coefficients between X_1 and the traffic of other satellites. As shown in Table 1, X_1 has a high spatio-temporal correlation with the traffic of these 32 groups of satellites (correlation coefficient greater than 0.80), but there are 14 groups with correlation coefficients above 0.85, so 0.85 was selected as the threshold. These 14 groups of spatio-temporally correlated traffic were used as input for predicting the traffic of X_1. The spatio-temporally correlated traffic reflects the cross-correlation and autocorrelation of the satellite network. Using spatio-temporally correlated traffic for prediction can eliminate interference from irrelevant traffic and improve the accuracy of the final results.

Table 2 provides a comparison of the prediction performance of different traffic prediction methods. The SVM-based method used in [6] cannot handle the temporal dynamics and nonlinearity present in LEO traffic data, resulting in poor performance when applied to LEO networks. The methods in [7] and [9] use deep learning models for traffic prediction, which are highly complex and dependent on hyperparameter selection. In addition, the limited amount of training data severely limits the performance of these methods. The proposed method considers the spatio-temporal correlation of satellite network traffic when predicting traffic load. By continuously adjusting the weights of the basic model, the proposed method achieved significantly improved prediction accuracy compared to traditional methods while reducing the training time.

Fig. 5 presents a comparison of the prediction errors of the proposed method and [9] for different numbers of testing samples. The proposed method continuously optimizes the weights of weak learners during the training process of the decision tree, turning them into strong learners and thereby improving the accuracy of the model's predictions. The prediction error of the proposed method was significantly lower than that of the method in [9], demonstrating the effectiveness of the proposed method.

Table 1. Person correlation coefficient between target satellite and other satellites

S	C	S	C	S	C	S	C
$X_1(t-9)$	0.937	$X_1(t-86)$	0.886	$X_1(t-315)$	0.842	$X_1(t-130)$	0.799
$X_1(t-13)$	0.927	$X_1(t-50)$	0.882	$X_1(t-128)$	0.833	$X_1(t-420)$	0.794
$X_1(t-10)$	0.905	$X_1(t-122)$	0.875	$X_1(t-189)$	0.831	$X_1(t-665)$	0.782
$X_1(t-15)$	0.897	$X_1(t-222)$	0.871	$X_1(t-255)$	0.828	$X_1(t-395)$	0.775
$X_1(t-1)$	0.894	$X_1(t-43)$	0.864	$X_1(t-113)$	0.819	$X_1(t-1285)$	0.771
$X_1(t-7)$	0.892	$X_1(t-27)$	0.859	$X_1(t-877)$	0.808	$X_1(t-785)$	0.753
$X_1(t-6)$	0.890	$X_1(t-74)$	0.855	$X_1(t-1275)$	0.801	$X_1(t-1430)$	0.749

Table 2. Traffic prediction comparison

Method	MAPE	MAE	Training time (sec)
Support vector machine	0.085	58.7	0.88
Recurrent neural network	0.038	35.8	125
Long short-term memory	0.031	32.7	275
Proposed method	0.017	22.8	1.25

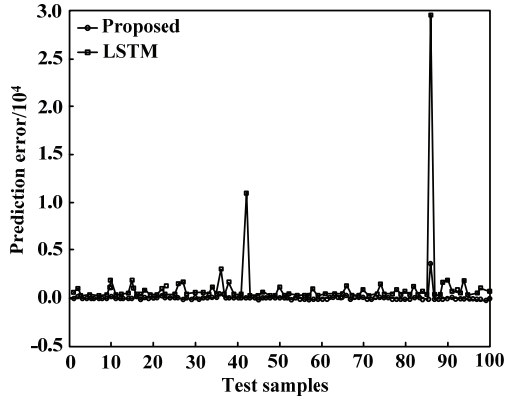

Fig. 5. Prediction error comparison.

5.3 Performance Analysis of Dynamic Cache Routing Strategy

Table 3 lists the packet loss rate performances of different strategies under various traffic loads. It can be seen from the table that the proposed DCAR strategy successfully reduced packet loss probability. The methods in [17] and [18] could not obtain an accurate state of the LEO network due to the introduction of additional signaling overhead and excessive traffic loads, and the performance deteriorated rapidly as the traffic load increased. The methods in [19] and [20] achieved significant performance improvements through load balancing and node congestion processing. By contrast, the proposed strategy can adjust the buffer queue length according to the traffic load ratio so that the ISL can provide more space to cache data packets under a high traffic load. From the results in Table 3, when the total traffic flow is high, the packet loss rate increases significantly. This is because the data packets that must be transmitted in the link exceed the capacity of the ISL, resulting in congestion at multiple nodes in the network.

Table 3. Packet loss rate comparison (unit: %)

Method	Total traffic load / (Tbit · day)$^{-1}$					
	1.8	**2.2**	**2.6**	**3.0**	**3.4**	**3.8**
LAOR	1.4	7.5	12.8	26.7	38.5	41.3
ALBR	1.2	6.3	11.7	24.3	36.1	39.7
SALB	0.9	3.4	6.5	17.8	25.6	30.5
SDLBR	0.7	2.9	5.8	15.4	26.3	31.0
Proposed DCAR	0.4	0.8	2.9	9.3	13.7	15.6

Table 4 lists the average end-to-end delays of different strategies under various total traffic loads. It can be seen that the proposed DCAR strategy improves the average end-to-end delay of data packets. The core idea of [19] and [20], and the proposed strategy is to perform rerouting of data packets when congestion occurs, i.e., to divert data packets to paths that do not include congested nodes for transmission. However, the advantage of the proposed DCAR strategy is that it provides sufficient cache space through cache queue allocation when the delay of rerouting data packets exceeds the queuing delay and avoiding the serious deterioration of queuing delay by changing the routing path when the delay of rerouting data packets is less than the queuing delay. Therefore, although the end-to-end delay increases with increasing traffic load, the proposed DCAR strategy shows an optimal average end-to-end delay performance.

Table 4. Average end-to-end delay comparison (unit: ms)

Method	Total traffic load / (Tbit · day)$^{-1}$					
	1.8	**2.2**	**2.6**	**3.0**	**3.4**	**3.8**
LAOR	125	143	162	181	195	215
ALBR	121	139	154	168	182	199
SALB	117	128	147	155	179	196
SDLBR	105	121	142	151	174	192
Proposed DCAR	98	115	138	146	171	182

As shown in Fig. 6, we compared the total system throughput under different numbers of terminal devices using the proposed method and methods in [19] and [20]. Fig. 6 shows the number of terminal devices on the horizontal axis and system throughput on the vertical axis. The system throughput decreased as the number of terminal devices increased. The system throughputs obtained by the two methods in [19] and [20] are very close because they have the same total resources and each terminal device occupies independent resources, resulting in small differences in the final results. The proposed method achieved a higher system throughput, demonstrating the effectiveness of using a gradient boosting tree algorithm for traffic prediction and dynamically adjusting cache queues based on the prediction results.

Fig. 7 shows the business distribution coefficient (BDC) performance of the three strategies under different total traffic amounts. A higher BDC indicates a more uniform traffic distribution over the entire system. It can be seen from Fig. 6 that the traffic distribution of the proposed dynamic cache allocation strategy is initially slightly better than that of [19] and [20]; however, with an increase in traffic load, the improvement ratio of the business distribution increases. This is because when the load is high, the proposed strategy can better balance the queuing and switching of data packets to obtain a higher BDC.

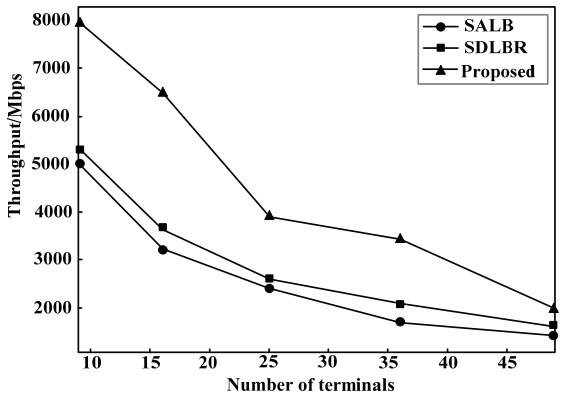

Fig. 6. System throughput comparison.

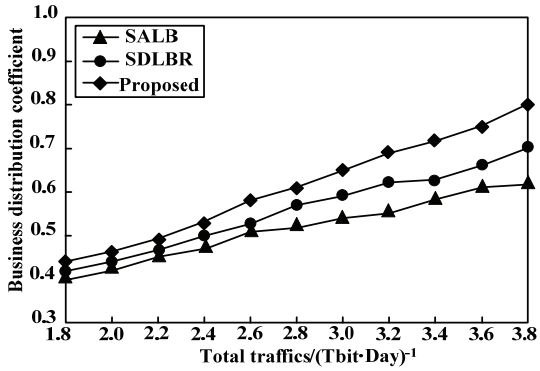

Fig. 7. Business distribution coefficient.

Now, we discuss practical applications. LEO satellites orbit approximately 750–1,500 km above Earth's surface, which is just above Earth's atmosphere and below the inner Van's radiation belt. LEO satellite networks can reuse wireless spectra more efficiently, thereby increasing the communication bandwidth. The round-trip delay between satellite and ground nodes is approximately 15 ms, which meets the network delay requirements of most real-time applications. At the same time, the terminal size of the ground user node is equivalent to that of an ordinary handheld terminal and consumes less power. For user nodes, the same LEO satellite can provide services for approximately ten minutes at most. The most notable feature of LEO is the provision of voice and low-speed data services with transmission rates of 2.4 k and 7.2 k, respectively.

6. Conclusion

To address the data packet routing problem in LEO satellite systems, the spatio-temporal characteristics of the traffic load of each satellite node in an LEO system were analyzed, and a traffic prediction method based on the GBRT algorithm was proposed. Based on this prediction method, a DCAR strategy based on actual traffic loads was proposed, dividing the system into two stages: initialization and system operation. The cache resources occupied by the four ISLs of the satellite nodes are dynamically allocated

based on the traffic prediction results. Simulation results show that the proposed method can improve the packet loss rate and end-to-end delay and has certain theoretical value and application prospects in LEO satellite systems.

The round-trip delay between medium-orbit (MEO) satellites and the ground is long. Compared with the round-trip delay between ground user nodes and LEO satellites, which is only several milliseconds, the round-trip delay from ground user nodes to MEO satellite nodes is approximately 100 ms, even without forwarding by LEO satellites. Therefore, the round-trip delay caused by MEO satellites is greater than that caused by LEO satellites. Similar to the traditional land-based public mobile communication network, only before the communication process between user nodes is established, the network needs to obtain the current location and access information of the target user node through the mobility management system. From the perspective of routing strategies, the dynamic route caching strategy used in this study is also applicable to MEO satellites.

Data Availability

The data used to support the findings of this study are included within the article.

Acknowledgement

This research was funded by the National Natural Science Foundation of China (No. 6156010183), Guizhou Province Education Department Projects of China (KY[2017]031 and KY[2020]007).

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. F. M. Montoya, et al., “Satellite communications in the new space era: a survey and future challenges,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 70-109, 2021. https://doi.org/10.1109/COMST.2020.3028247

[2] M. Hoyhtya, S. Boumard, A. Yastrebova, P. Jarvensivu, M. Kiviranta, and A. Anttonen, “Sustainable satellite communications in the 6G era: a European view for multi-layer systems and space safety. *IEEE Access*, vol. 10, pp. 99973-100005, 2022. https://doi.org/10.1109/ACCESS.2022.3206862

[3] P. Wang, J. Zhang, X. Zhang, Z. Yan, B. G. Evans, and W. Wang, “Convergence of satellite and terrestrial networks: a comprehensive survey,” *IEEE Access*, vol. 8, pp. 5550-5588, 2019. https://doi.org/10.1109/ACCESS.2019.2963223

[4] S. K. Routray, R. Tengshe, A. Javali, S. Sarkar, L. Sharma, and A. D. Ghosh, “Satellite based IoT for mission critical applications,” in *Proceedings of 2019 International Conference on Data Science and Communication (IconDSC)*, Bangalore, India, 2019, pp. 1-6. https://doi.org/10.1109/IconDSC.2019.8817030

[5] V. S. Chippalkatti, "Review of satellite based Internet of Things and applications," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 12, pp. 758-766, 2021.

[6] Z. Tian and S. Li, "A network traffic prediction method based on IFS algorithm optimised LSSVM," *International Journal of Engineering Systems Modelling and Simulation*, vol. 9, no. 4, pp. 200-213, 2017. https://doi.org/10.1504/IJESMS.2017.087553

[7] N. Ramakrishnan and T. Soni, "Network traffic prediction using recurrent neural networks," in *Proceedings of 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2018, pp. 187-193. https://doi.org/10.1109/ICMLA.2018.00035

[8] L. Nie, Z. Ning, M. S. Obaidat, B. Sadoun, H. Wang, S. Li, L. Guo, and G. Wang, "A reinforcement learning-based network traffic prediction mechanism in intelligent Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2169-2180, 2021. https://doi.org/10.1109/TII.2020.3004232

[9] S. Nihale, S. Sharma, L. Parashar, and U. Singh, "Network traffic prediction using long short-term memory," in *Proceedings of 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 338-343. https://doi.org/10.1109/ICESC48915.2020.9156045

[10] X. Cao, Y. Li, X. Xiong, and J. Wang, "Dynamic routings in satellite networks: an overview," *Sensors*, vol. 22, no. 12, article no. 4552, 2022. https://doi.org/10.3390/s22124552

[11] H. Wu, J. Li, H. Lu, and P. Hong, "A two-layer caching model for content delivery services in satellite-terrestrial networks," in *Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, USA, 2016, pp. 1-6. https://doi.org/10.1109/GLOCOM.2016.7841557

[12] S. Liu, X. Hu, Y. Wang, G. Cui, and W. Wang, "Distributed caching based on matching game in LEO satellite constellation networks," *IEEE Communications Letters*, vol. 22, no. 2, pp. 300-303, 2018. https://doi.org/10.1109/LCOMM.2017.2771434

[13] G. Zhong, J. Yan, and L. Kuang, "QoE-driven social aware caching placement for terrestrial-satellite networks," *China Communications*, vol. 15, no. 10, pp. 60-72, 2018. https://doi.org/10.1109/CC.2018.8485469

[14] B. Soret and D. Smith, "Autonomous routing for LEO satellite constellations with minimum use of inter-plane links," in *Proceedings of 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1-6. https://doi.org/10.1109/ICC.2019.8761787

[15] Y. Huang, S. Wu, Z. Kang, Z. Mu, H. Huang, X. Wu, A. J. Tang, and X. Cheng, "Reinforcement learning based dynamic distributed routing scheme for mega LEO satellite networks," *Chinese Journal of Aeronautics*, vol. 36, no. 2, pp. 284-291, 2023. https://doi.org/10.1016/j.cja.2022.06.021

[16] I. Del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta astronautica*, vol. 159, pp. 123-135, 2019. https://doi.org/10.1016/j.actaastro.2019.03.040

[17] S. Karapantazis, E. Papapetrou, and F. N. Pavlidou, "On-demand routing in LEO satellite systems," in *Proceedings of 2007 IEEE International Conference on Communications*, Glasgow, UK, 2017, pp. 26-31. https://doi.org/10.1109/ICC.2007.14

[18] Z. Wu, G. Hu, F. Jin, B. Jiang, and Y. Fu, "Agent-based dynamic routing in the packet-switched LEO satellite networks," in *Proceedings of 2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, Nanjing, China, 2015, pp. 1-6. https://doi.org/10.1109/WCSP.2015.7341005

[19] X. Li, F. Tang, L. Chen, and J. Li, "A state-aware and load-balanced routing model for LEO satellite networks," in *Proceedings of 2017 IEEE Global Communications Conference (GLOBECOM)*, Singapore, 2017, pp. 1-6. https://doi.org/10.1109/GLOCOM.2017.8254443

[20] N. Li, X. H. Zhao, and K. Yao, "Semi-distributed load balancing routing algorithm based on LEO satellite networks," in *Proceedings of SPIE 11848: International Conference on Signal Image Processing and Communication (ICSIPC 2021).* Bellingham, WA: International Society for Optics and Photonics, 2021, pp. 390-397. https://doi.org/10.1117/12.2600123

[21] Y. Zhang, Q. Wu, Z. Lai, and H. Li, "Enabling low-latency-capable satellite-ground topology for emerging LEO satellite networks," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, London, UK, 2022, pp. 1329-1338. https://doi.org/10.1109/INFOCOM48880.2022.9796886

[22] B. Du, F. Liu, X. Sun, R. Song, and L. Wang, "A prediction method of LEO satellite orbit control effect based on multiple regression analysis model," in *Proceedings of 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, Nanjing, China, 2021, pp. 1-6. https://doi.org/10.1109/PHM-Nanjing52125.2021.9612824

[23] C. Bentejac, A. Csorgo, and G. Martinez-Munoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937-1967, 2021. https://doi.org/10.1007/s10462-020-09896-5

[24] S. Balasundaram and S. C. Prasad, "Robust twin support vector regression based on Huber loss function," *Neural Computing and Applications*, vol. 32, pp. 11285-11309, 2020. https://doi.org/10.1007/s00521-019-04625-8

[25] B. Feng, Y. Huang, A. Tian, H. Wang, H. Zhou, S. Yu, and H. Zhang, "DR-SDSN: an elastic differentiated routing framework for software-defined satellite networks," *IEEE Wireless Communications*, vol. 29, no. 6, pp. 80-86, 2022. https://doi.org/10.1109/MWC.011.2100578

[26] P. Zuo, C. Wang, Z. Yao, S. Hou, and H. Jiang, "An intelligent routing algorithm for LEO satellites based on deep reinforcement learning," in *Proceedings of 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Norman, OK, USA, 2021, pp. 1-5. https://doi.org/10.1109/VTC2021-Fall52928.2021.9625325

Yang Yang **https://orcid.org/0000-0002-4998-0260**

He received the B.Sc. degree in Electrical Engineering and Automation from Guizhou Normal University, Guiyang, China in 2020. He is currently working toward the master's degree in the Department of Computer Science and Technology, Guizhou Normal University, Guiyang, China. His research interests include space-based internet of things, information centric networks, machine learning and satellite communication.

Shengbo Hu **https://orcid.org/0000-0002-7891-2451**

He received the B.Sc. degree in communication engineering from Southeast University, Nanjing, China in 1985, the M.Sc. degree in communication engineering from National University of Defense Technology, Changsha, China in 1992 and the Ph.D. degree in communication engineering from Chongqing University, Chongqing, China in 2006. Currently, he is a professor and the dean with the School of Big Data and Computer Science, Guizhou Normal University, China. He has been funded as the member of the Academic Committee of the Key Laboratory of Electronic and Information Technology for Complex Space Systems, CAS.

Guiju Lu **https://orcid.org/0009-0008-3370-2822**

She received the B.Sc. degree in Electrical Engineering and Automation from Guizhou Normal University, Guiyang, China in 2021. She is currently working toward the master's degree in the Department of vocational and technical education, Guizhou Normal University, Guiyang, China. Her research interests include artificial intelligence, machine learning, Internet and edge computing.