


RESEARCH

Open Access



Research of 5G HUDN network selection algorithm based on Dueling-DDQN

Jianli Xie^{1*} , Binhan Zhu¹ and Cuiran Li¹

*Correspondence:
xiejl@mail.lzjtu.cn

¹ School of Electronic
and Information Engineering,
Lanzhou Jiaotong University,
Lanzhou 730070, China

Abstract

Due to the dense deployment and the diversity of user service types in the 5G HUDN environment, a more flexible network selection algorithm is required to reduce the network blocking rate and improve the user's quality of service (QoS). Considering the QoS requirements and preferences of the users, a network selection algorithm based on Dueling-DDQN is proposed by using deep reinforcement learning. Firstly, the state, action space and reward function of the user-selected network are designed. Then, by calculating the network selection benefits for different types of services initiated by users, the analytic hierarchy process is used to establish the weight relationship between the different user services and the network attributes. Finally, a deep Q neural network is used to solve and optimize the proposed target and obtain the user's best network selection strategy and long-term network selection benefits. The simulation results show that compared with other algorithms, the proposed algorithm can effectively reduce the network blocking rate while reducing the switching times.

Keywords: 5G heterogeneous ultra-dense network, Deep reinforcement learning, QoS, Network selection, Dueling-DDQN

1 Methods/experimental

In order to solve the network selection problem in 5G heterogeneous ultra-dense network (HUDN) environment, a network selection algorithm based on Dueling Double Deep Q Network (DDQN) is proposed. The deep reinforcement learning (DRL) method was introduced to model and execute the algorithm, and its applicability was verified in the coexistence environment of multiple networks such as wireless local area network (WLAN), long time evolution (LTE), and 5G. We designed the state, action space, and reward function of the user selection network. We calculated the network selection benefits for different types of services initiated by the user, and established the weight relationship between the user's different services and network attributes using analytic hierarchy process (AHP). Finally, we used a deep Q neural network to solve and optimize the proposed goal, obtaining the user's optimal network selection strategy and long-term network selection benefits.

Compared with other intelligent algorithms, the proposed network selection algorithm effectively reduces the number of switches and improves the efficiency of network resource utilization.

2 Introduction

With the rapid development of wireless communication technology and the large-scale popularity of mobile intelligent terminals, a single network cannot meet the needs of users. The future wireless network will be a heterogeneous network composed of different technologies and types [1–3]. As one of the key technologies of 5G communication, heterogeneous ultra-dense network technology was explicitly included in the “5G Concept White Paper” released by IMT-2020 [4–6]. In the HUDN environment, the network topology becomes more complex, and the service types of users develop in a more diversified trend. To ensure the QoS of the user services, it is necessary to research more efficient and reasonable network selection algorithms for different business requirements of users. The network selection in a heterogeneous network can be divided into two types. The first is horizontal handover, which occurs between the same network type, and the second is vertical handover, which occurs between different network types [7]. As an important part of vertical handover, network selection is difficult to implement in the complex network topology of HUDN [8–10]. Inaccurate and inefficient selection algorithms will cause the selected network to fail to meet user needs and users to hand over frequently. Therefore, the study of more efficient and reasonable network selection algorithms has become a hot issue in the current research field of heterogeneous networks.

Based on the above research needs, the research motivation of this paper is to study a network selection algorithm suitable for HUDN. Therefore, a network selection algorithm is proposed based on Dueling-DDQN, which successfully takes into account several factors. The proposed algorithm not only considers the different service attributes of the user with different network preferences, also uses a comprehensive utility function to calculate the user’s instant profits of the network selection, effectively distinguishes the different business types of users, improves the total rewards of the user terminal. Compared with the existing methods, the proposed algorithm can reduce blocking probability and improve the utilization of network resources.

The rest of this paper is presented as follows: Sect. 3 presents related work on network selection algorithms. Section 4 introduces the relevant system model. Section 5 describes the structure and content of the proposed network selection algorithm in detail. Section 6 describes the simulation experiments and results analysis in detail. Section 7 concludes this paper, pointing out the inadequacies and directions for future research.

3 Related work

At present, the research on network selection has achieved some results that can be divided into the following categories:

- (a) Network selection algorithm based on a single indicator: Ahujia et al. [11] proposed that the signal strength received by the mobile terminal is used as the only indicator for network selection, and the network with the highest signal strength is selected to be accessed by the user. Zhao et al. [12] proposed a network selection algorithm based on the predicted signal strength, and the signaling process of network handover was designed to reduce the handover delay. In [13], the network delay was used

as the criterion for the user to select the network, and the network with the lowest delay was selected for access by the user.

- (b) Network selection algorithm based on multi-attribute decision-making (MADM): This type of network selection algorithm considers multiple attribute indicators on the network side in the design stage and integrates various factors to select the network for users to access. In [14], a network selection algorithm based on the improved Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) was proposed. This solves the problem of abnormal network ranking and selects the best network for users to access. Yu et al. [15] proposed an MADM network selection algorithm based on the combination of the fuzzy analytical hierarchy process (FAHP), entropy-weight-method (EWM) and TOPSIS to select the network with the largest comprehensive utility value for users to access. Gaur et al. [16] proposed a specific threshold-based MADM network selection algorithm, which selects the best network access for users while reducing unnecessary handovers by setting QoS thresholds for user-specific applications. In [17], a network selection algorithm for dynamic weight optimization was proposed that selects the network access with the lowest computational cost for users by constructing a cost function.
- (c) Network selection algorithm based on fuzzy logic (FL): fuzzy logic was adopted in [18] to select the best access network between 3G and Wi-Fi, and the input linguistic variable of this fuzzy logic model is Wi-Fi and 3G network signal strength and network load. A fuzzy decision-making system was proposed [19]. The relevant parameters of the switching decision are input into the fuzzy decision-making system. After quantifying the decision parameters, the set decision rules can effectively reduce the ping-pong handover.
- (d) Network selection algorithm based on utility function: Due to the diversity of user services, different users have different satisfaction with the same network attribute parameters. Therefore, some algorithms use a utility function to measure the user's satisfaction with the currently selected network parameters [20, 21]. This type of algorithm converts specific network parameter values into utility values and then selects the network with the highest comprehensive utility value through the set calculation rules.

The above related research mainly selects the best network for users based on the current state of network attributes. Although the network selection algorithm based on a single index is relatively simple to implement, few factors are considered, and without considering other network attributes, it is difficult to meet user's QoS requirements such as bandwidth and packet loss of different services. Especially when the heterogeneous network environment is complex, the signal strength of each network is not clearly distinguished, which may easily lead to a ping-pong handover effect. Although the algorithm based on MADM and utility function selects the network for the user with the highest network score and comprehensive utility as the goal, in the HUDN environment, the network is densely covered, and the network dynamics are further enhanced. This is prone to problems such as insignificant differences in network scores and user ping-pong handovers. As the number of users increases, a more serious load occurs in the

current optimal network, and it is difficult to guarantee the QoS of users. Although an algorithm based on fuzzy logic can deal with inaccuracy, nonlinearity and other problems, the design of the membership function requires certain experience. With an increase in the number of switching decision parameters, the number of fuzzy rules also increases, and the complexity becomes higher. These algorithms do not consider the impact on mobile users due to changes in network attributes after selecting a network from a long-term perspective. Therefore, some network selection algorithms based on users' long-term interests have been proposed.

- (e) Network selection algorithm based on Markov Decision Proposed (MDP): Xie et al. [22] proposed a network selection algorithm based on MDP and used the GA-SA algorithm to solve the proposed model, enabling users to access the network with the best long-term benefits. Khodmi et al. [23] constructed an MDP-based decision-making model to solve the problem of handover decisions in which a Steinberg competition model was introduced to balance the load of the macro-base station and the relay base station. According to the constructed model, the reward function for the user to switch the network is given, and the user selects the base station with the highest reward to access. Yang et al. [24] analyzed the different requirements of real-time and nonreal-time business for network-side attributes, designed a reward function according to the maximum and minimum values of network-side attributes, and proposed a network selection algorithm based on MDP, which reduced the network blocking rate.
- (f) Network selection algorithm based on reinforcement learning (RL): Although a network selection algorithm based on MDP can maximize the benefits of mobile users in the long term, it still has some defects. For example, the state transition probability cannot be known in advance in some cases, and the algorithm does not easily converge due to the large state space in complex problems. He et al. [25] proposed a network selection algorithm based on Q-learning, which maximizes user satisfaction by mapping the QoS attributes of the network to Quality of Experience (QoE) parameters and then constructing a reward function for user selection of the network. Liu et al. [26] proposed a combined fuzzy logic reinforcement learning handover algorithm that integrates the advantages of subtractive clustering and a Q-learning framework into the traditional fuzzy-logic-based handover algorithm. The optimal switching strategy is obtained by using the Q-learning algorithm. Although the network selection algorithm based on Q-learning does not need to know the state transition probability in advance and can store the state-action space pair through the Q value table, with the dense deployment of base stations in the HUDN environment, the dimension of the state space continues to increase. The Q-value table is unable to store all state-action space pairs. Therefore, Sun et al. [27] proposed using the deep reinforcement learning framework to solve the problem that the algorithm does not easily converge due to the large state space in the network selection process and designed a deep Q network (DQN) selection algorithm. Cao et al. [28] proposed a DRL-based user access control algorithm for open wireless access networks. Configuring a DRL-based user access control scheme on the carrier intelligence controller to maximize throughput avoids problems such as

frequent user switching. Yang et al. [29] proposed an end-to-end network handover algorithm based on DRL, which takes user service time, interruption times and handover cost as constraints of the reward function, which effectively improves the utilization of network resources.

Although the application of DRL in network selection algorithms has achieved good results, these studies did not design reward functions based on the relationship between the network attributes and the user services, and the designed reward function directly affects the final performance of the algorithm. Some studies did not take into account the user's business preferences. Therefore, to solve the problems in the above research, we proposed a network selection algorithm based on deep reinforcement learning that selects the most suitable network for different user services based on a deep neural network, utility function, AHP, standard deviation and other methods. With the goal of maximizing the accumulated reward value of the user-selected network, the state, action space and reward function of the user-selected network are designed. By calculating the network selection benefits for different types of services initiated by users, the analytic hierarchy process is used to establish the weight relationship between the different user services and the network attributes. Finally, a deep Q neural network is used to solve and optimize the proposed target and obtain the user's best network selection strategy and long-term network selection benefits. The main contributions of the proposed algorithm are as follows:

- (a) A network selection algorithm aimed at the long-term benefit is proposed to avoid the problem of a high network blocking rate caused by multiple users accessing the same network. At the same time, dueling network mechanism is introduced to avoid the problem that the algorithm selects the network incorrectly due to the insignificant difference of network score values in the HUDN environment.
- (b) Different business types of users are distinguished, and a utility function is used to construct different reward functions for the different user business types. This avoids errors in the results of network selection caused by unreasonable reward function settings. The cumulative reward value of the user's choice of network increases, and the number of handover times decreases.
- (c) Considering the user's preferences for different services, to avoid subjective preference weight settings, the standard deviation method is introduced to calculate the objective preference weight, and the comprehensive preference weight is obtained through reasonable calculation.

4 System model

In order to differentiate from previous studies on heterogeneous network selection [30], the network density of the studied network selection problem is highlighted. The system model in this paper not only increases the type and number of networks, but also adds the number of networks. The consideration of 5G new business, so as to reflect the general scenario of multiple services and multiple networks, has certain extensibility and versatility. The HUDN considered is composed of LTE macro-base stations,

wireless local area networks and 5G micro-base stations, as shown in Fig. 1. Among them, macro-base stations provide users with wide-area network services, micro-base stations provide users with small-scale and high-quality network services, and WLANs provide users with low-cost network services. The wireless access network adopts orthogonal frequency division multiplexing (OFDM).

technology. All networks are unified and converged to access the core to allow users to access the internet. User candidate networks are represented as the set $N = \{n_1, n_2, n_3, \dots, n_i, \dots, n_n\}$, where n_i is the i th candidate network.

Each network in the system has five network attributes, is bandwidth (B), delay (D), jitter (J), price (E), and packet loss (P). The user terminal business considers 5G Communication under the new business, such as 4k Ultra HD video, industrial remote control, telecommuting, smart home, and so on. These business information flows are divided into three categories, namely eMBB, URLLC, and mMTC. At the same time, the key performance indicators (KPIs) of these three types of new services are similar to traditional user service KPIs, mainly including bandwidth, delay, jitter, and price. For example, the eMBB scenario mainly meets the needs of future-oriented mobile Internet services. Its typical application, 4K ultra-clear video call, as an example, requires a bandwidth of at least 75 Mbps and a delay of less than 100 ms [31]. The URLLC service needs to meet the requirements of low latency and high reliability. Taking the typical remote control application as an example, it needs at least 50 ms delay to make the reliability reach 99.99%,

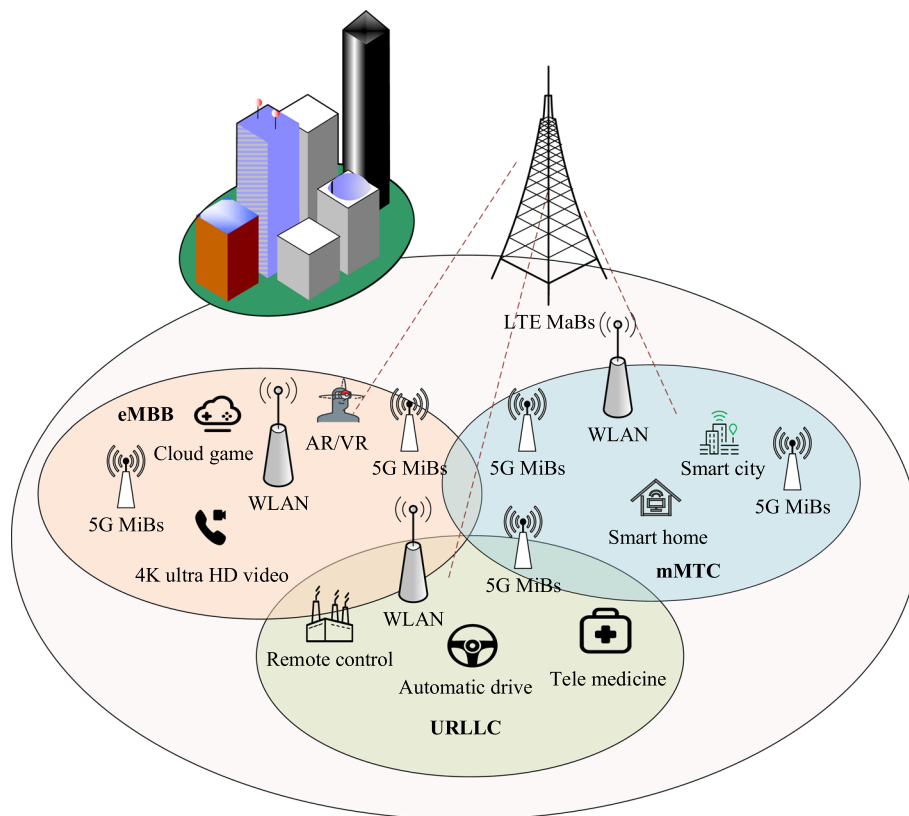


Fig. 1 System model

Table 1 QoS requirements of three types of services

Network attributes	Service types		
	eMBB	URLLC	mMTC
B(Mbps)	75	10	0.1
D(ms)	60	40	–
J(ms)	–	30	–
E	–	–	2
P (10 ⁻⁶)	10	–	30

and some key services need less than 30 ms delay and at least 10 Mbps bandwidth [32]. The mMTC scenario aims to provide reliable connections for devices with low power consumption, low cost, and small data packets. Taking its typical application of smart home as an example, it has low tolerance for price and packet loss rate [33]. Table 1 presents the different QoS requirements of the three types of services, and these values provide the relevant basis for the design of the network selection algorithm and utility function.

5 The network selection algorithm

5.1 Introduction to DRL

Reinforcement learning consists of three parts: state, action space and reward function. The goal of reinforcement learning is to learn a strategy π through online training, take an action a_t based on the current state s_t of the environment, obtain the reward r_t of environmental feedback and then go to the next state s_{t+1} . This process is repeated continuously to obtain the maximum cumulative long-term return, as shown in formula (1). In addition, for the optimal strategy π^* all satisfy $r^{\pi^*} > r^\pi$, where γ^t is the discount factor, and $r(t|s=s_t, a=a_t)$ is the instant reward for taking action a_t at time t .

$$r^\pi = E_\pi \left[\sum_{t=1}^T \gamma^t r(t|s = s_t, a = a_t) \right] \quad (1)$$

The Q-learning algorithm is a classical reinforcement learning algorithm that uses a Q-value table to store the optimal strategy and select the best action. However, in a complex environment, it is difficult to search the Q-value table to select the optimal strategy due to the large state space and action space. Therefore, by combining deep learning (DL) with reinforcement learning and fitting the Q value through deep neural networks (DNNs), we obtain DRL, the general process is as follows [34].

Firstly, at each moment, the intelligent agent interacts with the environment to obtain a current environmental state and uses DL to perceive the environmental state, in order to obtain a specific feature representation of the state;

Secondly, evaluate the value function of each action based on the rewards received, and adopt a strategy to map the current state to the corresponding action;

Finally, when the action is taken, the environment responds to it and enters the next state. By continuously cycling through the above processes, the best strategy to achieve the goal can ultimately be achieved. Shown in Fig. 2 is the schematic diagram of DRL.

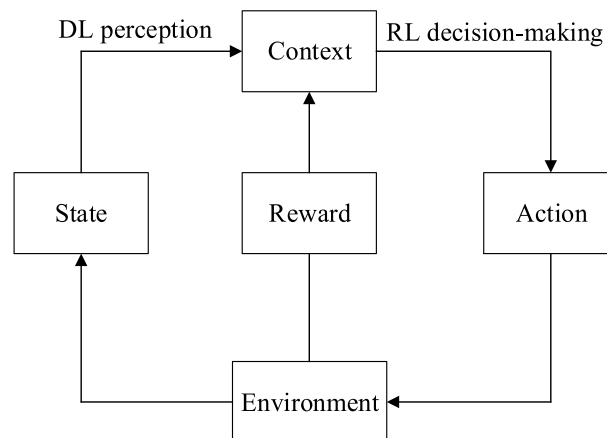


Fig. 2 The schematic diagram of DRL

5.2 State, action space, and reward function

Combined with the studied network selection problem in the HUDN environment, the DRL-based network selection model elements are defined as follows:

- (a) Agent: The agent considered here is a mobile user covered by HUDN.
- (b) State space: The state space considered here consists of network QoS attributes M (bandwidth B , delay D , jitter J , price E , packet loss rate P), service type G currently initiated by the user and current selection network of n_i . Then, at decision time t , the state space S^t can be expressed as the set shown in (2);

$$S^t = \{M^t, G^t, n_i^t\} \tag{2}$$

- (c) Action space: The user is under the coverage of multiple networks and can only choose one network to access at the time of handover decision t . Then, the action space can be expressed as the set of (3), where a_t is the action taken by the mobile user at the moment of switching decision t , and n_i represents the i th candidate network.

$$a_t = \{a | a \in (n_1, n_2, \dots, n_i, \dots, n_n)\} \tag{3}$$

- (d) Reward function: According to the three application scenarios of 5G divided by ITU, consider the services initiated by users as eMBB, URLLC and mMTC. Considering the QoS requirements of user services, the reward function can be defined as shown in formula (4), where $j \in \{B, D, J, E, P\} \times \{n_1, n_2, \dots, n_i, \dots, n_n\}$ is expressed as the QoS attribute value of different networks, w_j is the user's preference weight for different network attributes, and $r_j(s, a)$ is the instant reward for users to choose a network under different network QoS attributes. $C(s, a)$ is the network handover cost. When the selected network meets the QoS required by the service, $C(s, a) = 0$; otherwise, $C(s, a) = -1$.

$$r(s, a) = \sum_{j=1}^m w_j r_j(s, a) + C(s, a) \tag{4}$$

5.3 Determination of network parameter attribute weights

In the HUDN environment, all networks can be connected and used by users. Combined with the three major application services under 5G communication mentioned above, different service types have different requirements for network performance, and users who initiate different services have different preferences for network attributes. According to feedback from users, eMBB services have greater demands on network bandwidth and delay, while URLLC services are more sensitive to network delay and jitter and mMTC services have higher requirements for network packet loss rate and price. Therefore, to effectively distinguish the different needs of different services for the network, it is necessary to assign different weights of network attribute parameters to different services when constructing the reward function [35, 36]. In this paper, the AHP is used to assign subjective weights to different network attributes. The AHP structure is shown in Fig. 3.

By constructing a preference matrix, the sensitivity of different services to different network attribute parameters can be effectively represented. Here, a 1–9 scaling method is introduced to construct a discriminant matrix by pairwise comparison of different factors at the same level, which is expressed as $Z = (z_{ij})_{m \times m}$, where z_{ij} represents the comparison result between the i th factor and the j th factor, and $z_{ij} = 1/z_{ji}$. Because the constructed preference matrix is based on subjective judgment, the consistency ratio index CR can be calculated by formula (5) and (6), where λ_{max} is the maximum eigenvalue of the preference matrix, m is the number of preference matrix indicators, and RI is the random consistency index of the preference matrix. When $m = 5$ and $RI = 1.12$, if $CR < 0.1$, then the constructed preference matrix can pass the consistency test and satisfy the condition of subjective weighting.

$$CI = \frac{\lambda_{max} - m}{m - 1} \tag{5}$$

$$CR = \frac{CI}{RI} \tag{6}$$

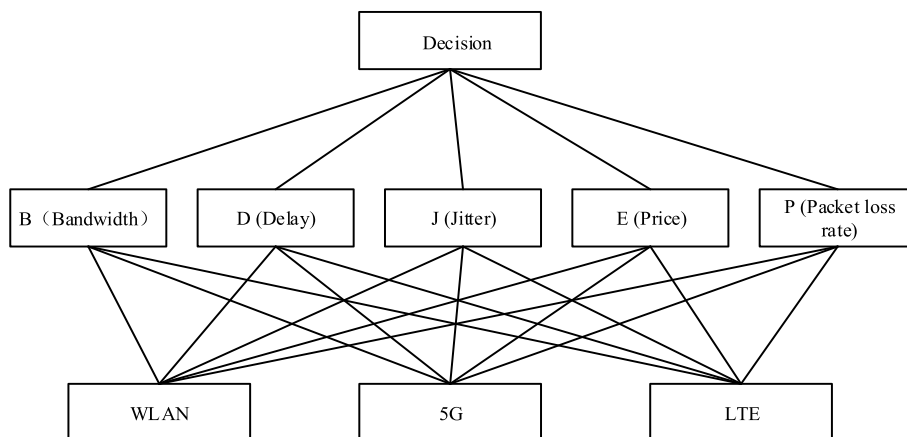


Fig. 3 user preference hierarchy structure

here we consider that all networks have 5 state attributes: bandwidth B, delay D, jitter J, price E and packet loss rate P, so $i, j, m = 1, 2, \dots, 5$. The subjective weights of network attributes are calculated by formula (7), and the discriminant matrices of the three types of services obtained are shown in Tables 2, 3 and 4.

$$w_j^s = \frac{\left(\prod_{i=1}^5 z_{ij}\right)^{1/m}}{\sum_{i=1}^5 \left(\prod_{j=1}^5 z_{ij}\right)^{1/m}} \tag{7}$$

Since the weights calculated by the AHP method are based on user preferences to weight the network attribute parameters, the weight results obtained are too subjective. Therefore, the mean square error method is used to calculate the objective weights of network attribute parameters, where $R = (r_{ij})_{n \times m}$ is the normalized matrix obtained after dimensionless X , and the objective weight of the network attributes can be calculated by the formula (8), where $\bar{r}_j = 1/n \sum_{i=1}^n r_{ij}$.

$$w_j^o = \sqrt{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2} / \sum_{j=1}^m \sqrt{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2} \tag{8}$$

Table 2 eMBB business discrimination matrix

eMBB	B	D	J	E	P
B	1	5	4	7	6
D	1/5	1	2	3	4
J	1/4	1/2	1	2	3
E	1/7	1/3	1/2	1	1/2
P	1/6	1/4	1/3	2	1

Table 3 URLLC business discrimination matrix

URLLC	B	D	J	E	P
B	1	1/4	1/2	3	2
D	4	1	2	4	6
J	2	1/2	1	2	3
E	1/3	1/4	1/2	1	1/3
P	1/2	1/6	1/3	3	1

Table 4 mMTC business discrimination matrix

mMTC	B	D	J	E	P
B	1	1/2	1/2	1/5	1/4
D	2	1	3	1/2	1/2
J	2	1/3	1	1/3	1/3
E	5	2	3	1	2
P	4	2	3	1/2	1

Then, the comprehensive weight of network attributes w_j can be obtained by formula (9), where w_j^s and w_j^o are the subjective and objective weights of network attributes, respectively.

$$w_j = \frac{(w_j^s w_j^o)}{\sum_{j=1}^m (w_j^s w_j^o)^{1/2}} \tag{9}$$

5.4 Construction of comprehensive reward function

In some papers that proposed DRL-based network selection algorithms, the reward value for user selection of the network is usually calculated using a piecewise function [37, 38]. However, due to the diversity of mobile user services, different types of services have different network requirements. Using a single reward function may lead to inaccurate network selection by users, and even the selected network cannot meet the QoS requirements of user services. The utility function can effectively represent the interest relationship between the user’s business and the network attributes, so the utility function is used to calculate the utility value to measure the user’s satisfaction with each attribute of the selected network.

Three types of services are considered here, which are initiated by users (eMBB, URLLC and mMTC). The QoS attributes of the network include bandwidth (B), delay (D), jitter (J), price (E) and packet loss rate (P). Then, the user instant reward can be defined as formula (10):

$$\sum_{j=1}^m w_j r_j(s, a) = w_B r_B(s, a) + w_D r_D(s, a) + w_J r_J(s, a) + w_E r_E(s, a) + w_P r_P(s, a) \tag{10}$$

$r_B(s, a)$, $r_D(s, a)$, $r_J(s, a)$, $r_E(s, a)$, $r_P(s, a)$ are the utility functions of bandwidth, delay, jitter, price and packet loss rate, respectively. The network bandwidth is a benefit attribute, so for eMBB and URLLC services, the bandwidth utility function adopts the sigmoid function as $f(x) = \frac{1}{1+(x/a)^{-b}}$. For mMTC services, the bandwidth utility function adopts an exponential function $u(x) = 1 - e^{-hx}$, where the parameters a are used to adjust the threshold of the function, and b and h are used to adjust the function steepness. The network delay utility function adopts a sigmoid function, but it is a cost attribute, so the sigmoid function used needs to be rewritten as $h(x) = \frac{(\frac{x}{a})^{-b}}{1+(\frac{x}{a})^{-b}}$. For eMBB and URLLC services, the network jitter utility function is designed using a logarithmic function as $p(x) = 1 - (m + k \ln(x + l))$, for mMTC services, the network jitter utility function is designed using a linear function $g(x) = cx + d$. where m, l, d is used to adjust the function threshold, and k, c is used to adjust the steepness of the function. For the network price utility function, a piecewise function is used to design, as shown in formula (11):

$$z(x) = \begin{cases} 1, & 0 \leq x \leq i \\ \frac{j-x}{j-i}, & i < x \leq j \\ 0, & x > j \end{cases} \tag{11}$$

The parameters i, j represent the minimum and maximum network prices acceptable to the business, respectively. Since the three types of services eMBB, URLLC and mMTC can tolerate a certain packet loss rate, the packet loss rate utility function is designed as a linear function. The utility functions and parameter settings of the three types of services corresponding to different network attributes are shown in Fig. 4 and Table 5.

Therefore, the comprehensive reward function of each type of business can be expressed as formula (12).

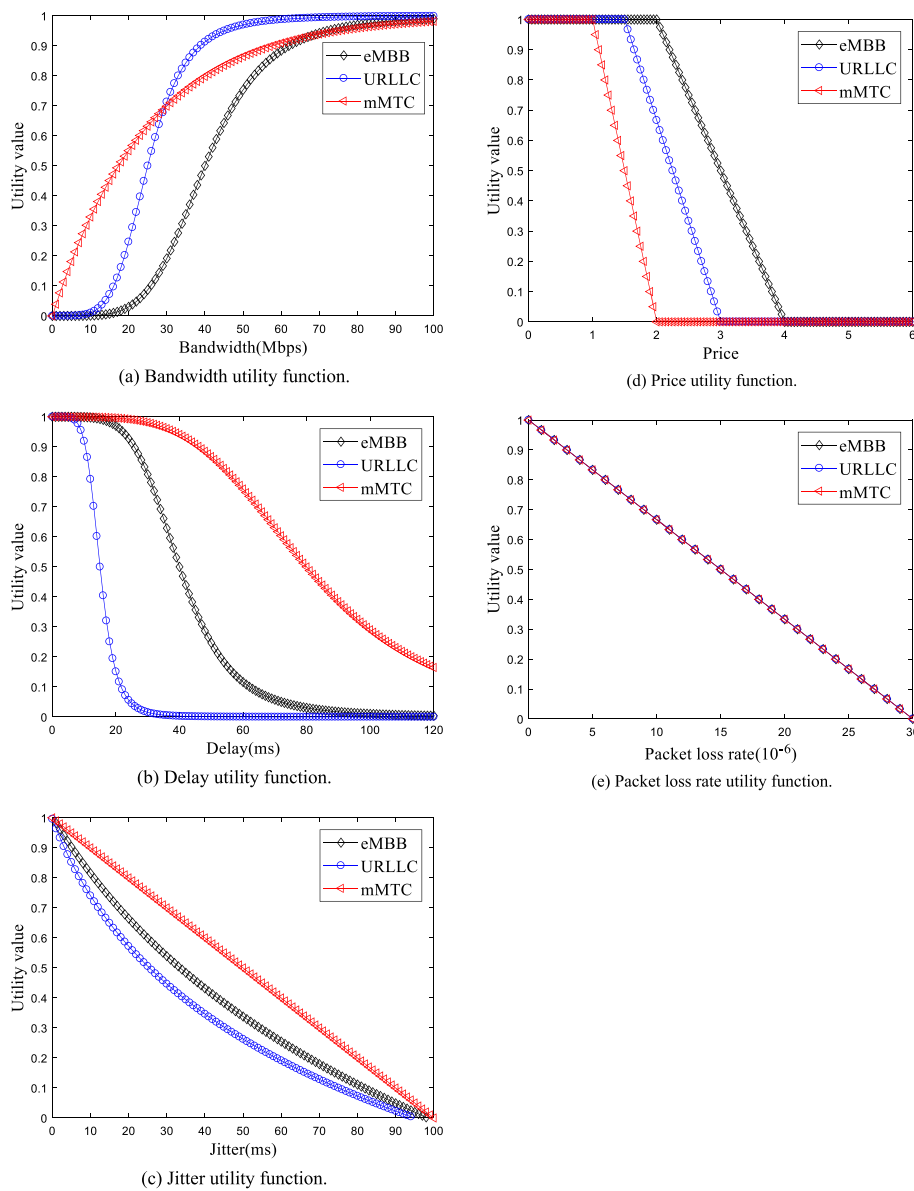


Fig. 4 Three types of business utility functions

Table 5 Utility function parameter settings

	eMBB	URLLC	mMTC
Bandwidth	Sigmoid function $a = 40, b = 5$	Sigmoid function $a = 25, b = 5$	Exponential function $h = 0.04$
Delay	Sigmoid function $a = 40, b = 5$	Sigmoid function $a = 15, b = 6$	Sigmoid function $a = 80, b = 4$
Jitter	Logarithmic function $m = -2.67$ $k = 0.75, l = 35$	Logarithmic function $m = -1.35$ $k = 0.5, l = 15$	Linear function $c = -0.01, d = 1$
Price	Piecewise function $i = 2, j = 4$	Piecewise function $i = 1.5, j = 3$	Piecewise function $i = 1, j = 2$
Packet loss rate	Linear function $c = -1/30, d = 1$	Linear function $c = -1/30, d = 1$	Linear function $c = -1/30, d = 1$

$$\left\{ \begin{array}{l}
 r_{\text{eMBB}}(s, a) = \omega_{\text{eMBB}}^B * f_{\text{eMBB}}(x) + \omega_{\text{eMBB}}^D * h_{\text{eMBB}}(x) \\
 \quad + \omega_{\text{eMBB}}^J * p_{\text{eMBB}}(x) + \omega_{\text{eMBB}}^E * z_{\text{eMBB}}(x) \\
 \quad + \omega_{\text{eMBB}}^P * g_{\text{eMBB}}(x) \\
 r_{\text{URLLC}}(s, a) = \omega_{\text{URLLC}}^B * f_{\text{URLLC}}(x) + \omega_{\text{URLLC}}^D * h_{\text{URLLC}}(x) \\
 \quad + \omega_{\text{URLLC}}^J * p_{\text{URLLC}}(x) + \omega_{\text{URLLC}}^E * z_{\text{URLLC}}(x) \\
 \quad + \omega_{\text{URLLC}}^P * g_{\text{URLLC}}(x) \\
 r_{\text{mMTC}}(s, a) = \omega_{\text{mMTC}}^B * u_{\text{mMTC}}(x) + \omega_{\text{mMTC}}^D * h_{\text{mMTC}}(x) \\
 \quad + \omega_{\text{mMTC}}^J * g_{\text{mMTC}}(x) + \omega_{\text{mMTC}}^E * z_{\text{mMTC}}(x) \\
 \quad + \omega_{\text{mMTC}}^P * g_{\text{mMTC}}(x)
 \end{array} \right. \quad (12)$$

In summary, the calculation of the reward value of the user selecting the network can be expressed as follows:

Input: QoS attribute value of each network; service type G currently initiated by the user; minimum QoS required by the current service;

Output: Comprehensive reward value

1. Initialize various parameters
2. According to the different business types initiated by the user, formula (4) and formula (12) are used to calculate the comprehensive reward value $r(s, a)$
3. Return $r(s, a)$

Algorithm 1 User selection network comprehensive reward value calculation

5.5 Network selection algorithm based on Dueling-DDQN

In HUDN, due to the dense deployment of the network, the state space and action space become very large. And when the number of networks increases, in a certain state, there may be cases where the scores of different networks are not obvious. Therefore, when the agent chooses a certain action and obtains a better reward value, it will not be able to evaluate because in the current state, each time a network with a high score still gets a higher reward value because the agent action is selected appropriately. To solve these problems, this paper introduces a deep Q neural network and uses the Dueling-DDQN algorithm to solve the DRL-based network selection problem. The specific algorithm structure is shown in Fig. 5.

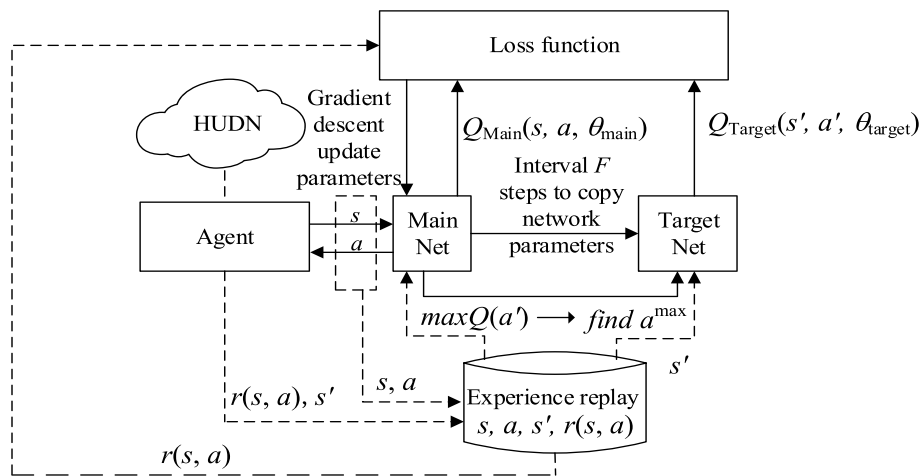


Fig. 5 Structure of network selection algorithm based on Dueling-DDQN

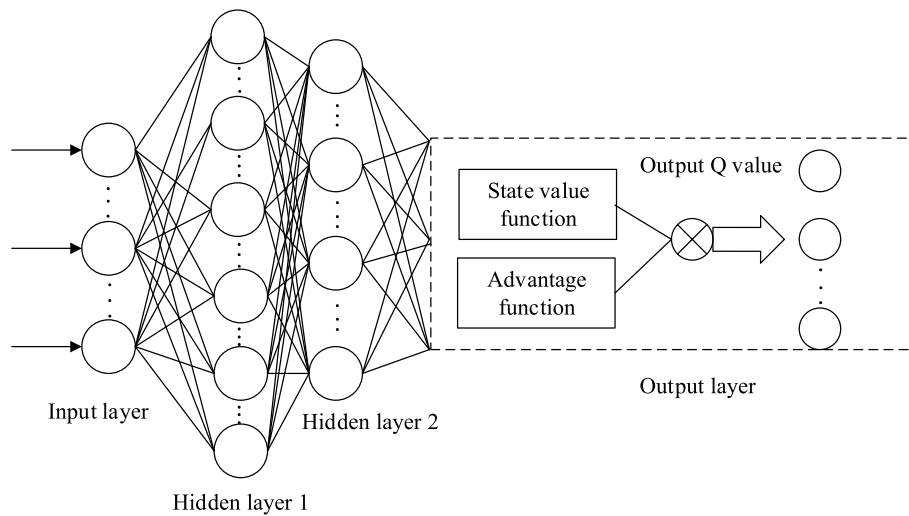


Fig. 6 Structure of main-net and target-net

Dueling-DDQN is a deep reinforcement learning algorithm based on dueling architecture mechanism in which Main-Net and Target-Net are the same structure [39]. The specific structure is shown in Fig. 6. It is mainly composed of input layer, hidden layer 1, hidden layer 2, and output layer, because the agent needs to obtain system state information, and the system state consists of eight networks, and five network attributes, and it consists of the user’s current business type and the selected network, so the number of neurons in the input layer is set to $5n + 2$, the hidden layer 1 is 128, the hidden layer 2 is 64. The output layer is composed of a state value function and an advantage function. The state value function is used to represent the state value of the current environment, while the advantage function represents the reward value that can be brought by selecting an action in the current state, respectively, expressed as $V(s; \theta, \beta)$ and $A(s, a; \theta, \alpha)$. Finally, combining the two, the Q value obtained by taking actions in each state is obtained, as shown in formula (13). Among them, θ is the parameters of the input layer

and the hidden layer, β and α are the neural network parameters of the two branches of the output layer, respectively.

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + A(s, a; \theta, \alpha) \tag{13}$$

By decoupling the calculation of the Q value, Main-Net is used to select the action taken by the next state with the highest Q value and then uses Target-Net to calculate the highest Q value that can be obtained after taking the next action to avoid the problem of overestimating the Q value.

Among them, the optimal action selection of Main-Net can be expressed as (14), and according to the Bellman formula [40], the Q value update method in Main-Net is shown in formula (15). θ_{main} is the parameter in Main-Net, s' is the next state, α is the learning rate and γ is the discount factor, r is the immediate reward value.

$$a^{\max}(s, \theta_{\text{main}}) = \arg \max_a Q_{\text{Main}}(s, a, \theta_{\text{main}}) \tag{14}$$

$$Q_{\text{Main}}(s, a, \theta_{\text{main}}) = Q_{\text{Main}}(s, a, \theta_{\text{main}}) + \alpha \left(r + \gamma \max_a Q_{\text{Main}}(s', a, \theta_{\text{main}}) - Q_{\text{Main}}(s, a, \theta_{\text{main}}) \right) \tag{15}$$

The Q value calculated by Target-Net according to the action can be expressed as formula (16), which is the fitting Q value obtained by DNN approximation, θ_{target} is the parameter of Target-Net.

$$Q_{\text{Target}}(s', a', \theta_{\text{target}}) = r + \gamma Q_{\text{Target}}[s', a^{\max}(s, \theta_{\text{main}}), \theta_{\text{target}}] \tag{16}$$

The mean square deviation method is used to calculate the loss function. The Q value obtained by Main-Net and the Q value obtained by Target-Net are used to calculate the error. The obtained loss function is shown in formula (17). The gradient descent method is used to optimize the neural network parameters and backpropagated to the main Q network to minimize the loss function and train the network.

$$L(\theta) = E \left[\left(Q_{\text{Target}}(s', a', \theta_{\text{target}}) - Q_{\text{Main}}(s, a, \theta_{\text{main}}) \right)^2 \right] \tag{17}$$

To increase the convergence speed of the algorithm proposed, a strategy ϵ -greedy is adopted for the selection of the agent strategy. The specific formula is shown in the following formula (18). During each round of learning and training, the agent will randomly select the strategy with the probability of ϵ . Or use the probability of $1 - \epsilon$ to select the strategy that can obtain the maximum return in the current state. Among them, ϵ is the set exploration rate, which will continue to decrease with the increase in the number of training rounds until it is 0 during the training process.

$$\pi^*(s) = \begin{cases} \arg \max_a Q(s, a), p = 1 - \epsilon \\ \text{random}(\pi), p = \epsilon \end{cases} \tag{18}$$

At the same time, the experience replay mechanism is introduced here. After each round of network selection decision-making, a quadruple is generated, which is determined by the current environment state s , the action a currently taken by the agent, the

reward value $r(s, a)$ obtained, and the environment state s' at the next moment. These data are stored in the replay experience pool as experience data. When the amount of data in the replay experience pool reaches the set threshold, a batch of historical data is randomly selected from the replay experience pool for model training to speed up model training and algorithm convergence speed.

The process of the network selection algorithm based on Dueling-DDQN is as follows:

Input: state space S , action space A , initial exploration rate ε , learning rate α , discount factor γ , target Q network parameter update frequency F , training times M , decision time interval t , replay experience pool capacity N ;

Output: network selection strategy;

1. Empty the replay experience pool D and initialize the parameters of the main Q and target Q networks $\theta_{\text{main}} = \theta_{\text{target}}$
2. for episode=1: M
3. Initialize the environment state $S=s$;
4. for $t=1:T$
5. Use the current state s as input in the main Q network to obtain the Q value output corresponding to all the selection actions in the main Q network, and use the strategy ε -greedy to select the action a according to formula (18) in the current output Q value;
6. Execute the action a , obtain the environmental state s' at the next moment, and calculate the user's comprehensive reward value $r(s, a)$ according to the process shown in algorithm 1;
7. Store the quadruple $(s, a, s', r(s, a))$ in the replay experience pool D ;
8. if the capacity D is higher than N
9. Randomly extract the m experience sample from the experience pool D , $(s_i, a_i, s'_i, r_i(s, a), i = 1, 2, \dots, m)$, and calculate the current target Q value by formula (16);
10. Minimize formula (17) by gradient descent, and use backpropagation to update all parameters θ_{main} of the main Q network;
11. end if
12. if $T\%F=0$,
13. Update the target Q network parameters $\theta_{\text{main}} = \theta_{\text{target}}$;
14. end if
15. if s' is the terminal state or when the maximum number of training rounds is reached
16. End the training;
17. else
18. go to step 2;
19. end if
20. end for
21. end for

Algorithm 2 Network selection algorithm based on Dueling-DDQN

6 Simulation results and analysis

To verify the effectiveness of the proposed algorithm, simulation experiments are carried out on the proposed algorithm through the PyCharm platform. The programming language is based on Python 3.7 and uses TensorFlow GPU 2.6.0 to build a neural network framework. The computer used for training and testing in simulation experiments

Table 6 Experiment parameter settings

Parameter name	Parameter value
Initial exploration rate ϵ	0.1
Learning rate α	0.0002
Replay experience pool capacity N	100,000
discount factor γ	0.6
Number of macro-base stations	1
Number of WLANs	2
Number of micro-base stations	5
Batch size	128
Network parameter update frequency F	100

Table 7 Network attribute parameter setting

Network parameters	Bandwidth (Mbps)	Delay (ms)	Jitter (ms)	Price	Packet loss rate (10^{-6})
LTE	100	10–70	10–40	2–5	0–20
5G	1000	5–30	5–25	3–6	0–20
WLAN	600	20–100	25–60	1–3	0–30

has a central processing unit of i5-9300HF, a memory of 16 GB, and a graphics processing unit of GTX1660Ti. It is assumed that mobile users are under the coverage of the WALN network, LTE network and 5G network. The specific network attribute parameters and related experimental parameters refer to the settings in references [41–43], as shown in Tables 6 and 7. Except for the fixed value of network bandwidth, other network attribute values fluctuate randomly within a certain range. Different types of services are randomly generated by the user terminals and arrive at the network according to the Poisson process. Among them, the probability of the terminal initiating the three types of services is $P_{\text{eMBB}} = 50\%$, $P_{\text{URLLC}} = 30\%$, $P_{\text{mMTC}} = 20\%$. The terminal network selection interval is 5 s. The AHP method is used to obtain the weight results of different services on different network attributes, as shown in Fig. 7. At the same time, the value of the experience playback pool is set to 100,000, this is because there are three types of networks in the simulation setting, a total of eight networks, and each network has five network attributes of bandwidth, delay, jitter, price, and packet loss rate, and the attribute values fluctuate randomly within a certain setting range, so the state space of the entire system is very large. If the experience playback pool capacity is set too small during simulation, the neural network will not be able to learn enough state space during training. In this state, the most appropriate action cannot be selected, thus affecting the performance of the algorithm. For each experience sample taken from the experience playback pool, the batch size is 128. At the same time, the discount factor γ and the learning rate α are also two important hyperparameters that affect the final performance of the algorithm. The discount factor γ represents the importance of future rewards to current rewards, the learning rate α represents the allowable error value when updating the Q value, the exploration rate ϵ represents the probability of the agent randomly selecting actions. As shown in Figs. 8, 9, and 10, under different discount factor, learning

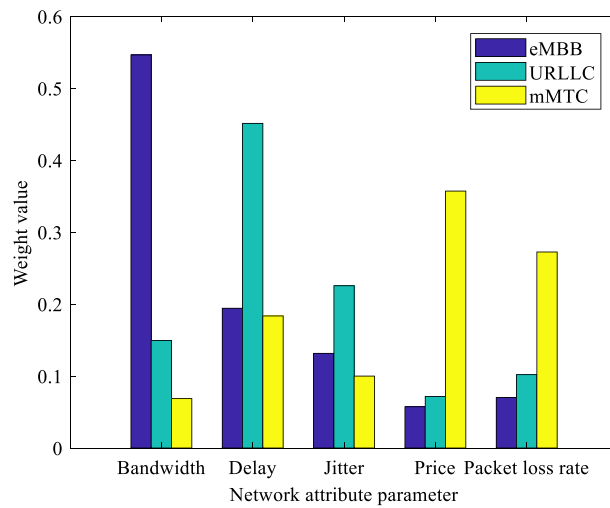


Fig. 7 Network attribute weights

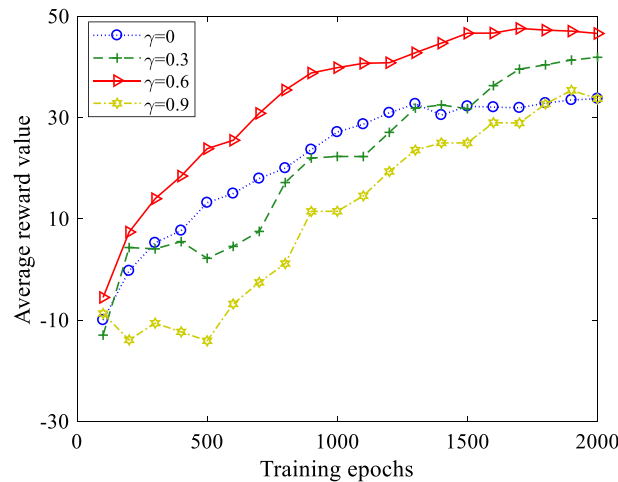


Fig. 8 The impact of different discount factor on reward values

rate and exploration rate, the average reward value received by the intelligent agent are different, when the discount factor is 0.6, the learning rate is 0.0002 and the exploration rate is 0.1, the algorithm performance is the best.

The simulation data are the average result obtained after training 2000 epochs and then running 800 epochs. And compared with MADM-SAW algorithm [44], Q-learning algorithm [45] and DDQN algorithm [29], where MADM-SAW is a multi-attribute decision-making algorithm that first normalizes the attribute parameters of each network. After parameter normalization, each normalized parameter is multiplied by a set weight value and added up. By comparing the normalized weighted sum of different networks, the network with the highest weighted sum is selected as the target network for access. Q-learning and DDQN algorithms are RL algorithms. In RL-based network selection algorithms, the goal of the agent is to train to obtain the optimal network selection strategy. Firstly, the agent takes action at to interact with the current network state s_t and

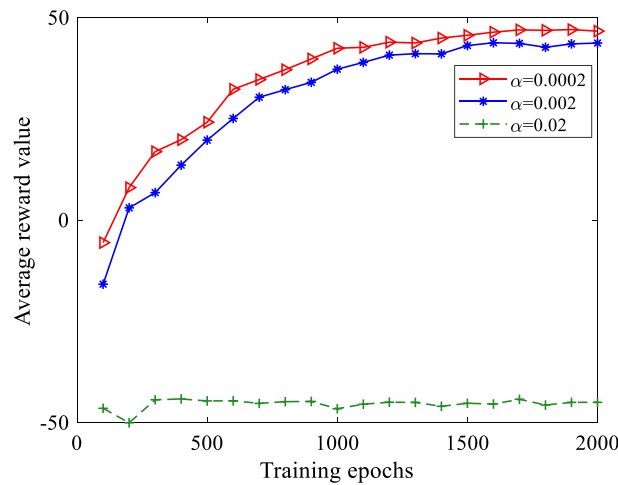


Fig. 9 The impact of different learning rates on reward values

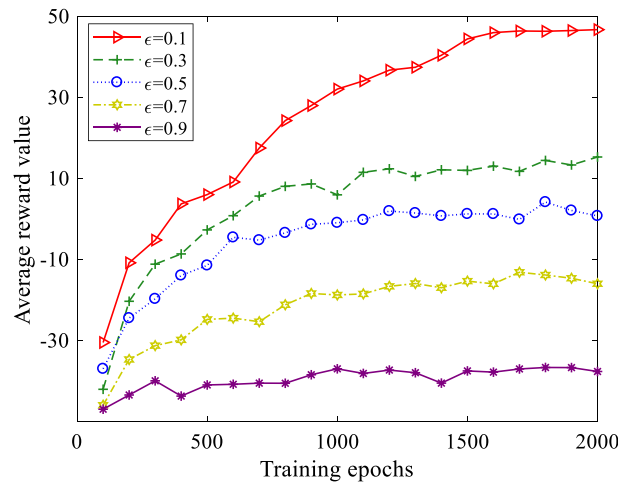


Fig. 10 The impact of different exploration rates on reward values

receive a reward r from the current environment feedback. Then, the intelligent agent continuously adjusts the action to be taken at the next moment based on the received reward value until it reaches the maximum reward value. Finally, the optimal network strategy is output.

Figure 11 shows the comparison of the average reward value obtained by the algorithm proposed in this paper with the other three algorithms in 800 experiments, respectively. It can be seen from the comparison experiments that the average reward value obtained by the proposed algorithm higher than the other three algorithms, respectively. This is because the proposed algorithm uses DNN to fit the Q value, which avoids the problem that the Q-learning algorithm cannot adapt to the large state and action space due to the limited size of the Q value table. Compared with the DDQN algorithm, the proposed algorithm adopts a decoupled Q value calculation method, effectively improving the cumulative reward for users choosing the network. At every decision moment, the

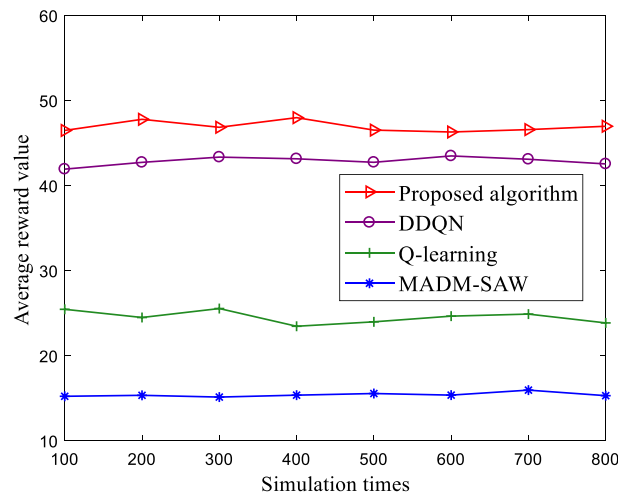


Fig. 11 Average reward value

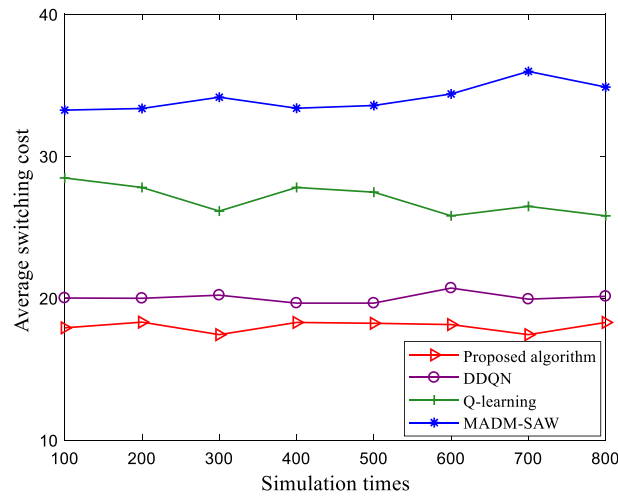


Fig. 12 Average switching cost

proposed algorithm can effectively adapt to the dynamic changes of HUDN and select the most suitable network for users to access. In the complex network environment of HUDN, MADM-SAW only uses simple weighted network parameters to provide terminal with access. The network is selected so that it has the lowest average cumulative reward value.

Figure 12 shows the average switching cost under different algorithms. Compared with the other three algorithms, since there is no special consideration for the QoS requirements of different service types, a reasonable switching cost function is designed for user’s terminal. Therefore, the other three algorithms finally get the average switching cost is higher than the proposed algorithm.

Figure 13 shows a comparison of the average single switching delay of different algorithms, which reflects the complexity of the algorithms. It is not difficult to see from Fig. 12 that the proposed algorithm has a slightly higher latency than the MADM-SAW

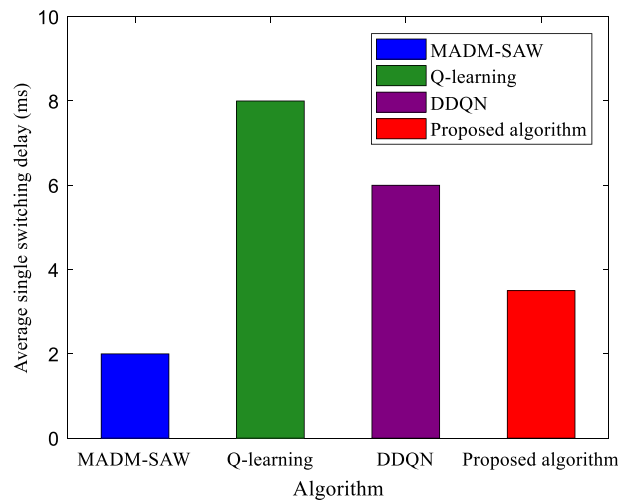


Fig. 13 Average single switching delay

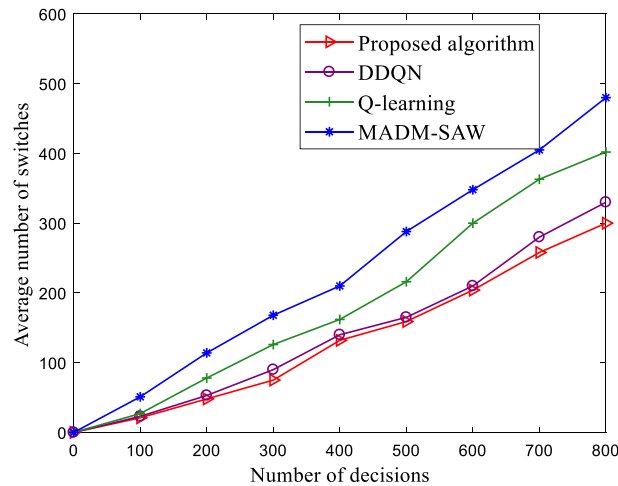


Fig. 14 Average number of switches

algorithm. This is because the MADM-SAW algorithm only performs simple weighting operations on network parameters, sacrificing algorithm performance for faster running speed, which can easily lead to low user satisfaction in complex scenarios. The switching delay of Q-learning and DDQN algorithms is higher than that of the proposed algorithm, because the proposed algorithm distinguishes different business types and avoids querying state action pairs from the Q-value table, reducing the preparation time before switching. Therefore, the switching delay of the proposed algorithm is lower than that of these two algorithms.

Figure 14 shows the average switching times of different algorithms. As seen from the figure, compared with the other three algorithms, the average switching times of the proposed algorithm are always lower. This is because the proposed algorithm considers the cost of network switching and simultaneously introduces a strategy ϵ -greedy for the user to select the network strategy so that the user can effectively select the network at

each decision-making moment, thereby reducing the number of switches and effectively reducing unnecessary switching. The other three algorithms fail to consider the cost of the terminal switching network and cannot effectively distinguish terminal service types, which leads to frequent network switching and intensifies the ping-pong handover effect. Therefore, the proposed algorithm can effectively reduce the number of network handovers.

Figure 15 shows the average handover failure probability of the four algorithms under different user numbers. It is not difficult to see from the figure that the average handover failure probability of the proposed algorithm is lower than that of the other three algorithms. This is because the proposed algorithm sets corresponding reward functions for different services initiated by users based on different QoS requirements, which maximally avoids the possibility of switching failure when users choose a network due to network performance not meeting their business needs. For the other three algorithms, due to their failure to consider the actual QoS requirements of user services and the dynamic changes in the HUDN environment, therefore the handover failure rate is higher than the proposed algorithm.

Figure 16 shows the changes in the network blocking rate of the four algorithms as the number of users increases. When the number of users is low, there is little difference in the performance of the four algorithms because when the number of users is very low, the network bandwidth can meet the needs of the service, and user terminal services will not be blocked. However, with an increase in the number of users, the network blocking rate of the four algorithms increases, the blocking rate of the proposed algorithm is not significantly different from that of the DDQN algorithm, but is significantly lower than that of the Q-learning algorithm and MADM-SAW algorithm. This is because the algorithm proposed in this paper introduces a replay experience mechanism on the basis of distinguishing different business types of user terminal. At the same time, the use of Deep Q Network can contain numerous state-action pairs to select the network with the best long-term income for users to access, effectively avoiding network congestion caused by users accessing a network at the same time. For the Q-learning algorithm, due

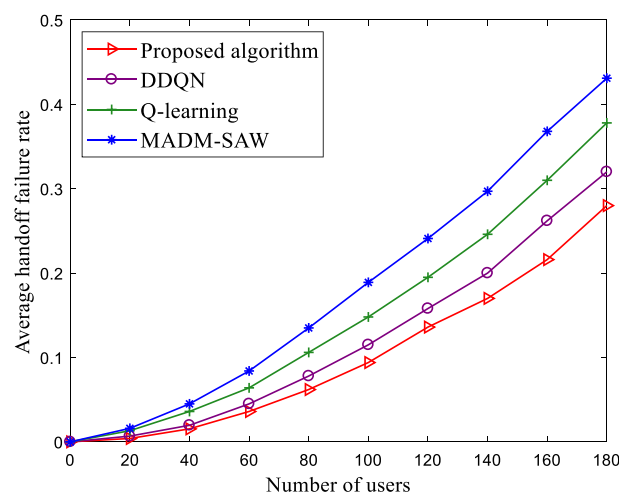


Fig. 15 Average handoff failure rate

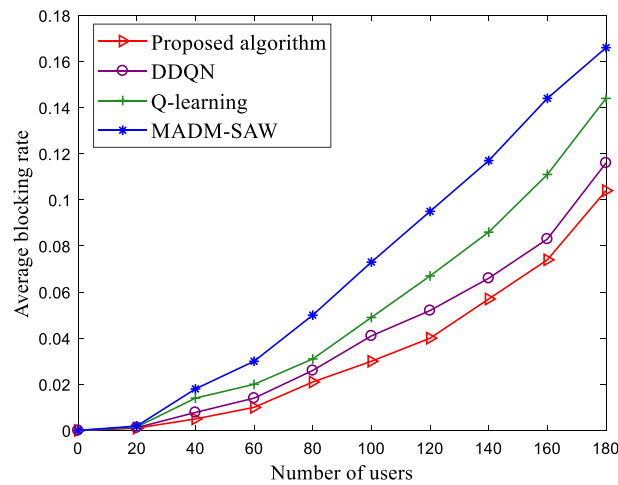


Fig. 16 Average blocking rate

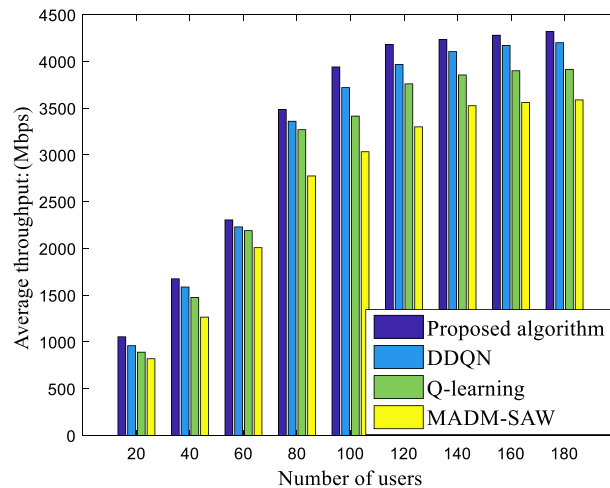


Fig. 17 Average throughput

to the limited storage capacity of the internally maintained Q value table, it cannot effectively accommodate various states and actions that should be taken. Therefore, when the number of users increases, the network blocking rate increases significantly. For the MADM-SAW algorithm, since it adopts the strategy of weighted network parameters for prioritization, it is easy to cause a large number of user terminals to access the same network with the best performance. Therefore, when the number of users increases, the network blocking rate is the highest.

Figure 17 shows the change in the average throughput of the network under the different number of users. When the number of users is low, the performance difference of the four algorithms is not obvious. This is because when the number of users is low, the system network resources are sufficient, and services are not easily blocked. When the number of users is greater than 120, the average network throughput obtained by the four algorithms increases slowly. This is because when the number of users increases, the available network bandwidth resources decrease, but at the same number of users,

the average network throughput obtained by the algorithm in this paper is greater than that of the other three algorithms. This is because the other three algorithms cannot effectively make reasonable network selection decisions for users when the number of optional networks is large. The MADM-SAW algorithm always selects the network access with the best theoretical performance. When the number of users increases, the network bandwidth resources are preempted by the first arriving service, and the service requests that arrive later are easily refused service by the currently selected network. This makes the overall network resource utilization low, resulting in the inability to effectively improve the total throughput. The Q-learning algorithm cannot contain the state space of many networks because of its limited Q value table, at the same time, the DDQN algorithm cannot independently evaluate the environment and the value of taking actions. Therefore, when selecting networks for users, these two algorithms the overall HUDN performance cannot be effectively considered, the throughput is lower than that of the proposed algorithm. The algorithm in this paper aims at the user's long-term network selection benefits and fully considers the different requirements of users on QoS for different services, reducing the possibility of network congestion and thus effectively improving the total network throughput.

7 Conclusion

This paper proposed a network selection algorithm based on Dueling-DDQN. For new services emerging under the 5G communication scenario, different reward functions were designed, and the advantages and disadvantages of the network selection strategy were determined using long-term decision-making benefits obtained by the user's network selection. The simulation results showed that the algorithm reduces the number of network switches and realizes the efficient use of network resources while guaranteeing maximization of the benefits of network selection for users. At the same time, due to the difficulty in predicting user movement trajectories, the proposed algorithm fails to effectively consider the issue of switching between networks while users are moving. In the future, under high user mobility conditions, the same layer and cross-layer interference between heterogeneous networks can be considered, and further research on network selection issues can be conducted.

8 Results and discussion

This paper proposes a novel network selection algorithm, which effectively improves the utilization efficiency of network resources and improves the user communication experience. However, the research results still have certain limitations:

- (1) The network selection studied in this paper only considers users choosing one network for access. With the development of intelligent communication products, more communication devices will support parallel access to multiple networks in the future. Therefore, the next step of research should focus on selecting multiple networks for parallel access.
- (2) The network selection studied in this paper is conducted under the condition of low user mobility, and the next step is to study the network selection of users in high-speed mobile states.

Abbreviations

HUDN	Heterogeneous ultra-dense network
QoS	Quality of service
MADM	Multi-attribute decision-making
TOPSIS	Technique for order preference by similarity to an ideal solution
FAHP	Fuzzy analytical hierarchy process
EWM	Entropy-weight-method
FL	Fuzzy logic
ITU	International telecommunication union
eMBB	Enhanced mobile broad band
URLLC	Ultra-reliable low latency communications
mMTC	Massive machine-type communications
MDP	Markov decision proposed
QOE	Quality of experience
DL	Deep learning
DRL	Deep reinforcement learning
DQN	Deep Q network
AHP	Analytic hierarchy process

Author contributions

JX contributed to the proposal of the concept and modeling. BZ contributed to the design and algorithm implement of the paper. CL contributed to result analysis of the paper.

Funding

This work was partially supported by Nation Science Foundation of China (62161016), and the Science and Technology Plan of Gansu Province (20JR10RA273).

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 13 June 2023 Accepted: 1 November 2023

Published online: 06 November 2023

References

1. M.M. Hasan, S. Kwon, S. Oh, Frequent-handover mitigation in ultra-dense heterogeneous networks. *IEEE Trans. Veh. Technol.* **68**(1), 1035–1040 (2019)
2. H. Yu, Y. Ma, J. Yu, Network selection algorithm for multiservice multimode terminals in heterogeneous wireless networks. *IEEE Access* **7**, 46240–46260 (2019)
3. S. Baghla, S. Bansal, An approach to energy efficient vertical handover technique for heterogeneous networks. *Int. J. Inf. Technol.* **10**(3), 359–366 (2018)
4. F. Jiang, C. Feng, H. Zhang, A heterogenous network selection algorithm for internet of vehicles based on comprehensive weight science direct. *Alex. Eng. J.* **60**(5), 4677–4688 (2021)
5. R. Honarvar, A. Zolghadrasli, M. Monemi, Context-oriented performance evaluation of network selection algorithms in 5G heterogeneous networks. *J. Netw. Comput. Appl.* **202**, 103358 (2022)
6. B. Priya, J. Malhotra, *5GNet: An Intelligent QoE Aware RAT Selection Framework for 5G-Enabled Healthcare Network*, vol. 14 (Springer, Berlin, Heidelberg, 2023), pp.8387–8408
7. G. Liang, X. Guo, G. Sun et al., Multi-attribute access selection algorithm for heterogeneous wireless networks based on uncertain network attribute values. *IEEE Access* **10**, 74071–74081 (2022)
8. P. Satapathy, J. Mahapatro, An adaptive context-aware vertical handover decision algorithm for heterogeneous networks. *Comput. Commun.* **209**, 188–202 (2023)
9. B.S. Khan, S. Jangsher, N. Hussain, M.A. Arafah, Artificial neural network-based joint mobile relay selection and resource allocation for cooperative communication in heterogeneous network. *IEEE Syst. J.* **16**(4), 5809–5820 (2022)
10. T.M. Duong, S. Kwon, Vertical handover analysis for randomly deployed small cells in heterogeneous networks. *IEEE Trans. Wirel. Commun.* **19**(4), 2282–2292 (2020)
11. K. Ahuja, B. Singh, R. Khanna, Network selection algorithm based on link quality parameters for heterogeneous wireless networks. *Optik* **125**(14), 3657–3662 (2014)
12. F. Zhao, H. Tian, G. Nie, and H. Wu, Received signal strength prediction based multi-connectivity handover scheme for ultra-dense networks, in *proc. Asia-Pac. Conf. Commun. (APCC), Ningbo, China* (2018), pp. 233–238
13. A. Kaswan, P. K. Jana, and M. Azharuddin, A delay efficient path selection strategy for mobile sink in wireless sensor networks, in *proc. Int. Conf. Adv. Comput., Commun. Inf. (ICACCI), Udupi, India* (2017), pp. 168–173
14. M. Alhabet, L. Zhang, Multi-criteria handover using modified weighted TOPSIS methods for heterogeneous networks. *IEEE Access* **6**, 40547–40558 (2018)

15. H.W. Yu, B. Zhang, A heterogeneous network selection algorithm based on network attribute and user preference. *Ad Hoc Netw.* **72**, 68–80 (2018)
16. G. Gaur, T. Velmurugan, P. Prakasam, S. Nandakumar, Application specific thresholding scheme for handover reduction in 5G ultra dense networks. *Telecommun. Syst.* **76**(1), 97–113 (2021)
17. M. Pradeep, P. Sampath, An optimized multi-attribute vertical handoff approach for heterogeneous wireless networks. *Concurr. Comput. Pract. Exp.* **31**(20), e5296 (2019)
18. N. Abbas, J.J. Saade, A fuzzy logic based approach for network selection in WLAN/3G heterogeneous network, in *Proc. Annu. IEEE Consumer Commun. Netw. Conf., (CCNC), Las Vegas, NV, USA* (2015), pp. 631–636
19. B. Naeem, R. Ngah, S.Z.M. Hashim, Reduction in ping-pong effect in heterogeneous networks using fuzzy logic. *Soft. Comput.* **23**(1), 269–283 (2019)
20. R.K. Goyal, S. Kaushal, A.K. Sangaiah, 'The utility based non-linear fuzzy AHP optimization model for network selection in heterogeneous wireless networks'. *Appl. Soft Comput.* **67**, 800–811 (2018)
21. X. Wu, Q. Du, 'Utility-function-based radio-access-technology selection for heterogeneous wireless networks'. *Comput. Electr. Eng.* **52**, 171–182 (2016)
22. J. Xie, W. Gao, C. Li, Heterogeneous network selection optimization algorithm based on a Markov decision model. *China Commun.* **17**(2), 40–53 (2020)
23. A. Khodmi, S.B. Rejeb, N. Nasser, and Z. Choukair, MDP-based handover in heterogeneous ultra-dense networks, in *Proc. Int. Conf. Inf. Networking (ICOIN), Jeju Island, Korea (South)* (2021), pp. 349–352.
24. B. Yang, X. Wang, Z. Qian, A multi-armed bandit model-based vertical handoff algorithm for heterogeneous wireless networks. *IEEE Commun. Lett.* **22**(10), 2116–2119 (2018)
25. L. He, D. Jiang, C. Wei, A QoE-based dynamic energy-efficient network selection algorithm. *Wirel. Netw.* **27**(1), 3585–3595 (2020)
26. Q. Liu, C.F. Kwong, S. Wei et al., Reinforcement learning-based joint self-optimisation method for the fuzzy logic handover algorithm in 5G HetNets. *Neural Comput. Appl.* **35**, 1–17 (2021)
27. J. Sun, Z. Qian, X. Wang, ES-DQN-based vertical handoff algorithm for heterogeneous wireless networks. *IEEE Commun. Lett.* **9**(8), 1327–1330 (2020)
28. Y. Cao, S.Y. Lien, Y.C. Liang, et al., Federated deep reinforcement learning for user access control in open radio access networks, in *Proc. IEEE Int Conf Commun. (ICC), Montreal, QC, Canada* (2021), pp. 1–6
29. F. Yang, W. Wu, X. Wang, Y. Zhang and P. Si, Deep reinforcement learning based handoff algorithm in end-to-end network slicing enabling HetNets. in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC), Nanjing, China* (2021), pp. 1–7
30. P. Dhand, S. Mittal, G. Sharma, An intelligent handoff optimization algorithm for network selection in heterogeneous networks. *Int. J. Inf. Technol.* **13**(5), 2025–2036 (2021)
31. H. Yin, L. Zhang, S. Roy, Multiplexing URLLC traffic within eMBB services in 5G NR: fair scheduling. *IEEE Trans. Commun.* **69**(2), 1080–1093 (2020)
32. J.S. Wey, J. Zhang, X. Lu, et al. Real-time investigation of transmission latency of standard 4K and virtual-reality videos over a commercial PON testbed, in *Optical Fiber Communications Conference & Exposition (IEEE, 2018)*
33. R. Liu, G. Yu, J. Yuan et al., Resource management for millimeter-wave ultra-reliable and low-latency communications. *IEEE Trans. Commun.* **69**(2), 1094–1108 (2021)
34. K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath, Deep reinforcement learning: a brief survey. *IEEE Signal Proc. Mag.* **34**(6), 26–38 (2017)
35. P. Dhand, S. Mittal, G. Sharma, An intelligent handoff optimization algorithm for network selection in heterogeneous networks. *Int. J. Inf. Technol.* **13**(5), 2025–2036 (2021)
36. R. Luo, S. Zhao, and Q. Zhu, Network selection algorithm based on group decision making for heterogeneous wireless networks, in *Proc. IEEE 9th Int. Conf. Commun. Softw. Netw. (ICCSN), Guangzhou, China* (2017), pp. 397–402
37. M. Wu, W. Huang, K. Sun, and H. Zhang, A DQN-based handover management for SDN-enabled ultra-dense networks, in *Proc. IEEE 92nd Veh Technol Conf (VTC2020-Fall), Victoria, BC, Canada* (2020), pp. 1–6
38. Y. Xu, W. Xu, Z. Wang, J. Lin, S. Cui, Load balancing for ultra dense networks: A deep reinforcement learning-based approach. *IEEE Internet Thing J.* **6**(6), 9399–9412 (2019)
39. Z. Wang, T. Schaul, M. Hessel, et al. Dueling network architectures for deep reinforcement learning, in *International conference on machine learning* (PMLR, 2016), pp. 1995–2003
40. Z. Wang, L. Li, Y. Xu, H. Tian, S. Cui, Handover control in wireless systems via asynchronous multiuser deep reinforcement learning. *IEEE Internet Thing J.* **5**(6), 4296–4307 (2018)
41. X. Tan, G. Chen, H. Sun, Vertical handover algorithm based on multi-attribute and neural network in heterogeneous integrated network. *EURASIP J. Wirel. Commun.* **2020**(1), 1–21 (2020)
42. G. Liang, H. Yu, X. Guo, Y. Qin, Joint access selection and bandwidth allocation algorithm supporting user requirements and preferences in heterogeneous wireless networks. *IEEE Access* **7**, 23914–23929 (2019)
43. A. Zhu, M. Ma, S. Guo et al., Adaptive multi-access algorithm for multi-service edge users in 5G ultra-dense heterogeneous networks. *IEEE Trans. Veh. Technol.* **70**(3), 2807–2821 (2021)
44. A. Debnath, N. Kumar, Simple additive weighted algorithm for vertical handover in heterogeneous network, in *2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS)* (IEEE, 2020), pp. 1–2.
45. L. He, D. Jiang, C. Wei, A QoE-based dynamic energy-efficient network selection algorithm. *Wirel. Netw.* **27**(5), 3585–3595 (2021)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jianli Xie was born in Gansu Province in China in 1972. He received a BSc. Degree from Sichuan University in 1993, and a Master degree and a Dr. degree from the School of Electronics and Information

Engineering, Lanzhou Jiaotong University in 1999 and 2014, respectively. From 1993 to 1996, he worked as an engineer in Lanzhou General Machinery Plant. From 1999 to 2007, he was a senior hardware engineer in Beijing Great Dragon company. Now he is working in Lanzhou Jiaotong University as a professor. He has published over 20 scientific papers in his research area till now. His research interests include railway wireless communication and cognitive radio techniques.

Binhan Zhu was born in Hunan Province in China in 1999. He is currently studying for a master's degree at the School of Electronics and Information Engineering, Lanzhou Jiaotong University. His research interests include heterogeneous wireless networks and wireless communication technologies.

Cuiran Li was born in Shanxi Province in China in 1975. She received bachelor's and master's degrees from the Department of Communication Engineering, Lanzhou Jiaotong University, in 1996 and 1999, respectively, and a Dr. degree from the School of Electronics and Information Engineering, Beijing Jiaotong University, in 2003. She is now working in Lanzhou Jiaotong University, as a professor and an IEEE member. She has published a book and over 40 scientific papers in her research area till now. Her research interests include railway mobile communication, wireless sensor network and cooperative communication technology.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
