Hindawi
Security and Communication Networks
Volume 2021, Article ID 5587656, 13 pages
https://doi.org/10.1155/2021/5587656

Research Article

A Reinforcement Learning-Based Configuring Approach in Next-Generation Wireless Networks Using Software-Defined Metasurface

Fatemeh Aliannejad, Esmaeel Tahanian, Mansoor Fateh, and Mohsen Rezvani

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

Correspondence should be addressed to Esmaeel Tahanian; e.tahanian@shahroodut.ac.ir

Received 12 January 2021; Revised 5 April 2021; Accepted 13 April 2021; Published 27 April 2021

Academic Editor: Mamoun Alazab

The next generation of wireless networks including Five and Six Generations (5*G* and 6*G*) can provide very high data rates as a demand for the Internet of Everything (IoE) system which connects millions of people and billions of machines. To reach such a high data rate, the wireless networks should work at high enough frequencies, such as millimeter and THz bands, which in turn suffer from a large attenuation and acute multipath fading. The idea of coating any objects in the environment with Software-Defined Metasurfaces (SDMs) was presented to control these effects by managing the electromagnetic properties of the environment. Since the programmable environment can be changed during the communication, for example, a sudden obstacle appears, this management should be adaptive. This paper presents the use of a reinforcement learning (RL) algorithm for dynamically configuring such an environment. In other words, when a change happens in the environment, for example, an obstacle blocks some EM waves, the agent receives a large punishment, and therefore a new action is selected. In our model, the transmitted electromagnetic waves and the tiles are considered as the agents and states, respectively. Moreover, the actions of each tile include absorbing or reflecting the impinging waves in a specific direction. We utilize the Q-learning technique to establish proper wireless links between the users and the access point (AP) by controlling the state of the tiles in an environment covered by the SDMs. Evaluation of the proposed model for different scenarios, including emerging sudden obstacles, indicates its potential to provide a proper signal level for all the users and improve the average received power up to 12% in comparison with the related works.

1. Introduction

With the increase of the smart wireless devices usage and the emergence of the idea of the Internet of Everything (IoE), wireless communication systems have been faced with a demand for higher data rates as well as a better Quality of Service (QoS). Accordingly, after deploying 5G around the world, academic and industrial efforts have started to conceptualize 6G [1–3]. To provide such a high data rate communication, these new generations of wireless systems should work at higher frequencies such as millimeter wave or even THz bands. On the other hand, the larger attenuation due to the free space path loss and lower penetration depth as well as acute multipath fading limit the use of such high-frequency electromagnetic (EM) waves [4–8]. To alleviate these drawbacks, the researchers have encouraged using the coating technologies to turn the propagation environments into a programmable media. Despite the regular environments in which the EM waves undergo various interactions including reflections, diffractions, and scattering, there is a complete control over EM waves impinging upon surfaces in the programmable wireless environment (PWE) and it is possible to reengineer their direction [9].

Relays [10, 11], reflectarrays [12–15], and metasurfaces [9, 16–19] are the most popular coating technologies for PWEs. Relays contain an array of low-cost antennas which passively either reflect or attenuate impinging EM waves on

the wall, configuring the environment to improve wireless networks operating nearby. Obviously, this approach cannot adaptively reconfigure the PWE. The second approach, reflectarrays, comprises a number of $\lambda/2$ patch antennas in a 2D grid arrangement. Moreover, there are some active elements such as PIN diodes to alter the phase of the reflected EM wave. The common functionality of reflectarrays is wave steering and absorption, especially reachable at the far field. Another approach is using metasurface that is similar to reflectarray but with a higher density of meta-atoms. Having high enough density lets the metasurface produce any EM profile, even in the near field [20]. We use the Software-Defined Metasurface (SDM) approach [16] in the remainder of this paper.

In an SDM-based PWE, all of the walls, doors, furniture, and other objects are coated with SDM tiles. Each tile has some networked hardware, control elements, and adaptive meta-atom metasurfaces and can receive external commands and set the states of its control elements to match the intended EM behavior [16]. Furthermore, tiles have environmental sensing and reporting capabilities to discover the communicating devices within the environment. In addition, there is a server that senses the EM profile and active users present in the environment and adaptively configures the matching functionality for each SDM tile.

Adaptively configuring the functionality of the tiles is a major challenge in an SDM-based environment, especially when the number and the position of the users can be constantly changing. However, the environment also can be changed; for example, an obstacle can suddenly block the directed EM waves and therefore some of the wireless links are failed. The employed approaches to control the wireless communication in the PWEs including solving an optimization problem [16] or neural-network-based technique [9] cannot dynamically reconfigure the functionality of tiles as soon as the environment changes.

In this paper, we propose an approach based on reinforcement learning (RL) to adaptively configure and control the indoor wireless communication environment. In our learning model, agents are the transmitted EM waves, the tiles play the role of states, and the different functionalities of tiles are considered as valid actions. By applying the reinforcement learning algorithm, the tiles can make a better decision, while the EM waves gradually interact with the environment. So, when a change happens in the environment, for example, an obstacle blocks some EM waves, the agent receives a large punishment and therefore a new action is selected. Since there is usually more than one user, we model the wireless communication environment as a multiagent RL problem. We consider different scenarios, namely, providing permanent coverage for the NLOS region, providing particular connections for each user when the number of users increases, providing connection in presence of obstacles, and finally providing connection after the sudden emergence of some obstacles, to evaluate the effectiveness of our approach. Simulation results indicate that the RL approach improves the average received power up to 12% in comparison with related works.

The main contributions of this paper are as follows:

(i) We properly model the PWE to apply the RL technique

(ii) We introduce a novel multiagent RL-based approach to configure PWE

(iii) Different scenarios are considered to model what is happening in a real PWE

The rest of this paper is organized as follows. Section 2 reviews the related works. The preliminary background for SDM is presented in Section 3. Section 4 provides the proposed RL model to configure the tiles. The paper is evaluated via ray-tracing-based simulations in Section 4.1. Finally, the conclusion, as well as future works, is presented in Section 5.

2. Related Work

The idea of an SDM-based communication environment is to coat the planar objects with units from a kind of software-controlled metasurfaces called SDM tiles [16]. As can be seen in Figure 1, the tiles consist of a set of switch elements controlled by a set of networked controllers and a gateway that provides intertile and external connectivity. Using the gateway, the network of the controllers receives commands from the configuration server to change the current state of the switches and therefore configure the tiles for yielding the desired functionality. In addition, the gateway has the role of a bridge to provide the required power for the tiles. According to Figure 1, the tiles form a network with grid topology in which at least one of the gateways is connected to the environment configuration server for accumulating environmental sensed data to discover the communicating devices within the environment and propagating the proper commands within the tile network. These commands contain at least the type of action such as steering or absorbing and the address of the intended tile gateway. It is worth noting that the translation of action to a tile switch element configuration is usually done using a populated lookup table during the tile design/manufacturing process [16].

Due to the development of controlling techniques for the EM waves behavior, the researchers are attracted to SDM-based PWEs, especially for the realization of the next generations of wireless networks. The most related literature focus on designing PWE tile unit [8, 13, 18, 21, 22] rather than the adaptive configuration approaches for it. In this section, we review the studies that concentrated on proposing the approaches to adaptively configure the SDM-based PWEs.

Liaskos et al. introduced the idea of using SDMs, for the first time, to have a programmable wireless environment, especially for the next generations of wireless networks [16]. They provided coverage in an NLOS region considering 12 receivers in that area and adequately set the states of the tiles by solving an optimization problem using the genetic algorithm (GA) technique. Moreover, power maximization over the NLOS region is considered as the criterion for the genetic algorithm. Although this approach leads to adequate

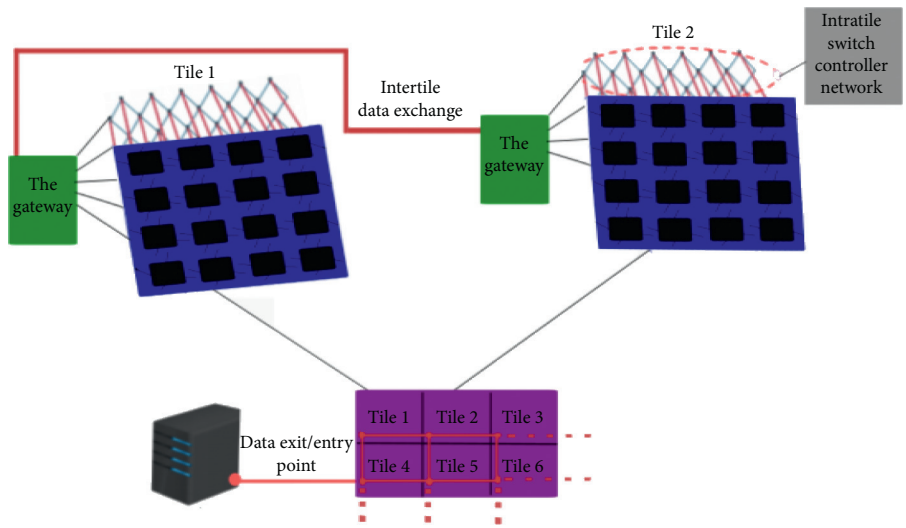

Figure 1: The component of a typical SDM tile.

coverage in the NLOS region, it cannot adaptively reconfigure the functionality of the tiles as soon as the environment changes. For example, when an object suddenly emerges and blocks some of the wireless links, a new optimization problem should be solved, which will be time-consuming.

Liaskos et al. in another work [9] presented an approach based on machine learning algorithms, in particular neural networks, to adaptively configure the tiles for a set of users. The authors model the wireless propagation as a neural network with walls as layers, tiles as nodes, and their cross interactions as links. This research considered a problem with only one user and the extension case with multiple users has not been discussed.

In another study, Liaskos et al. [17] model PWEs as a graph and describe their workflow and performance objectives as path finding problems. Unlike the above-mentioned works, this reference presented a network-layer solution to configure PWEs for multiple users and objectives. Nevertheless, the proposed approach in this reference is time-consuming to reconfigure the states of the tiles when the environment changes.

Our proposed RL-based approach in this paper can adaptively configure the PWEs that serve multiple users. In addition, when the environment changes, for instance, an obstacle blocks the path of some EM waves, the tiles can report this blockage to the configuration server. After recalculating the Q-table by considering the new reward/punishment value, the blocked EM waves can immediately find a new path. Table 1 summarizes the main adaptive configuration approaches for PWE tiles in the literature.

3. System Model and Problem Formulation

In this section, we present the reinforcement learning model of an SDM-based communication system working at millimeter-wave frequencies for the next generations of wireless networks. By applying reinforcement learning, the agent can make a better decision, while it gradually interacts with the environment [23, 24]. We utilize the Q-learning technique, which is one of the value-based reinforcement learning approaches, to establish proper wireless links between the users and the access point (AP) by controlling the state of the tiles in an environment covered by the SDMs. Q-learning searches for an optimal state-action policy, that is, a sequence of actions that maximize the expected discounted reward [25]. This optimal policy defines how the agent selects an action with the regard to its state. As a typical action-selection policy, the agent chooses the action with the highest Q-value with the probability of $1-\epsilon$ and acts stochastically with the probability of ϵ. In other words, there should be a tradeoff between exploration and exploitation. The exploration tries to execute the actions that are not executed before when the agent is at a specific state. At the beginning of the learning process, the agent has no experience of the environment and therefore it needs to obtain some rewards and punishments from the environment. So, in these conditions, the exploration is dominant and the value of ϵ is near one. Of course, as the agent interacts with the environment more and more and obtains some experiences, the value of ϵ is reduced and, therefore, the exploitation becomes dominant. The Q-value updating rule according to the Bellman equation [25] is given as follows:

$$Q_{(s_t,a_t)} \longleftarrow Q(s_t,a_t) + \alpha\left[r_{(t+1)}(s_t,a_t) + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\right], \tag{1}$$

where s_t is the state at time t, α is the learning rate, γ is the discount factor, and a' denotes one of the valid actions when the agent has state s_{t+1}.

Our reinforcement learning model for the SDM-based communication system has the following elements:

(i) Agent: An agent is the transmitted electromagnetic wave (signal) between a user and an AP.

(ii) State: The tiles in our model play the role of states in a reinforcement learning problem. At the beginning of each episode, according to the position of each

Table 1: Summary of the main adaptive configuration approaches for PWE tiles.

Reference	Approach	Discussion
Liaskos et al. [16]	Genetic algorithm (GA)	This approach cannot adaptively reconfigure the tiles functionality as soon as the environment changes.
Liaskos et al. [9]	Neural network	This approach considered a problem with only one user.
Liaskos et al. [17]	Graph model	This approach is time-consuming to reconfigure the states of the tiles when the environment changes.
This work	RL	This approach can adaptively reconfigure the states of the tiles when the environment changes in presence of multiple users.

user, its signal is received by a proper tile (e.g., the nearest tile) as the first state. Afterward, the agent (signal), with regard to the optimal policy, moves to the next states (tiles) until it arrives to the goal (AP). It should be noted that each tile can receive the EM waves with specific angles. So, when the agent is at state s_t, some of the states can be valid as the next state s_{t+1} for the agent. In addition, each state can be used by only one agent during the communication. So, the actions that lead to a transition from one state to these previously used states become invalid. It is worth noting that the Q-values for different states are stored in the server.

(iii) Action: For each tile, the wave can be transmitted with some predefined angles. Therefore, for each state, the action can be either steering the waves (agents) according to one of these valid angles or absorbing the waves. The action at time t is denoted as $a_t \in A$, where A is the action set. The server selects one of these valid angles according to the Q-values.

(iv) Reward: For each hop, that is, transition from one tile to another, we consider a negative reward (punishment) r_h. This punishment makes the agent reach the goal with fewer hops and therefore prevents the agent from oscillating. In addition, if the agent arrives at the goal by selecting a valid action, we consider a large reward denoted by r_g. On the other hand, if the path of an agent is blocked by an obstacle, the agent receives a large punishment r_o. So, we can write

$$r_{t+1}(s_t, a_t) = \begin{cases} r_h, & \text{if the agent goes to the next state,} \\ r_g, & \text{if the agent goes to the goal,} \\ r_o, & \text{if the path of agent is blocked by an obstacle.} \end{cases} \tag{2}$$

3.1. Multiagent Model. As previously mentioned, it is common to have more than one user communicating with the AP. For such a scenario, we should extend the above-mentioned single-agent RL algorithm to the multiagent one. A popular approach for a multiagent problem is Nash-Q learning [26, 27], in which the agents can be assumed to play a game to obtain maximum payoff. Instead of finding the maximum Q-value for each agent, the Nash Equilibrium property is employed to find the optimal set of actions for each state [27]. Nevertheless, this approach cannot be efficient for our problem because, in addition to a large number of users, the number of valid actions for each agent can be large. For example, if the numbers of agents and their valid actions are 10 and 12, respectively, then a set of 120 equations should be solved to obtain the Nash Equilibrium for the mixed strategy. Therefore, determining the Nash Equilibrium for such a problem with this amount of calculations, which should be repeated each time for transition from one state to another, is considerably time-consuming. Moreover, whenever a new user wishes to make a connection with the AP, the Nash Equilibrium should be calculated for all the users together, which can lead to a change in the state sequences of the previously connected users and therefore increase the delay for them.

Since it is important to reduce the delay of establishing communication between the transmitter and receiver as much as possible, we should consider another approach for our problem. We can simplify the model by considering that the users usually do not simultaneously request a connection at the same time. In other words, we can consider this problem as a single-agent model that is repeated for each user, and the environment is modified after each iteration. It is worth noting that this modification is needed because each tile can be used only by one agent. Therefore, at the end of each episode for every agent, some of the states are invalid for the other agents. In this model, all agents have a common Q-table and update it [27]. So, we can rewrite equation (1) by adding a subscript i to denote agent i and by defining the Q-values as a function of all agents' actions [27]:

$$\begin{aligned} Q\left(s_t^i, a_t^i\right) \leftarrow & Q\left(s_t^i, a_t^i\right) \\ & + \alpha\left[r_{t+1}^i\left(s_t^i, a_t^i\right) + \gamma \max_{a_t'^i} Q\left(s_{t+1}^i, a_t'^i\right) - Q\left(s_t^i, a_t^i\right)\right]. \end{aligned} \tag{3}$$

Algorithm 1 indicates the proposed reinforcement learning-based communication for the next generation of wireless networks. This algorithm takes the number of agents (n_a), the number of states (n_s), the acceptable error (e), and ϵ as input. Since each tile can only be used once to steer, we define set U to indicate the tiles that have been used before. In addition, ΔQ indicates the variation of the Q-table after finishing each episode and we set its initial value to some

Input: n_a:number of agents, n_s:number of states, e: acceptable error, ϵ
Output: Final Q table
$\varepsilon_i \longleftarrow \varepsilon$, $\forall i \in 1, 2, \ldots, n_a$
n_g: the set of agent which at the episode have not reached to the goal
$n_g \longleftarrow 1, 2, \ldots, n_a$
$S(s, a)$: return the next state corresponding to current state s and performing the action a
U:the set of states used by all of the agents
ΔQ: Variance of Q-table
$\Delta Q \leftarrow 2e$
while $\Delta Q > e$ **do**
  $U \leftarrow \varnothing$
  while $n_g \neq \varnothing$ **do**
    for $i \in n_g$ **do**
      generate a random number r in [0, 1]
      if $r > \epsilon_i$ **then**
        Select action of agent $i(a_i)$ according to Q-table
        if $S(s_i, a_i) \in U$ **then**
          select another action
        end
      end
      else
        Select action of agent $i(a_i)$ randomly
        if $S(s_i, a_i) \in U$ **then**
          select another action randomely
        end
      end
      Calculate $r_i(s_i, a_i)$ by equation (2)
      Update $Q(s_i, a_i)$ by equation (3)
      Update n_g
      $\epsilon_i \leftarrow \epsilon_i * c$
    end
  end
  Store $Q(s_i, a_i)$
  Calculate ΔQ
end

Algorithm 1: The process for finding the paths for multiple users in a PWE using the proposed RL algorithm.

value greater than e. For each episode, all the agents should reach the goal. Therefore, if the episode is finished for some of the agents, others should proceed to reach the goal. After all agents have reached the target, the episode is finished and the variation of the Q-values is calculated over a batch of recent episodes. If this variation is lower than a threshold e, the learning process has been completed. Each agent has its unique ε value that is equal to one at the beginning of the algorithm and then decreases after each agent's action by a constantly decreasing factor c.

Finally, we can predict the execution time of the proposed algorithm to configure the PWE. Suppose that R is the data rate of the wireless links, L is the length of pilot packets in *bits* sent to update Q-table before data transmission phase, d_p is the total propagation delay between the user and AP, and d_{proc} is processing delay per tile. If the average number of episodes required for the proposed RL algorithm to converge is N_{episode} and the average number of tiles activated to forward the EM waves for each user is n_{tile}, the total time needed before starting the data communication, denoted by T, can be estimated as follows:

$$T = \left[\frac{(n_{\text{tile}} + 1)L}{R} + d_p + n_{\text{tile}} d_{\text{proc}}\right] N_{\text{episode}}. \quad (4)$$

Because the next generations of wireless networks work at high frequencies, the value of R is considerably high. On the other hand, for an indoor environment, d_p will be negligible. We calculate T in Section 4.3 when some parameters such as n_{tile} and N_{episode} can be estimated.

4. Performance Evaluation

In this section, we evaluate adopting reinforcement algorithm to establish wireless communication links in an environment covered by SDM. We consider an indoor space, similar to that used in [9, 16], with the dimensions of 15 m × 10 m × 3 m. A wall with a length of 12 m and a thickness of 1 m exists in the middle of the room, which creates two separate sections, namely, line-of-sight (LOS) and non-line-of-sight (NLOS) regions. All the walls have been covered by 1 m × 1 m tiles. So, according to the dimensions of the room, there are 193 tiles. An AP working at 60 GHz with 100 dBmW transmitting power is

located at the height of 2 m somewhere in the LOS region. In our model, the tiles can steer or absorb the received EM waves from the directions formed by a combination of angles $\{-60^\circ, -45^\circ, -30^\circ, -15^\circ, 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ\}$ in the elevation plane and 0° to 360° with a step of 15° in the azimuth plane. We assume that no power is dissipated when the tile steers the wave and, on the other hand, all the power of the wave is absorbed while it is absorbing. To evaluate the performance of the proposed RL-based system, we develop a 3D ray-tracing simulator by Python.

We consider the reward function as equation (5) in which we have made differences between the next states in LOS and NLOS regions. If we assume that the EM waves propagate from the user in the NLOS region to the AP in the LOS region, we consider a larger value for r_{h2} to encourage the EM waves to enter the LOS region for arriving at the AP as soon as possible.

$$\text{Reward} = \begin{cases} r_{h1}, & \text{if the agent goes to the next state with } x > 5, \\ r_{h2}, & \text{if the agent goes to the next state with } x <= 5, \\ 100, & \text{if the agent goes to the goal,} \\ -100, & \text{if the path of agent is blocked by an obstacle.} \end{cases} \tag{5}$$

To obtain the optimal values for reward in our RL modeling, we calculate the average number of activated tiles in a scenario with three users located in NLOS region for different values of r_{h1} and r_{h2} . According to Figure 2(a), the minimum average number of activated tiles is achieved for $r_{h1} = -4$ and $r_{h2} = 6$. In addition, we search for the optimum values of γ and α according to a similar process. As can be seen in Figure 2(b), there is a lot to choose for these two parameters and we choose $\gamma = 0.9$ and $\alpha = 0.2$. The parameters related to the RL algorithm and the PWE which are used in the simulations have been listed in Tables 2 and 3, respectively.

In this section, we consider four different scenarios and evaluate our proposed method in comparison with the related works.

4.1. Scenario 1: Providing Coverage for the NLOS Region. An approach for providing coverage in an NLOS region is to put a large number of receivers in that area and adequately set the state of each tile. This setting is typically done by means of the optimization algorithms such as genetic algorithm (GA). For example, in [16], 12 receivers are uniformly distributed over the NLOS region and then, using the GA, the optimal tiles configurations are searched to maximize the minimum received power over these receivers. In this subsection, we use the proposed RL-based model to establish proper wireless links between these 12 receivers and the AP located at position $\{2.25, 3, 2\}$ m. So, we apply the proposed multiagent RL algorithm to properly configure the environment. Figure 3 shows the minimum, maximum, and average of the total received power by these receivers in comparison with the obtained results in [16]. According to this figure, our RL-based approach improves the mean and min values of the received powers up to 12% and 35%, respectively, after approximately 500 episodes. In addition, we consider a different number of randomly located users in the NLOS region and investigate the coverage ability of this scenario based on the prelearning, that is, the achieved Q-table corresponding to the previously mentioned 12 receivers. Table 4 shows the minimum, maximum, and average of the received power by the users. The mean values of the received powers are greater than 26.43 dBmW which obviously are large enough for the proper communication.

4.2. Scenario 2: Performance Evaluation with Varying Number of Users. In the second scenario, there is not any prelearning and the transmitted EM waves should find themselves the proper paths to their destinations. For such a scenario, we evaluate the performance of the system for a different number of users. Accordingly, we increase the number of users and put them randomly in the NLOS region. For each case, we calculate the max, mean, and min of the received powers and compare the results with those obtained in the first scenario (Table 4), as shown in Figure 4. According to this figure, when the PWE is particularly configured for some users with specific locations, as in Scenario 2, users generally receive more power compared to Scenario 1 in which the configuration is performed according to some predetermined receivers.

Moreover, we compare the results of the presented approach based on neural network in [9] with those obtained in our Scenario 2 for a single user as well as the regular propagation in a nonprogrammable environment. It is worth noting that, in this reference, the user can receive several signals, that is, five signals from different tiles. So, to do a fair comparison, we modify our approach by assuming five users at the same place, as shown in Figure 5, and reporting the summation of received powers by these users as the overall power. In addition, the transmitter works at 2.4 GHz with −30 dBmW transmitting power. Table 5 indicates that the proposed RL approach provides approximately the signal level reported in [9] for such an especial case.

4.3. Scenario 3: Performance Evaluation in Presence of Obstacles. In this subsection, we consider a more realistic problem in which some obstacles can block the path of the EM waves. In the first case, we evaluate the proposed system's performance by assuming that one, two, and three obstacles exist in the indoor environment, as can be seen in Figure 6. For each case, we repeated the simulation ten times. Figure 7 gives the average received power of all cases for a different number of users. According to this figure, by increasing the number of obstacles in the PWE, the average received power is generally decreased. Nevertheless, the average received power at least is around 21 dBmW which is an acceptable value in an indoor environment.

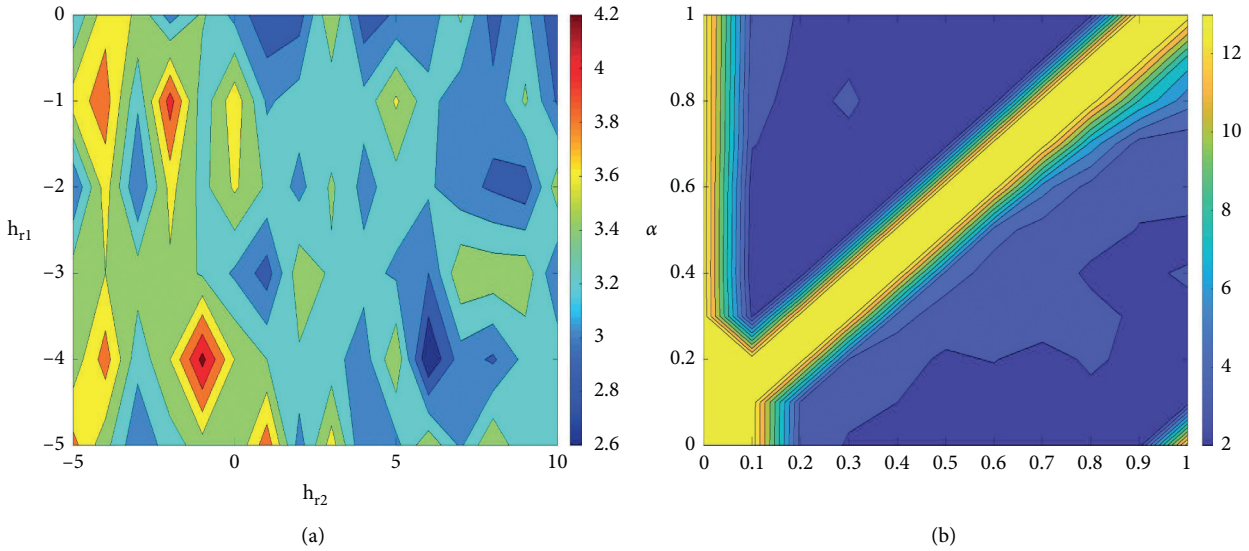

Figure 2: The average number of activated tiles for different values of (a) h_{r2} and h_{r1} and (b) γ and α.

Table 2: Parameters related to RL algorithm used in the simulation.

Parameters	Value
r_{h1}	−4
r_{h2}	6
r_g	100
r_o	−100
γ	0.9
α	0.2

Table 3: Parameters related to PWE used in the simulation.

Parameters	Value
Ceiling height	3 m
Tile dimensions	1 m × 1 m
Frequency	60 GHz
Tx power	100 dBmW
Antenna type	Omnidirectional with 2.5 dBi gain
Max ray bounces	20

Afterward, we investigate the ability of establishing wireless links between the users and the AP when some obstacles suddenly emerge after 500 episodes. Again, we consider three cases shown in Figure 6, with the difference that there are not any obstacles at the beginning of the communication, to evaluate the effect of environmental changes. Figure 8 illustrates the average received power of all cases for a different number of users. As we expected, the proposed RL-based approach adaptively configures the PWE, for the new conditions, to properly connect the AP and the receivers. In addition, we compare the number of episodes needed to converge our algorithm for both cases, namely, with and without environment changes, as indicated in Figure 9. Obviously, when the environment suddenly changes, the number of required episodes is increased.

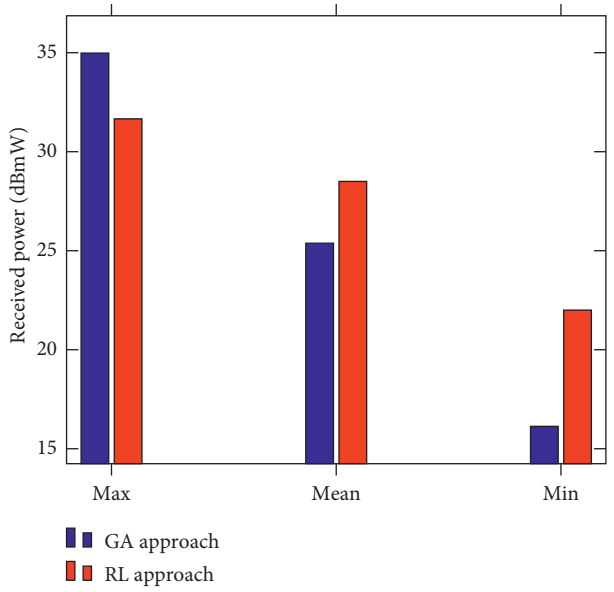

Figure 3: The max, mean, and min received powers by 12 receivers placed in the NLOS region calculated using GA and RL approaches.

Table 4: The minimum, maximum, and average of the total received power in dBmW by randomly placed users in Scenario 1, calculated according to the proposed RL approach.

Number of users	Max	Mean	Min
1	27.8	27.8	27.8
2	28.4	26.43	24.44
3	32.21	28.17	23.45
4	35.06	29.29	21.76
5	36.24	29.37	19.37

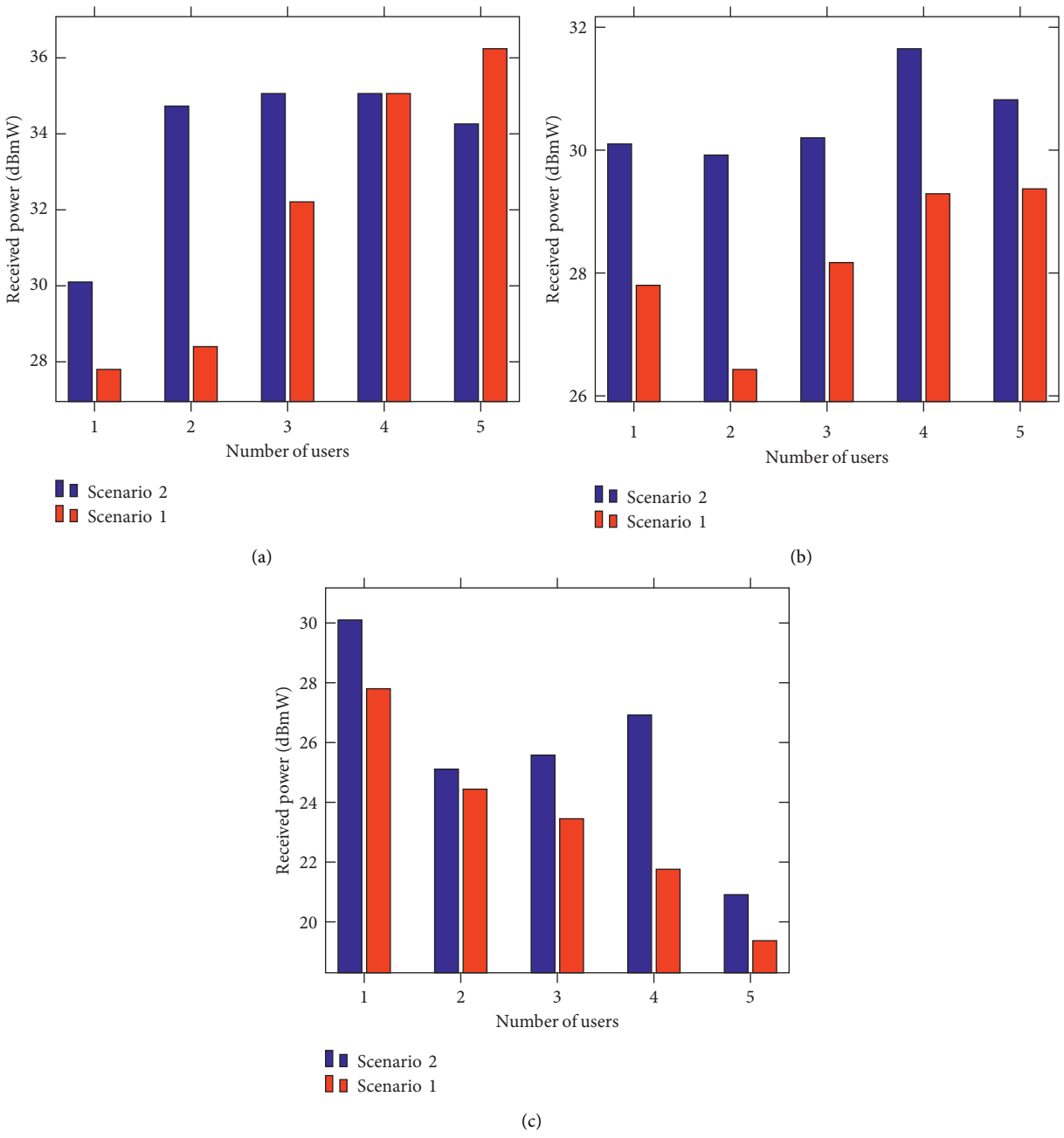

Figure 4: (a) Max, (b) mean, and (c) min of the received power for different number of users calculated by Scenario 2 in comparison with those obtained in Scenario 1.

4.4. Scenario 4: Performance Evaluation with Multipath Interference Cancellation. In a conventional wireless communication, the signal will reach the receiver not only via the direct path but also as a result of reflections from different objects. So, the overall signal at the receiver is a summation of the variety of signals being received. Since they all have different path lengths and therefore arrive with the different phases, the overall signal strength varies as a result of the different ways in which the signals will sum together. In this subsection, by using the obtained results in scenario 3, we can overcome the multipath phenomenon. In other words, we consider an obstacle at the place of each user to prevent other signals from passing through the receiver. Figure 10 shows two examples for such a scenario. Moreover,

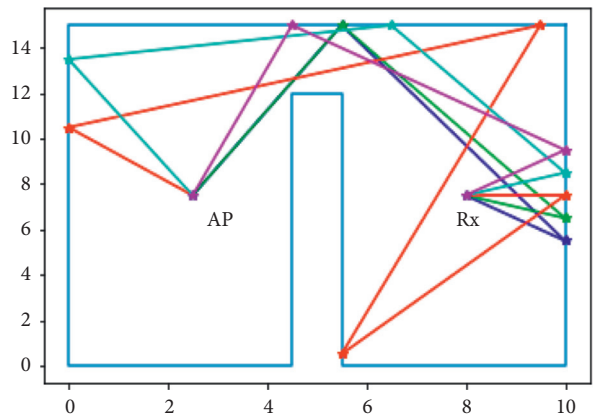

Figure 5: Trajectory of EM waves between a pair of receiver and transmitter according to Scenario 2. We have considered five users at the position (8, 7.5, 0.5). AP denotes access point.

Table 5: The total received power by a single user for three different approaches.

Approach	Received power (dBmW)
Regular propagation	−70.7
Neural network [9]	−49.87
This work	−51.7

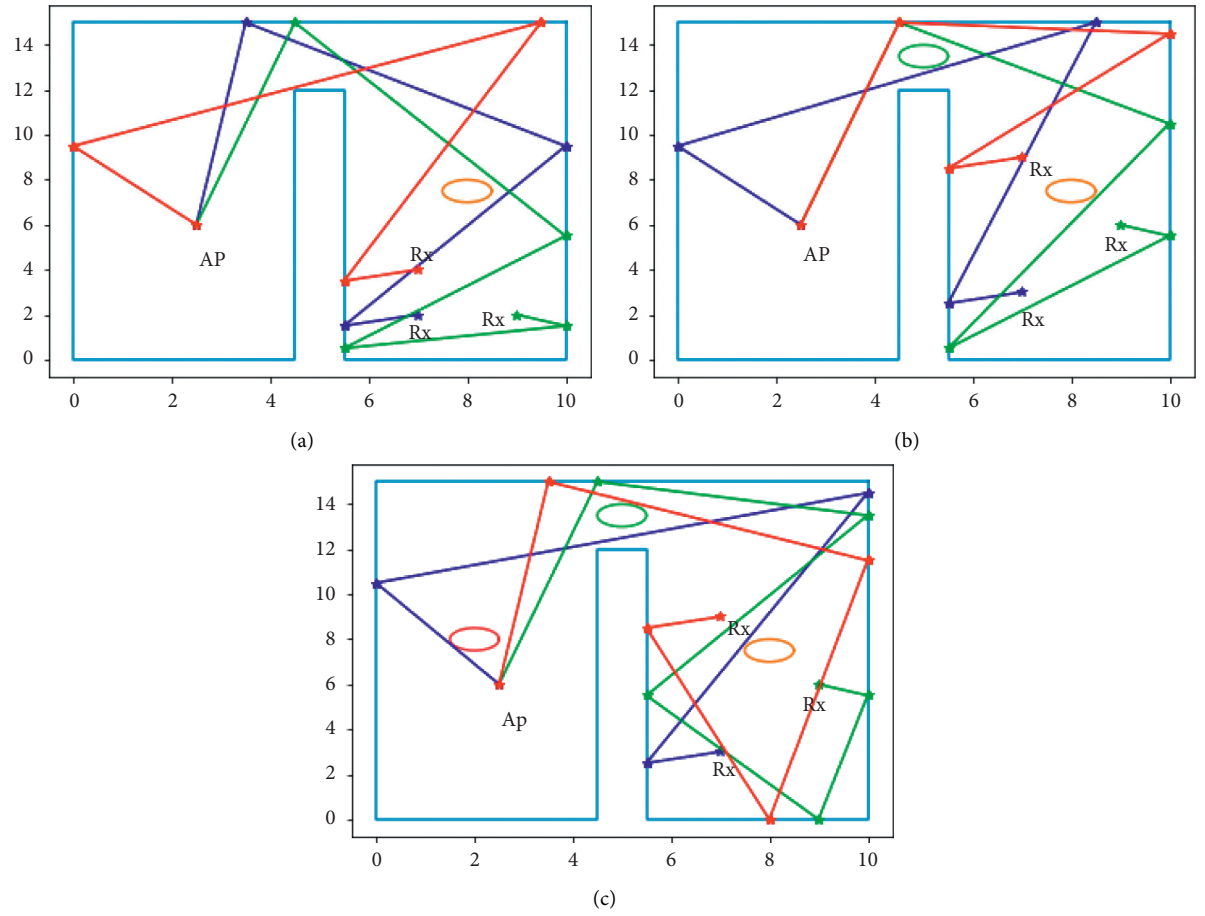

Figure 6: An example for wireless links in presence of (a) one obstacle (Case I), (b) two obstacles (Case II), and (c) three obstacles (Case III). AP denotes access point.

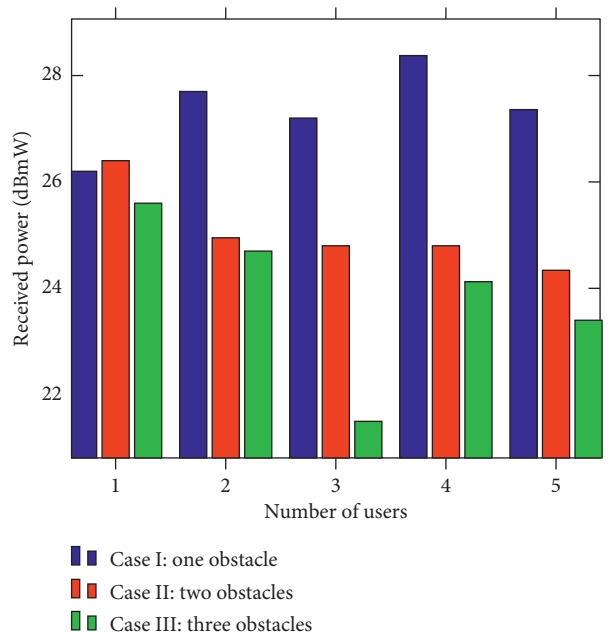

Figure 7: Comparison of the average received powers in presence of one, two, and three obstacles.

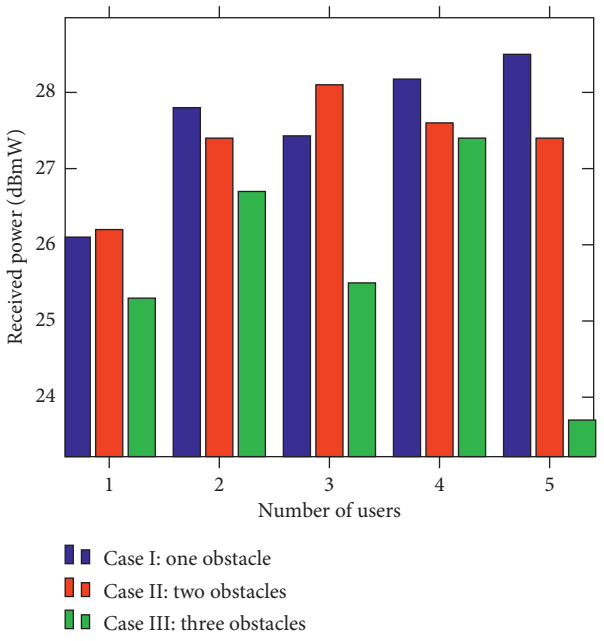

Figure 8: Comparison of the average received powers in presence of one, two, and three obstacles after environment changes.

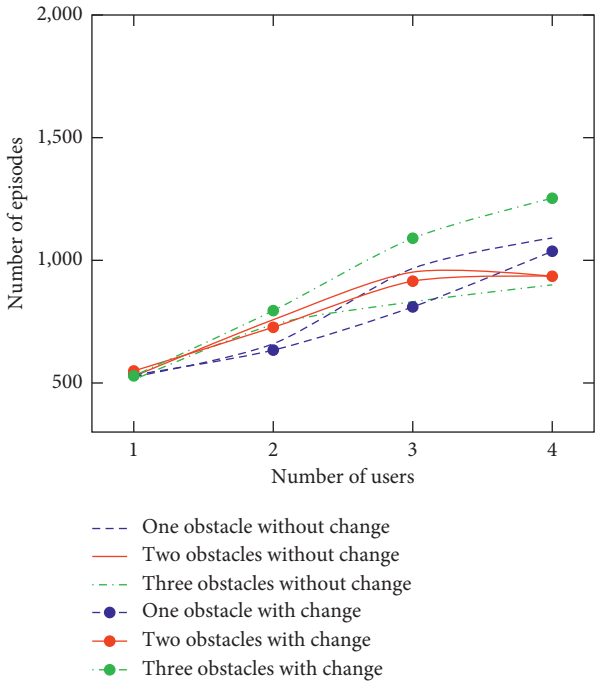

Figure 9: Comparison of the average received powers in presence of one, two, and three obstacles after environment changes.

Figure 11 illustrates the average received power as well as the number of required episodes for a different number of users for this scenario.

4.5. Time Complexity. Finally, predicting the execution time of each episode is worth noting. Suppose that the data rate of wireless links is about 100 Mbps and we use some 1000 bits pilot packets to update the Q-table as the initial phase of the communication. If we neglect the propagation delay (that is an acceptable assumption for an indoor communication) and consider 1 μs for processing delay in each tile, the time needed for a pilot packet to

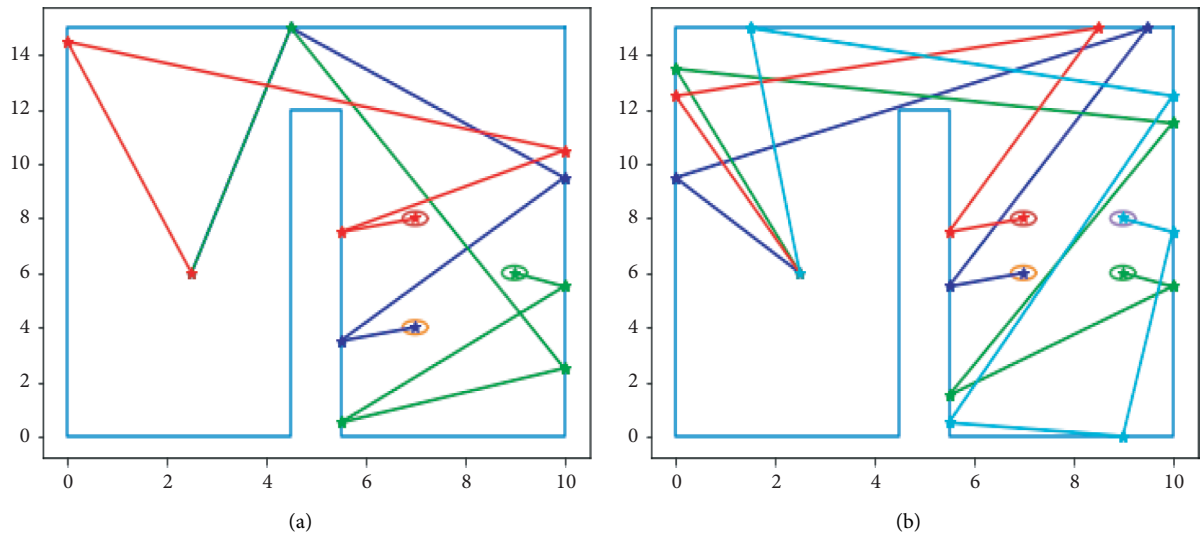

Figure 10: Two examples for Scenario 4. (a) Three users; (b) four users.

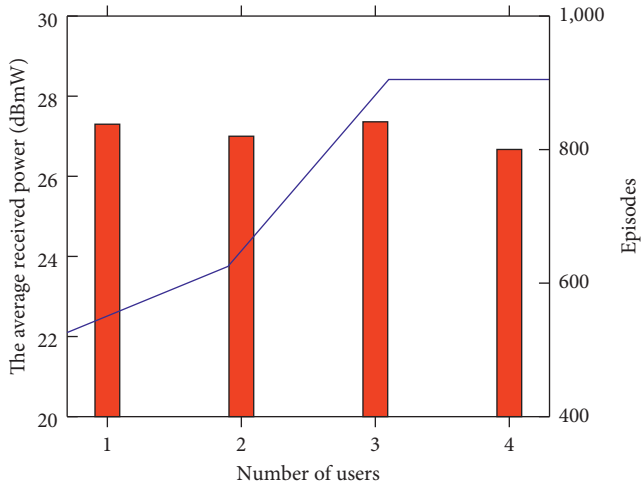

Figure 11: The average received power (red bar chart) as well as the number of episodes (solid blue curve) for Scenario 4.

arrive at the destination is about 55 μs. It should be noted that we consider that the average number of tiles activated to forward the EM waves is five (Figure 2). Now, if the number of episodes needed for converging our algorithm is 1000, according to equation (4), the total time for starting a communication will be about 55 ms.

5. Conclusion

We proposed an approach based on reinforcement learning (RL) to adaptively configure the indoor wireless communication environment in this paper. The proposed approach can be applied to a multiple user problem as well as a changing environment. To evaluate the performance of our method, we considered four different scenarios, namely, providing permanent coverage for the NLOS region, providing particular connections for a different number of users, providing connections in presence of obstacles, and finally providing connections after the sudden emergence of some obstacles. Our evaluation indicates that the proposed RL-based approach can properly provide wireless links for all the above-mentioned scenarios with enough signal level. The main concern about using the RL approach is the time required for updating the Q-table as the initial phase of the communication before starting the data communication. Our calculation shows that this delay time at the initial phase of communication for an indoor environment working at high frequencies can be negligible. It should be noted that when the environment is constantly changing, the time needed for updating the Q-table may disrupt the data communication.

As future work, we decide to use deep reinforcement learning (DRL), instead of the conventional RL algo-

rithm, to extend our approach for configuring more complicated environments. To this end, we can consider the IDs of the activated tiles and the current tile as well as the AP location as the neural network inputs. On the other hand, the output of the network can determine the reflection angle (proper action) to forward the EM waves to the next tile.

Data Availability

No data were used to support this study.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to declare.

References

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6 g: ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[2] M. J. Piran, N. H. Tran, D. Y. Suh, J. B. Song, C. S. Hong, and Z. Han, "Qoe-driven channel allocation and handoff management for seamless multimedia in cognitive 5 g cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6569–6585, 2016.

[3] A. Shahraki, M. Abbasi, M. Piran, M. Chen, S. Cui et al., "A Comprehensive Survey on 6G Networks: Applications, Core Services, Enabling Technologies, and Future Challenges," 2021, http://arxiv.org/abs/2101.12475.

[4] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[5] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmwave) for 5G: opportunities and challenges," *Wireless Networks*, vol. 21, no. 8, pp. 2657–2676, 2015.

[6] M. Numan, F. Subhan, W. Z. Khan et al., "A systematic review on clone node detection in static wireless sensor networks," *IEEE Access*, vol. 8, pp. 65 450–65 461, 2020.

[7] Q.-V. Pham, S. Mirjalili, N. Kumar, M. Alazab, and W.-J. Hwang, "Whale optimization algorithm with applications to resource allocation in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4285–4297, 2020.

[8] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, "Imcfn: image-based malware classification using fine-tuned convolutional neural network architecture," *Computer Networks*, vol. 171, Article ID 107138, 2020.

[9] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "An interpretable neural network for configuring programmable wireless environments," ", IEEE, in *Proceedings of the 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, Cannes, France, July 2019.

[10] R. Chandra and K. Winstein, "Programmable radio environments for smart spaces-hotnets-xvi dialogue," in *Proceedings of the ACM Workshop on Hot Topics in Networks*, Palo Alto, November 2017.

[11] A. Welkie, L. Shangguan, J. Gummeson, W. Hu, and K. Jamieson, "Programmable radio environments for smart spaces," in *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, pp. 36–42, Palo Alto, CA, USA, November 2017.

[12] M. A. Sheikh and S. A. Khan, "Performance optimization of unit-cell reflectarray antenna for future 5 g communications," *International Journal of Computer Applications*, vol. 975, p. 8887, 2017.

[13] L. Subrt and P. Pechac, "Intelligent walls as autonomous parts of smart indoor environments," *IET Communications*, vol. 6, no. 8, pp. 1004–1010, 2012.

[14] X. Tan, Z. Sun, J. M. Jornet, and D. Pados, "Increasing indoor spectrum sharing capacity using smart reflect-array," in *Proceedings of the 2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, May 2016, Kuala Lumpur, Malaysia.

[15] Q. Wu and R. Zhang, "Beamforming optimization for intelligent reflecting surface with discrete phase shifts," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7830–7833, IEEE, Brighton, UK, May 2019.

[16] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A novel communication paradigm for high capacity and security via programmable indoor wireless environments in next generation wireless systems," *Ad Hoc Networks*, vol. 87, pp. 1–16, 2019.

[17] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "On the network-layer modeling and configuration of programmable wireless environments," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1696–1713, 2019.

[18] H. Taghvaee, A. Cabellos-Aparicio, J. Georgiou, and S. Abadal, "Error analysis of programmable metasurfaces for beam steering," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 1, 2020.

[19] W. Tang, X. Li, J. Y. Dai et al., "Wireless communications with programmable metasurface: transceiver design and experimental results," *China Communications*, vol. 16, no. 5, pp. 46–61, 2019.

[20] G. Fairweather, A. Karageorghis, and P. A. Martin, "The method of fundamental solutions for scattering and radiation problems," *Engineering Analysis with Boundary Elements*, vol. 27, no. 7, pp. 759–769, 2003.

[21] S. Abadal, C. Liaskos, A. Tsioliaridou et al., "Computing and communications for the software-defined metamaterial paradigm: a context analysis," *IEEE Access*, vol. 5, pp. 6225–6235, 2017.

[22] M. Di Renzo, M. Debbah, D.-T. Phan-Huy et al., "Smart radio environments empowered by reconfigurable ai metasurfaces: an idea whose time has come," *EURASIP Journal on Wireless Communications and Networking*, vol. 1, pp. 1–20, 2019.

[23] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[24] C. Zhang and Z. Zheng, "Task migration for mobile edge computing using deep reinforcement learning," *Future Generation Computer Systems*, vol. 96, pp. 111–118, 2019.
[25] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
[26] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: an overview," in *Innovations in Multi-Agent Systems and Applications-1*, pp. 183–221, Springer, Berlin, Germany, 2010.
[27] Y. Shoham, R. Powers, and T. Grenager, "Multi-agent reinforcement learning: a critical survey," *Web Manuscript*, 2003.