IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 March 2023, accepted 2 May 2023, date of publication 11 May 2023, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3275567

RESEARCH ARTICLE

Secure Relay Selection With Outdated CSI in Cooperative Wireless Vehicular Networks: A DQN Approach

ESRAA M. GHOURAB[1], LINA BARIAH[1,2], (Senior Member, IEEE), SAMI MUHAIDAT[1,3], (Senior Member, IEEE), PASCHALIS C. SOFOTASIOS[1,4], (Senior Member, IEEE), MAHMOUD AL-QUTAYRI[5], (Senior Member, IEEE), AND ERNESTO DAMIANI[1], (Fellow, IEEE)

[1]KU Center for Cyber-Physical Systems, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates
[2]Technology Innovation Institute, Abu Dhabi, United Arab Emirates
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
[4]Department of Electrical Engineering, Tampere University, 33720 Tampere, Finland
[5]Systems-on-Chip (SoC) Center, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates

Corresponding author: Paschalis C. Sofotasios (p.sofotasios@ieee.org)

ABSTRACT Cooperative communications is a core research area in wireless vehicular networks (WVNs), thanks to its capability to provide a certain degree of fading mitigation and to improve spectral efficiency. In a cooperative scenario, the intercept probability of the system can be reduced by optimizing the relay selection scheme in order to select the optimal relay from a set of available relays for data transmission. However, due to the mobility of WVNs, the best relay is often selected in practice based on outdated channel state information (CSI), which in turn affects the overall system performance. Therefore, there is a need for a robust relay selection scheme (RSS) that guarantees a satisfactory overall achievable performance in the presence of an outdated CSI. Motivated by this and considering the advantageous features of autoregressive moving average (ARMA), the proposed contribution models a cooperative vehicular communication scenario with relay selection as a Markov decision process (MDP) and proposes two deep Q-networks (DQNs), namely DQN-RSS and DQN-RSS-ARMA. In the proposed framework, two deep reinforcement learning (RL)-based RSS are trained based on the intercept probability, aiming to select the optimal vehicular relay from a set of multiple relays. We then compare the proposed RSS with the conventional methods and evaluate the performance of the network from the security point of view. Simulation results show that DQN-RSS and DQN-RSS-ARMA perform better than conventional approaches, as they reduce intercept probability by approximately 15% and 30%, respectively, compared to the standard ARMA approach.

INDEX TERMS Secrecy capacity, cooperative communication, deep Q-network, outdated channel state information, reinforcement learning, relay selection.

I. INTRODUCTION

Fifth-generation and beyond (5G/B5G) wireless networks are considered for meeting the ever-increasing demand for global communication services and broad wireless coverage for a particularly large number of users [1]. Based on this and due to the inherent nature of the wireless environment along with the presence of several low-power wireless devices, it calls for sophisticated methods that will ultimately ensure the secure operation of B5G [2]. This becomes even more urgent because conventional cryptographic approaches are expected to be insufficient for the new technologies in B5G, and other security requirements are urgently needed [3]. In this context, physical layer security (PLS) has attracted considerable

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

attention in recent years thanks to its capability to enhance the overall security of wireless communication systems. PLS exploits the random nature of the wireless channel between transmitters and receivers, such as dedicated channel state information (CSI), to secure transmissions at the legitimate receiver as well as to degrade the quality of the received signal at the eavesdropper [4]. The core advantage of PLS methods is that they do not rely on encryption and decryption operations, so they overcome the difficulties of distributing and managing secret keys in extremely dense and large-scale heterogeneous networks. Moreover, compared to encryption-based methods, PLS approaches only need to execute relatively simple signal processing algorithms, resulting in a smaller overhead. Therefore, it can be concluded that the use of PLS in conjunction with traditional cryptographic methods could provide an additional layer of security that will further protect transmissions in wireless networks [5], [6].

As a kind of cooperative wireless network, vehicular relay networks are regarded as efficient wireless networks because they can reduce energy consumption, extend transmission range, and improve the achievable throughput [7]. Furthermore, PLS in cooperative vehicular relay networks has received considerable attention from both academia and industry. It is also recalled that in cooperative vehicular networks, the transmission overhead associated with CSI estimation depends on the number of relay nodes. Consequently, selecting the best relay based on the instantaneous CSI is challenging, so effective relay selection (RS) is still an open research area in dynamic wireless vehicular networks (WVNs) that require a real-time strategy. Particularly, in dynamic WVN, CSI for the best relay is usually outdated due to the channel feedback delay, so an efficient RS strategy that can cope with outdated CSI is required in WVN scenarios.

Recently, researchers used a channel delay model to characterize CSI inaccuracy and proposed a set of robust RS methods [8], [9]. Yet, while the delay model can develop a robust RS strategy, it cannot adapt to changes, which affects the achieved secrecy performance. As a method to solve this problem, researchers have proposed a self-learning RS scheme using the Q-learning scheme [10], [11].

Q-learning is a model-free reinforcement learning (RL) algorithm developed by the Markov decision process (MDP), which uses an iterative criterion to reach the optimal solution. Accordingly, a source node with learning capabilities can select the best relay for cooperative communication according to the prior system performance and the reward function of the observed state [11]. However, the Q-learning-based RS method stores the Q-value in a Q-table; hence, it can only solve the problem with a small state space. To this effect, since the storage capacity of a Q-table is limited, it cannot cover the entire state space when it is large. Therefore, the conventional RL algorithm cannot solve adequately such problems. As a result, Google DeepMind added deep learning (DL) to RL and developed the deep Q-network (DQN) aiming to solve this problem [12]. In particular, the DQN approach leverages the perceptual capabilities of DL to enable the RL algorithm to extract environmental features and solve associated Q-table problems.

A. RELATED WORK

The RS problems in relay-based cooperative communications have been analyzed extensively because RS is an effective method for increasing the communication range as well as improving the communication quality in emerging wireless networks [13], [14]. In wireless sensor networks (WSNs), for example, the authors in [15] proposed an adaptive forwarding-based RS approach, while [16] developed an energy-efficient cooperative communication model to improve the data transmission performance. In [17], the authors proposed several RS methods to improve security and counter strong eavesdroppers. Likewise, authors in [18] studied a cooperative wireless network with multiple relays, multiple eavesdroppers, and a transmitter-receiver pair to improve the overall system security and confirmed the network performance based on the achieved secrecy rate (SR) and secrecy outage probability (SOP). Moreover, the authors in [19] derived asymptotic and closed-form expressions for SOP aiming to maximize the considered energy harvesting for the case of partial/optimal RS approaches. In addition, the first game-theoretic RS approach that enhances PLS by selecting the optimal relay was proposed in [20]. Finally, the authors in [21] investigated RS techniques for a cooperative dual-hop network that uses the amplify-and-forward (AF) protocol in terms of reliability and security.

In general, the effectiveness of RS depends on numerous variables, including the availability and quality of channel characteristics. In particular, due to the time-varying nature of fading channels, there might be a time delay between relay selection and data transmission, rendering CSI to be outdated, which in turn degrades the overall network performance [22]. To this effect, two approaches are capable of addressing this problem: the simple approach is to use outdated CSI when selecting the best relay; while the second approach is to use the predicted CSI based on estimated measurements when selecting the best relay [23]. It is recalled here that by performing channel estimation, wireless communication systems can obtain an accurate CSI [24]. Therefore, researchers have developed several channel estimators, including the least square (LS) [25], maximum likelihood (ML) [26], and minimum mean-square error (MMSE) [27]. It is also noted that Jake's tap-gain method has proven to be a reliable choice for modeling Rayleigh fading channels [28]. In the same context, autoregressive (AR) and autoregressive moving average (ARMA) methods were employed to represent the Rayleigh variables of a time-varying channel [29]. For instance, the authors in [30] predicted V2V channels using a traditional low-complexity approach, i.e., an AR-based prediction approach.

It has been evident that the demand for learning wireless networks has grown exponentially in recent years [31]. For wireless networks, machine learning (ML) has been

proposed for spectrum sensing, resource allocation, and intrusion detection applications, to name but a few [31]. Therefore, in order to meet the requirements of new wireless technologies and improve RS methods, intelligent relay nodes are created by integrating ML into RS. For example, by combining PLS, RS, and ML, the authors in [32] transformed the RS problem into a multi-class classification problem, whereas, the analysis in [33] transformed the RS approaches into a prediction and decision-making problem. However, it must be noted that the availability of a large amount of historical data for training the ML algorithm is challenging, especially in rapidly changing fading environments.

In cooperative networks, the RL agent chooses the best relay in each time slot depending on the previous state observation and the action-reward feedback from the communication system. In this context, authors in [34] proposed a single-agent RL-based RS scheme for WSNs. Likewise, the authors in [35] employed unmanned aerial vehicles (UAVs) to relay an on-board unit's message and improve the bit error rate (BER) of vehicular ad-hoc networks (VANETs) against jammers, using an RL framework. In [36], the authors proposed a DQN-based RS scheme that combines a deep neural network (DNN) with the typical Q-learning algorithm to minimize the outage probability (OP), whereas authors in [45] developed a decision-based deep RL (DRL) approach to support RS as well as to improve the overall system performance. In addition, [38] considered an RS-based Q-learning approach to improve energy efficiency, whereas [39] divided RS and power optimization problems into two subproblems and solved them with a hierarchical RL architecture aiming to maximize the overall signal-to-noise ratio (SNR) and minimize the corresponding OP. Finally, in [40], the authors developed asynchronous DRL approaches to maximize the system throughput, whereas the authors in [41] proposed a dual DQN architecture to minimize the involved transmission delay.

B. CONTRIBUTIONS

Based on the above observations, this work proposes intelligent predictive algorithms for RS that maximize the security of vehicular wireless cooperative dual-hop networks in the presence of outdated CSI. The main contribution of this work is summarized below:

- We design DQN-RSS, which is a deep-Q-learning-based RS scheme for WVNs. In DQN-RSS, an optimal relay is selected from a plurality of relay candidates according to the corresponding intercept probability.
- We develop a DQN-RSS-ARMA for our system model, where the agent aims to use the estimated CSIs to predict the new one, select the optimal relay and then compare it with the DQN-RSS. The results show that DQN-RSS-ARMA achieves lower intercept probability and thus higher secrecy capacity compared to DQN-RSS.
- We compare the intercept probability of different approaches for relay selection, such as baseline, and ARMA, as well as approaches based on different combinations, such as participating relays and past outdated CSIs.

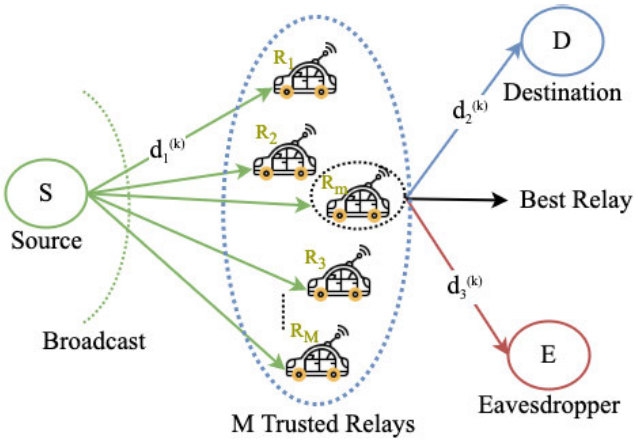


FIGURE 1. The RS strategy in a wireless vehicular network (WVNs).

To the best of the author's knowledge, the listed contributions have not been previously reported in the open literature.

The rest of the paper is organized as follows. The detailed system model is presented in section II. Section III presents the secrecy analysis of a cooperative model. Section IV presents the proposed Markov decision process formulation for relay selection. Section V illustrates the proposed intelligent relay selection algorithm. Section VI and Section VII present the performance metrics and discussions, and Section VIII concludes the paper.

II. SYSTEM MODEL

A. NETWORK MODEL

As shown in Fig. 1, we consider a vehicular wireless system with a source node (S), a destination node (D), and M trusted relays $R_m, m = 1, \ldots, M$, in the presence of an active eavesdropper. Any transmission for the $S - D$ link is possible only with the aid of one relay since a direct link between S and D is not available. Each node in the network is equipped with a single antenna and operates in half-duplex mode, whereas all channels in the network are assumed to experience independent Rayleigh fading.

We also let $[\boldsymbol{d}_{i,m}^{(k)} = [d_{1,m}^{(k)}, d_{2,m}^{(k)}, d_{3,m}^{(k)}]$ denote the distance vector at time k, where $d_{1,m}^{(k)}$ is the distance between the source S and relay m, $d_{2,m}^{(k)}$ is the distance between relay m and the destination D, and $d_{3,m}^{(k)}$ is the distance between the relay m and the eavesdropper E. To this effect, we also naturally consider that the distances $d_{2,m}^{(k)}$ change when the vehicular relays move. Therefore, the total distance matrix representing the distances from S to the relays, from the relays to D, and from the relays to E is defined by $\mathbf{D}_i^{m,n}$, of size $(M \times N)$, $i = 1, 2, 3,\ n = 1, 2, \ldots N,\ m = 1, \ldots, M$, column vectors are $d_{i,m}^{(k)}$, at time k.

The communication between S and D is realized in two-time slots. In the first time slot, S broadcasts the signal (X) to M relays with a transmission power of P_S, while in the

second time slot the best relay m forwards the transmitted information to D with a transmission power of P_R.

B. CHANNEL MODEL

Throughout the paper the channel gain vector is defined by $\mathbf{g}_{i,m}^{(k)} = [g_{1,m}^{(k)}, g_{2,m}^{(k)}, g_{3,m}^{(k)}]$ are the channel gains of the S-relay m, the relay m-D, and relay m-E in time slot k, respectively. The total channel gain matrix, which is composed of the previous data (i.e., outdated CSI) in all links, is represented by $\mathbf{G}_i^{m,n}, i = 1, 2, 3, \ n = 1, 2, \ldots N, \ m = 1, \ldots, M$ at time k, of size $(M \times N)$, as described in (1), as shown at the bottom of the page. The column elements of $\mathbf{G}_i^{m,n}$ at time k are defined as

$$g_i^{m,n} = \frac{|h_i^{m,n}|^2}{(d_i^{m,n})^a}, \quad i = 1, 2, 3, \\ m = 1, 2, ..M, \quad n = 1, 2, \ldots N \tag{2}$$

It is worth mentioning that $\boldsymbol{h_i^{m,n(k)}} = [\ h_1^{m,n(k)}, \ h_2^{m,n(k)}, \ h_3^{m,n(k)}]$ represents the channel coefficients of the S-relays, relays-D, and relays-E, links respectively, in time slot k. The channel coefficients include path loss, fading, and shadowing effects, whilst a represents the associated path loss exponent. Also, the system noise is modeled as additive white Gaussian noise with zero mean and variance of σ^2. In the time instant k, the SNR of the signals transmitted from S-relays, relays-D, and relays-E, are represented by a matrix $\boldsymbol{\gamma}$ of size $(M \times N)$ in (3), as shown at the bottom of the page. The entries of $\boldsymbol{\gamma}$ are: $\gamma_1^{m,n(k)} = \frac{P_S \, g_1^{m,n(k)}}{\sigma^2}, \ \gamma_2^{m,n(k)} = \frac{P_R \, g_2^{(k)}}{\sigma^2}, \ \gamma_3^{m,n(k)} = \frac{P_R \, g_3^{(k)}}{\sigma^2}$.

C. CHANNEL PREDICTION

In order to achieve reliable communication over channels with rapid time-varying characteristics, it is essential to minimize feedback delays and estimation errors. In particular, feedback delay causes CSI to become obsolete and degrades the precoder performance, especially in fast time-varying channels. This is also the case in wireless networks, such as vehicle-to-everything (V2X), which are characterized by dynamic environments and high mobility [42]. Therefore, channel prediction has been proposed as a potentially effective solution to this problem.

It is recalled that channel prediction techniques can be mainly classified into the parametric radio channel (PRC) model, the autoregressive (AR) model, and the basis expansion model (BEM) [43]. The PRC method represents a time-varying channel as a sum of complex sinusoids. These parameters are then estimated based on the known channel coefficients and are used for channel prediction. Traditional AR methods, on the other hand, predict the channel as a linear combination of the known channel coefficients using a linear or nonlinear MMSE filter. Consequently, AR depends on the correlation matrix of the channel [44]. However, conventional AR schemes are ineffective if the correlation function is unknown or if it varies over time. To this effect, adaptive AR schemes based on adaptive filtering techniques such as recursive least squares (RLS), least mean squares (LMS), and Kalman filtering have been developed to address this issue.

Based on the above, the best relay selected at time t according to the outdated CSI may not be the best relay at time $(t + \tau)$ of data transmission. In [45], the degree of mismatch is calculated by the correlation coefficient (ρ_o) between the outdated channel $(\hat{h}_n)$ at time t and the actual channel (h_n) at time $(t + \tau)$, where $0 < \rho_o < 1$. The expression for this correlation function is given by:

$$\rho_o = \frac{E\{\hat{h}_n h_n\}}{\sqrt{E\{|\hat{h}_n|^2\}E\{|h_n|^2\}}} \tag{4}$$

where $(\hat{h}_n)$ is the obsolete channel with variance $\sigma_{\hat{n}}$ and zero mean, which can be expressed as

$$\hat{h}_n = \sigma_{\hat{n}} \left(\frac{\rho_o}{\sigma_n} h_n + \epsilon \sqrt{1 - \rho_o^2} \right) \tag{5}$$

where ϵ is a random variable with a standard normal distribution and zero means. Using the traditional Jake's model with the assumption that the delay between the outdated and

$$G = \begin{pmatrix} g_1^{1,1} & g_1^{1,2} & \cdots & g_1^{1,N} & g_2^{1,1} & g_2^{1,2} & \cdots & g_2^{1,N} & g_3^{1,1} & g_3^{1,2} & \cdots & g_3^{1,N} \\ g_1^{2,1} & g_1^{2,2} & \cdots & g_1^{2,N} & g_2^{2,1} & g_2^{2,2} & \cdots & g_2^{2,N} & g_3^{2,1} & g_3^{2,2} & \cdots & g_3^{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1^{M,1} & g_1^{M,2} & \cdots & g_1^{M,N} & g_2^{M,1} & g_2^{M,2} & \cdots & g_2^{M,N} & g_3^{M,1} & g_3^{M,2} & \cdots & g_3^{M,N} \end{pmatrix} \tag{1}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1^{1,1} & \gamma_1^{1,2} & \cdots & \gamma_1^{1,N} & \gamma_2^{1,1} & \gamma_2^{1,2} & \cdots & \gamma_2^{1,N} & \gamma_3^{1,1} & \gamma_3^{1,2} & \cdots & \gamma_3^{1,N} \\ \gamma_1^{2,1} & \gamma_1^{2,2} & \cdots & \gamma_1^{2,N} & \gamma_2^{2,1} & \gamma_2^{2,2} & \cdots & \gamma_2^{2,N} & \gamma_3^{2,1} & \gamma_3^{2,2} & \cdots & \gamma_3^{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_1^{M,1} & \gamma_1^{M,2} & \cdots & \gamma_1^{M,N} & \gamma_2^{M,1} & \gamma_2^{M,2} & \cdots & \gamma_2^{M,N} & \gamma_3^{M,1} & \gamma_3^{M,2} & \cdots & \gamma_3^{M,N} \end{pmatrix} \tag{3}$$

the current CSI is (τ) and the maximum Doppler frequency is (f_d), the ρ_o value is expressed as

$$\rho_o = J_o(2\pi f_d \tau) \tag{6}$$

where $J_o(.)$ denotes the zero-order Bessel function of the first kind. In addition, f_d can be calculated according to:

$$f_d = \frac{v f_c}{c} \tag{7}$$

where v is the vehicle speed, which in our case is between (0-80) km/h, whereas f_c denotes the carrier frequency and c is the speed of light.

III. SECRECY ANALYSIS

Selecting a single relay m for information transmission from candidate relays improves the secrecy capacity of the system [46]. Therefore, in the present analysis, we select the optimal relay for transmission, taking into consideration the outdated CSIs. In particular, in the dual-hop communication process, S broadcasts information to all relays in the first hop, and then the selected relay m forwards the information to D in the second hop. Due to the broadcast nature of wireless communication, an eavesdropper could potentially intercept the data transmitted by this relay m. Therefore, the received signal at the m^{th} relay for the first hop is written as follows:

$$y_m = \sqrt{P_S}\ \ g_1^{(k)}\, X + n_m \tag{8}$$

Based on this, the signal received from node D and node E during the second hop can be represented as

$$y_D = \sqrt{P_R}\ \ g_2^{(k)}\, y_m + n_D \tag{9}$$

and

$$y_E = \sqrt{P_R}\ \ g_3^{(k)}\, y_m + n_E \tag{10}$$

respectively.

Therefore, the channel capacity of the channel in time slot k for the S-m link is given by

$$C_{S,m}^{(k)} = \frac{1}{2}\log_2\left(1 + \frac{P_S\ g_1^{m,n(k)}}{\sigma^2}\right) = \frac{1}{2}\log_2\left(1 + \gamma_1^{(k)}\right) \tag{11}$$

Without loss of generality, we assume that the considered cooperative system operates according to the AF relay protocol and that the S and D nodes are static, and has an amplification factor = 1. Also, all relays are moving in the same direction but at different velocities. Finally, the eavesdropper is active and we can locate it. Similar to [34] and [47], if the relay m is assigned to cooperate, the channel capacities for m-D and m-E are expressed as

$$C_{m,D}^{(k)} = \frac{1}{2}\log_2\left(1 + \frac{\gamma_1^{(k)}\ \gamma_2^{(k)}}{\gamma_1^{(k)} + \gamma_2^{(k)} + 1}\right) \tag{12}$$

and

$$C_{m,E}^{(k)} = \frac{1}{2}\log_2\left(1 + \frac{\gamma_1^{(k)}\ \gamma_3^{(k)}}{\gamma_1^{(k)} + \gamma_3^{(k)} + 1}\right) \tag{13}$$

respectively.

It is recalled that the secrecy capacity (C_s) and the intercept probability (IP) are the two basic performance metrics for PLS [13]. The achievable secrecy capacity is calculated as the difference between the secrecy capacity of the legitimate link and the eavesdropper link, namely

$$C_s = [C_{m,D} - C_{m,E}]^+ \tag{14}$$

where C_s is the secrecy capacity, $[x]^+ = \max(0, x)$. Based on the above, it follows that the secrecy capacity can be readily determined with the aid of (12), (13), and (14), yielding

$$C_s = \frac{1}{2}\log_2\left(\frac{1 + \dfrac{\gamma_1^{(k)}\ \gamma_2^{(k)}}{\gamma_1^{(k)} + \gamma_2^{(k)} + 1}}{1 + \dfrac{\gamma_1^{(k)}\ \gamma_3^{(k)}}{\gamma_1^{(k)} + \gamma_3^{(k)} + 1}}\right) \tag{15}$$

A. RELAY SELECTION

Based on the instantaneous end-to-end (EE) SNR of the m^{th}-relay in (12), the optimal selection policy activates the m relay, where:

$$\begin{aligned} m &= \arg\max_{m\in M} \gamma_{smd} \\ &= \arg\max_{m\in M} \frac{\gamma_1^{(k)}\ \gamma_2^{(k)}}{\gamma_1^{(k)} + \gamma_2^{(k)} + 1} \end{aligned} \tag{16}$$

with γ_{smd} denoting the end-to-end (EE) SNR from S to D via m. Notably, due to the mobility of the involved vehicles, the instantaneous SNR used for RS is an outdated version of (16). Therefore, our model predicts the CSI based on the estimated values to select the optimal relay. In particular, the proposed scheme selects the best relay m^* as $m^* = \arg\max_{m\in M} \hat{\gamma}_m$, where $\hat{\gamma}_m$ is the predicted CSI for the S-m and m-D transmission.

B. INTERCEPT PROBABILITY

It is recalled that the intercept probability occurs when the capacity of the main link drops below that of the wiretap link. Therefore, when m is chosen as the optimal relay for cooperative vehicular communications from (16), the intercept probability based on (15) can be expressed as

$$\begin{aligned} P_m^{(k)}(D, \boldsymbol{\gamma}) &= Pr(\max_{m\in M} C_s < 0) \\ &= Pr(\max_{m\in M} C_{mD} < C_{mE}) \end{aligned} \tag{17}$$

where C_{mE} denotes the channel capacity from the optimal relay (m) to E. According to [47], $P_m^{(k)}(D, \boldsymbol{\gamma})$ can be calculated as shown in (18)

$$P_m^{(k)}(D, \boldsymbol{\gamma}) = \prod_{i=1}^{M} Pr\left(\frac{g_1{}^{(k)} g_2{}^{(k)}}{1 + g_1{}^{(k)} g_2{}^{(k)}} < \frac{g_1{}^{(k)} g_3{}^{(k)}}{1 + g_1{}^{(k)} g_3{}^{(k)}}\right) \tag{18}$$

Without loss of generality, the fading coefficients $g_i^{(k)}$ are assumed to be independent and identically distributed.

Therefore, similar to [R1] , $P_m^{(k)}(D, \boldsymbol{\gamma})$, can be written as follows

$$P_m^{(k)}(D, \boldsymbol{\gamma}) = \prod_{i=1}^{M} Pr(g_2 < g_3) \tag{19}$$

IV. MDP FORMULATION FOR RELAY SELECTION

According to the system model described in the previous sections, the cooperative communication process is analogous to the state transition process. To progress to the next state ($\hat{s}^{(k)}$), the system chooses an action ($a^{(k)}$) and performs it in the current state ($s^{(k)}$). The state of the next time is associated only with the current state and action; therefore, the RS problem is modeled as an MDP [48]. Notably, an MDP is the best way to make decisions for stochastic dynamical systems based on a Markov process (MP), a large class of stochastic processes whose original model is a Markov chain.

In this paper, DQN-RSS and DQN-RSS-ARMA are developed as two different agents for the RS process in a vehicular cooperative network. The purpose of these agents is to select the optimal relay given an outdated CSI in a vehicular environment. In particular, DQN-RSS only considers previously transmitted signals that contain outdated versions of CSI, while DQN-RSS-ARMA employs ARMA-based channel prediction in the reward function to evaluate past and future versions of CSIs. Therefore, the reward-penalty function in DQN-RSS-ARMA gives it the ability to predict the channel based on the predictive behavior of the ARMA model.

Fig. 2 presents the components of RL in the context of the proposed framework, namely, state space (s), action space (a), and the long-term cumulative reward (r), respectively.

A. STATE ($s^{(k)}$)

The system state at time k consists of the following parts: 1) the distance matrix $\boldsymbol{D}_i = d_i^{(k)}, i = 1, 2, 3$; 2) the SNR signals $\boldsymbol{\gamma}$ with entries $\gamma_1^{(k)}$, $\gamma_2^{(k)}$ and $\gamma_3^{(k)}$; and 3) the intercept probability $P_m^{(k)}(D, \boldsymbol{\gamma})$. Thus, the state space can be written as $S = [D, \boldsymbol{\gamma}, P_m^{(k)}(D, \boldsymbol{\gamma})]$ of size $(N \times 6M + 1)$, where M is the number of relays, and N is the previous transmitted frames.

During the initial state setup, the system is in the initial system state $s_o^{(k)}$, which is a tuple consisting of $[D_o, \boldsymbol{\gamma_o}, P_o^{(k)}(D, \boldsymbol{\gamma})]$. In particular, D_o and $\boldsymbol{\gamma_o}$ are obtained when all relays are in the initial position before movement, and $P_o^{(k)}(D, \boldsymbol{\gamma})$ is the intercept probability resulting from this initial setup. Finally, the system reaches the final state $s_f(k)$ after convergence.

B. ACTION ($a^{(k)}$)

Based on the current state of the channel system, an action must be selected for execution. This action is defined as

$$\boldsymbol{A} = \{a^{(k)}\}, \quad k = 1, 2, 3, \ldots, K. \tag{20}$$

where $a^{(k)} \in A = \{0, 1, 2, 3, \ldots.., M\}$ of size $(1 \times M)$. In particular, when $a^{(k)} = 0$, it means that the system does not transmit any data in time slot (k) (i.e., there is no direct transmission). When $a^{(k)} = m$, the best relay m is selected to participate in cooperative communication in time slot (k).

C. REWARD ($r^{(k)}$)

In DQN, the definition of a reward function is extremely important. To this end, the reward $r^{(k)}$, which is a function of the received $P^{(k)}(m)$, varies depending on the state ($s^{(k)}$) and action ($a^{(k)}$). Therefore, in this paper, we present two different reward functions to evaluate their efficiency and robustness against eavesdropping.

1) DQN-RSS FRAMEWOK

In this framework, the immediate reward function is defined as the receiver's instantaneous intercept probability for a given state $s^{(k)}$ and action $a^{(k)}$, namely

$$r^{(k)}(s, a) = \exp(-P^{(k)}(D, \boldsymbol{\gamma})) - (csr \quad \lambda) \tag{21}$$

where csr is the cost of changing the selected relay, indicating the degree of energy consumption for the different actions based on the previous decisions, and λ is the transitions between relays at time $(k-1)$ and (k) according to

$$\lambda = \begin{cases} 0, & a^{(k)} = a^{(k-1)} \\ 1, & a^{(k)} \neq a^{(k-1)} \end{cases} \tag{22}$$

The choice of exp(.) is due to its concavity, which captures adequately the system reward. Decreasing $P^{(k)}(m)$ causes the reward function to grow rapidly.

2) DQN-RSS-ARMA FRAMEWORK

From Fig. 3, it can be seen that the reward function depends on the intercept probability, $(P_{AR}^{(k)})$, taking into account the behavior of the ARMA model. Specifically, the reward function compares the intercept probability of DQN-RSS-ARMA with the ARMA model at each time (k) to consider both outdated CSIs as well as the predicted ones. Whenever the DQN-RSS-ARMA model outperforms or equals the ARMA model, it is rewarded; otherwise, it is penalized. Therefore, the reward function for the proposed system can be expressed as follows:

$$r^{(k)}(s, a) = \begin{cases} -P^{(k)}(D, \boldsymbol{\gamma}) + 20 - (csr \quad \lambda), & P_m^{(k)} > P_{AR}^{(k)} \\ -P^{(k)}(D, \boldsymbol{\gamma}) + 10 - (csr \quad \lambda), & P_m^{(k)} = P_{AR}^{(k)} \\ -P^{(k)}(D, \boldsymbol{\gamma}) - 10 - (csr \quad \lambda), & \text{otherwise} \end{cases} \tag{23}$$

The proposed agents' goal is to maximize the cumulative reward by determining the best strategy π^*, which is typically an approximation function with adjustable parameters that map $(a^{(k)})$ given $(s^{(k)})$.

V. LEARNING-BASED RELAY SELECTION SCHEME

Despite the fact that dynamic programming techniques can be used to solve MDPs and determine the optimal

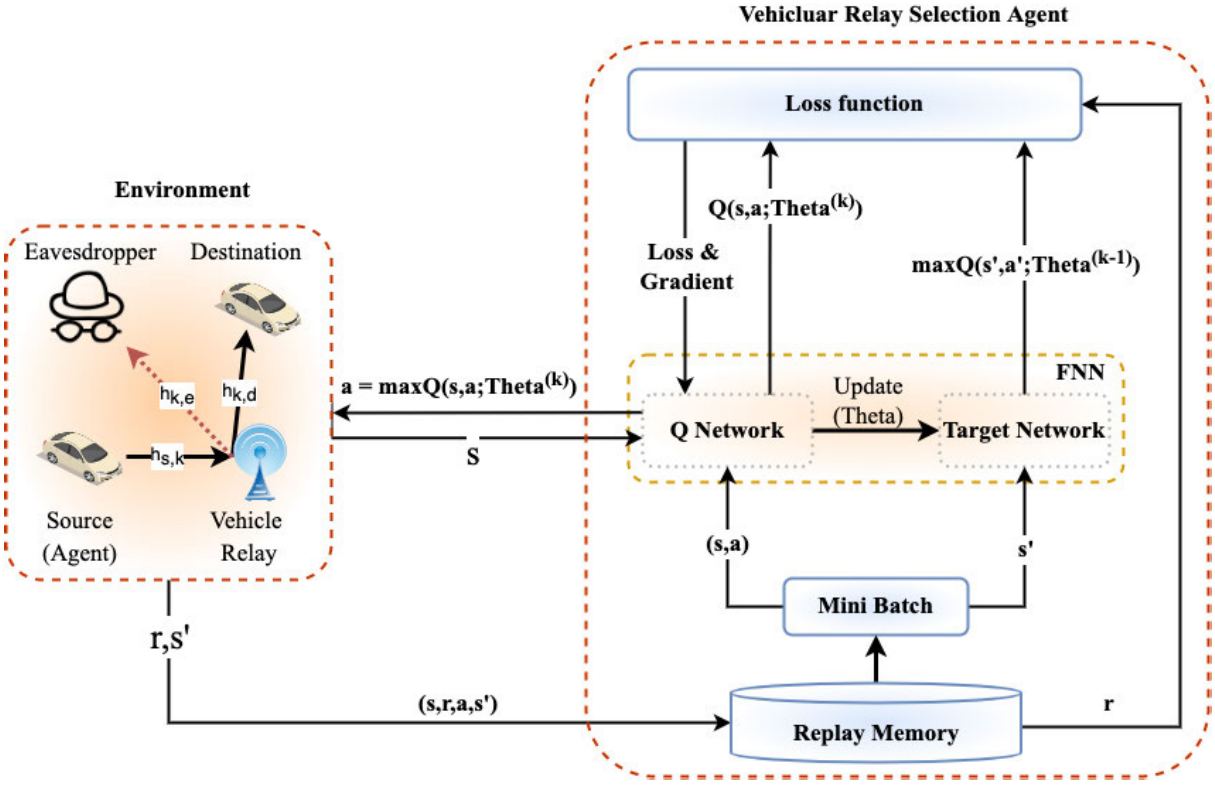


FIGURE 2. Proposed DQN-RSS framework.

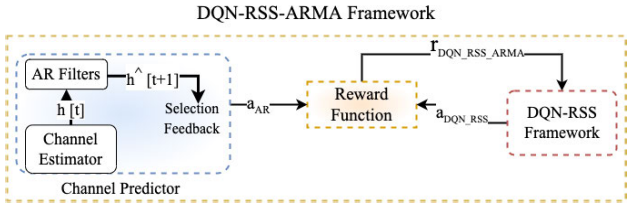


FIGURE 3. Proposed DQN-RSS-ARM framework.

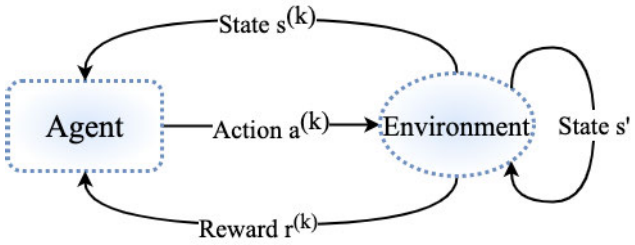


FIGURE 4. Basic components of RL.

strategy, $\pi^*(.)$, large MDPs result in a high degree of computational complexity. In particular, the system states and actions are discrete, and when an action is executed, the system state changes rapidly. Researchers are increasingly employing RL algorithms to overcome the constraint of dimensionality in large MDPs [49]. RL is a goal-oriented computing technique in which an agent interacts with an unknown dynamic environment in order to learn how to perform a task. The objective of the agent is to maximize the task's cumulative reward without being explicitly programmed and without human intervention [50].

In RL, the agent selects an action ($a^{(k)}$) and executes it in a given state ($s^{(k)}$). When the environment accepts this action, it moves to the next state ($\hat{s}^{(k)} = s^{(k+1)}$) as well as sends a reward signal ($r^{(k)}$) to the agent. Consequently, the agent chooses an action ($a^{(k+1)}$) based on the reward ($r^{(k)}$). The fundamental RL framework is presented in Fig. 4.

In the proposed approach, the environment consists of the structure described in Sec. II. The structure of the system is dynamic and unknown at the beginning of each training episode; however, it remains the same during the time steps of each episode. In the proposed framework, shown in Fig. 3, we develop a double-deep Q-network (DDQN) as a learning agent algorithm to solve the RS problem in the presence of obsolete channels. Nonetheless, when the DQN algorithm attempts to approximate a large function, it has a tendency to overestimate the action values. Therefore, this limitation is solved by the idea of DDQN, which generalizes the problem of approximating large functions and provides considerably better performance. This type of agent was employed in our proposed framework since the observation space is continuous and the action space is discrete. The implementation of the proposed DDQN algorithm is shown in **Algorithm 1**.

As Q-network approximators, the DDQN agent implements two independent feedforward neural networks (FNN): 1) the action-value function approximator $Q(s, a; \theta)$ and 2) the target action value function approximator $Q(s, a; \theta^-)$, where θ and θ^- are the current and prior parameters, respectively. To obtain an optimal strategy, θ and θ^- are updated in each iteration; however, $\pi^*(.)$ can be achieved only with $k \to \infty$. In particular, a target network is utilized to generate an action, i.e., the optimal RS, while a Q-network is employed to determine the Q-value of this action. As shown in Table 1, the Q-network is composed of four dense (i.e., fully connected - FC) layers, with a rectified linear unit (ReLU) activation function, $\sigma = \log(1 + e^x)$.

TABLE 1. DDQN critic network dimension.

Layer	Type	Dimension
Input layer	Feature input	$\|S\|$ features
Hidden layer 1	FC	$\|S\| \times 256$
Hidden layer 2	FC	256×256
Output layer	FC	$256 \times \|A\|$

In each time slot k, the agent (i.e., source node (S)) stores its interactive experience tuple $e^{(k)} = (s^{(k)}, a^{(k)}, r^{(k)}, \hat{s})$ in a replay memory (buffer) $\mathcal{D}$, which is defined as $\mathcal{D} = (e^{(1)}, e^{(2)}, \cdots, e^{(K)})$. The FNN parameter $\theta^{(k)}$ is then updated by randomly sampling the display memory $\mathcal{D}$. This demonstrates that given a state $s^{(k)}$, performing an action $a^{(k)}$ in the environment, results in a reward $r^{(k)}$, and changes the state of the environment to $\hat{s}^{(k)}$. The approximation of the Q-value should be able to approximate the state-action value function estimated by the DDQN. This state-action value function is given by the Bellman equation as

$$Q_{\pi^*}(s^{(k)}, a^{(k)}) = \mathbb{E}_s\Big[r^{(k)} + \gamma \max_{a \in A} Q_{\pi^*}(\hat{s}, \hat{a}) \; | s^{(k)}, a^{(k)}\Big] \quad (24)$$

where, $\hat{a} = a^{(k+1)}$.

During training, the DDQN updates the weights (i.e., θ and θ^-) at the end of each time interval to minimize the mean square error (MSE) of the target value in mini-batches every J time slot as follows:

$$\min_{\theta^{(k)}} \Big(L(\theta^{(k)})\Big) = \min_{\theta^{(k)}} \Big(\mathbb{E}_{s,a}[(y_t - Q_\pi(s^{(k)}, a^{(k)}; \theta^{(k)})^2]\Big) \quad (25)$$

where y_t denotes the estimated function value at k when $s^{(k)}$ is the current state and $a^{(k)}$ is the action performed, so y_t is written as follows:

$$y_t = \mathbb{E}_s\Big[r^{(k)} + \gamma \max_{a \in A} Q_\pi(\hat{s}, \hat{a}; \theta^{k-1}) \; | s^{(k)}, a^{(k)}\Big] \quad (26)$$

where $0 < \gamma < 1$ is the discount factor. In training, stochastic gradient descent (SGD) is used to update the weights. In SGD, the weights are initialized randomly and updated iteratively based on a learning rate η. The formula for updating the weights is expressed as:

$$\theta^k = \theta^k - \eta \; \nabla \; L(\theta^{(k)}) \quad (27)$$

Algorithm 1 DQN-Based RS Scheme for Outdated Channels

Initialize γ, η, $\boldsymbol{A}$, $\mathcal{D}$;
Initialize the Q-network with random weights θ;
Initialize the target Q-network parameters by $\theta = \theta^-$;
Initialize the first state s^0;
For $k = 1, 2, 3, \ldots, K$ **do** S node will broadcast the message to all involved relays;
Select $a^{(k)} \in \boldsymbol{A}$ by ϵ-greedy policy;
Receive $D, \boldsymbol{\gamma}$ and the intercept probabilities ($P_m^{(k)}$ & $P_{AR}^{(k)}$) from node D;
Obtain reward function ($r^{(k)}$) for DQN-RSS from (21) and DQN-RSS-ARMA from (23);
Update $\mathcal{D}$ as ($\mathcal{D} \leftarrow \mathcal{D} + \{s^{(k)}, a^{(k),r^{(k)}}, s^{(k+1)}\}$);
For j=1,2,3,....J **do** Select ($s^{(j)}, a^{(j),r^{(j)}}, s^{(j+1)}$) from $\mathcal{D}$ randomly;
$y^{(j)} \leftarrow r^{(j)} + \gamma \max_{a \in A} Q(s^{(j+1)}, a^{(j+1)}; \theta)$;
calculate θ via (27);
Update FNN weights by $\theta^{(k)}$.
Optimal Policy π^*

VI. PERFORMANCE METRICS

In this section, we discuss the performance metrics, such as convergence and intercept probability improvement that are used to measure the performance of the proposed DQL-based framework solutions for the RS problem, given the outdated CSI in a vehicular network.

On the one hand, convergence is the number of episodes in which the total targeted reward is achieved. The convergence time increases as the number of relays (M) increases since both the size of the state space (i.e., the size of the input) and the size of the action space (i.e., the size of the output) increase. Table 4 shows that the input to the action space increases linearly with the size of M, resulting in a linear convergence time.

On the other hand, by using the intercept probability expression given in (19), we obtained the improvement of the intercept probability with variable system model parameters to demonstrate the generalization of the optimal policy ($\pi(.)$), obtained with the proposed RS agents.

VII. RESULTS AND DISCUSSIONS

Based on the performance metrics described in Sec. VI, this section evaluates the performance of the proposed intelligent RS solutions over the vehicular network.

Unless otherwise stated, the system parameters for the proposed frameworks are listed in Table 2, while the RL hyperparameters are depicted in Table 3. During training, these hyperparameters are adjusted to achieve an optimal strategy for RS agents. The impact of some key parameters on the security of the proposed system is quantified and the obtained results are compared with the benchmark solutions. These results are obtained as an average of 200 runs.

The area of the simulation environment is a vehicular network with dimensions of a lane (1000m $\times$ 20m) in which S, D, M relays, and an eavesdropper are placed

TABLE 2. System model parameters.

Parameter	Value
Area Dimensions	1000m × 20m
No. of relays (M)	[2:2:16]
Previous CSI Data (N)	[0,5,10]
Channel Type	Flat fading channel
Source Position	[0,0]
Destination Position	[0,randi(1000)]
Eavesdropper Position	$[randi(20), randi(1000)]$
Path Loss exponent (a)	3
Source & Relay power	20 dB
Noise power	-10 dB
Autoregression order (P)	M
Vehicle Velocity	[0 - 80] kmps
Simulator	MATLAB R2022A

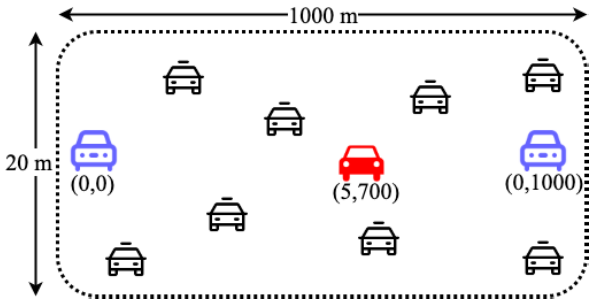


FIGURE 5. Example of the initial simulation scenario with $M = 8$ relays.

TABLE 3. RL hyperparameters.

Parameter	Value
No. of episodes	400
No. of time-steps	200
learning rate η	0.6
Discount factor (Γ)	0.9
Initial exploration rate (ϵ)	1
Exploration decay	0.005
Minimum exploration rate (ϵ_{min})	0.01
Relay switching cost (csr)	[0, 0.01, 0.02]
Target critic smooth factor	1e-3
Target critic update frequency	1
Experience buffer size ($\mathcal{D}$)	10000
Minimum batch size	64
Averaging window size	50
Target average reward	3200

randomly. Specifically, the S, D, and E are located at (0, 0), (0, randi(1000)), and (randi(20), randi(1000)), respectively. It is also assumed that the direct links, namely $S - D$ and $S - E$, are blocked by obstacles. We assumed that one D node moves over time in the simulation area and that the used mobility model for relay distriution is a random waypoint

TABLE 4. DQN-based approaches complexity.

Process	Complexity
Size of action space $\|A\|$	$\mathcal{O}$(M)
Size of state space $\|S\|$	$\mathcal{O}$(N+6M+1)
Overall Run-time	$\mathcal{O}$(N+7M) $\approx$ $\mathcal{O}$(7M)

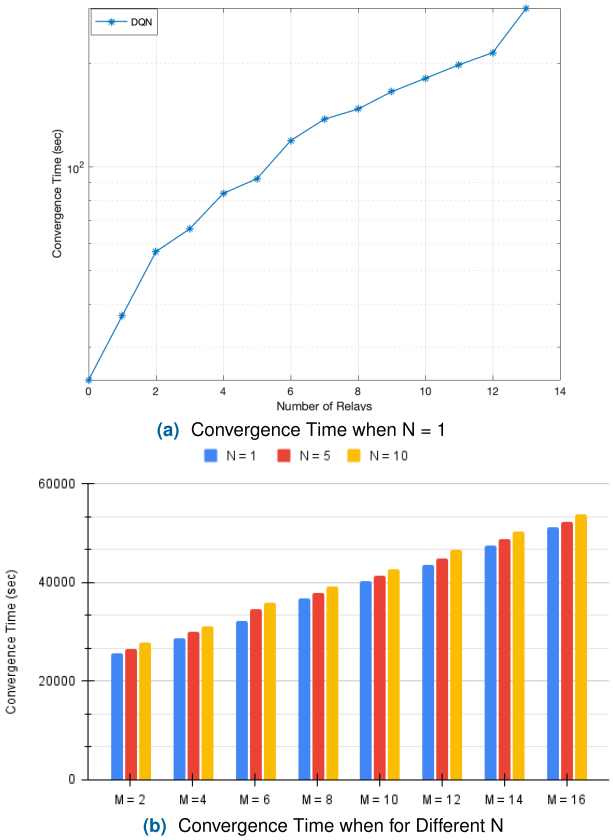
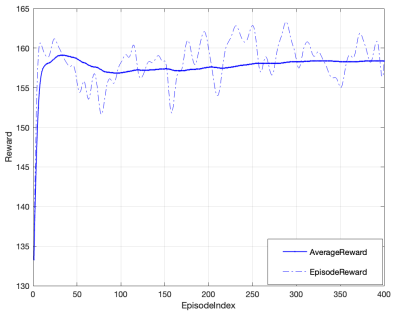


FIGURE 6. Convergence time of the proposed frameworks as a function of the number of vehicular relays M and previous data N.

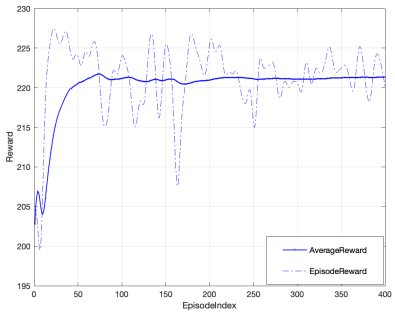
(RWP) model with a velocity distribution of 0-80 km/h. Also, we considered the traffic in the lane to be of a single direction. The example of the initial simulation scenario with $M = 8$ relays is shown in Fig. 5. At each time step, the number of relays and the positions of the D nodes change as well.

The path loss coefficient of the channels is assumed to be $a = 3$. The transmit power of the (S) node and (M) nodes are assumed to be 20 dB. The power of the white Gaussian noise is -10 dB. According to the IEEE 802.15.4 standard protocol, we assume that the operating frequency for all nodes is 2.4 GHz. The learning rate η and discount factor γ are set to 0.6 and 0.9, respectively. Simulations were performed to evaluate the performance of the proposed learning-based predictive RS schemes; ARMA scheme; random RS scheme, and conventional RS scheme.[1]

[1]The random scheme means that the RS policy is random.

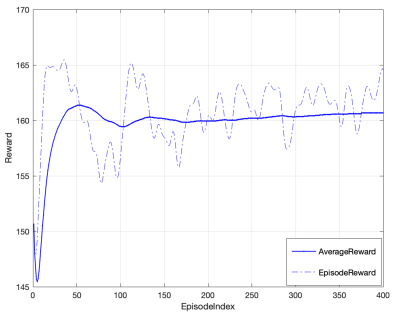


(a) Episode Reward for DQN-RSS with M=4, N=10, csr=0

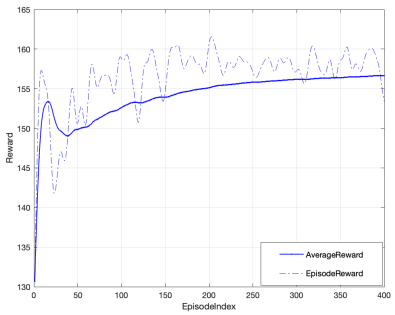


(b) Episode Reward for DQN-RSS with M=14, N=10, csr=0.01

FIGURE 7. Performance convergence of the reward function for the proposed DQN-RSS approach with different specifications.



(a) Episode Reward for DQN-RSS-ARMA with M=4, N=10, csr=0



(b) Episode Reward for DQN-RSS-ARMA with M=14, N=10, csr=0.01

FIGURE 8. Performance convergence of the reward function for the proposed DQN-RSS-ARMA approach with different specifications.

A. CONVERGENCE

Fig. 6 depicts the convergence time for training the proposed DQL-based frameworks as a function of the number of vehicular relays (M). It can be observed that the convergence time for training also increases with the number of relays (M) and previous data (N). This is confirmed by the calculated computational complexity, which is depicted in Table 4.

From Fig. 7 and Fig. 8, we can see that the proposed selection criteria achieve optimal policy (π^*) after convergence in the vehicular cooperative communication system. This confirms the convergence capability of the proposed models (i.e., DQN-RSS and DQN-RSS-ARMA). In essence, the proposed framework outperforms the ARMA approach with a higher reward function and a lower intercept probability.

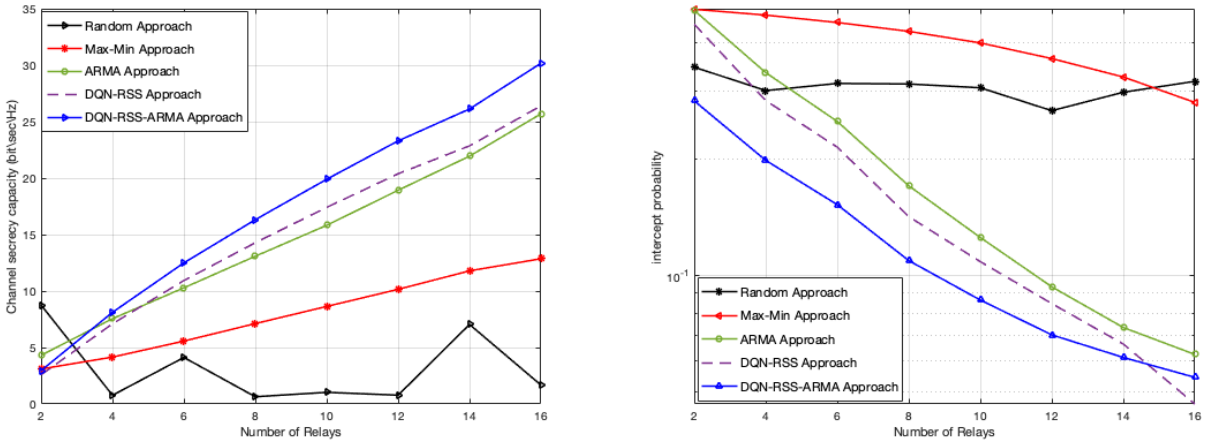
For example, Fig. 7 and Fig. 8 show the reward function of the proposed RS model as a function of time during training with $M = [4, 14]$, $csr = [0, 0.01]$, and $N = 10$, for DQN-RSS and DQN-RSS-ARMA, respectively. The reward function increases with time, as shown in the graphs, indicating that the agents select the optimal relay and that the proposed agent outperforms the autoregression method. Therefore, it is evident that the overall security of the vehicular network is improved. Further to this, after training the proposed model we tested the agent selection in different networks to determine the overall security and compared it with other approaches, which confirms its effectiveness.

B. PERFORMANCE COMPARISON

Fig. 9 compares the average secrecy capacity and intercept probability of different RS approaches, including random, ARMA, max-min, and the proposed DQN approaches (i.e., DQN-RSS and DQN-RSS-ARMA). As the number of relays increases, it can be seen that the DQN-based approaches achieve remarkable performance improvements. For example, when $M = 10$, the secrecy capacity of the proposed method is nearly 30% higher than that of the ARMA approach and almost 4 times higher than that of the max-min selection approach. Moreover, the intercept probability of the proposed approaches is almost 60% lower than that of the conventional models. Fig. 9 also shows that the proposed DQN-RSS-ARMA framework is better than the DQN-RSS framework in terms of both secrecy capacity and intercept probability. However, both of the proposed frameworks exhibit better performance than the conventional approaches. Consequently, the DQN relay selection approaches can effectively improve the security of vehicular networks.

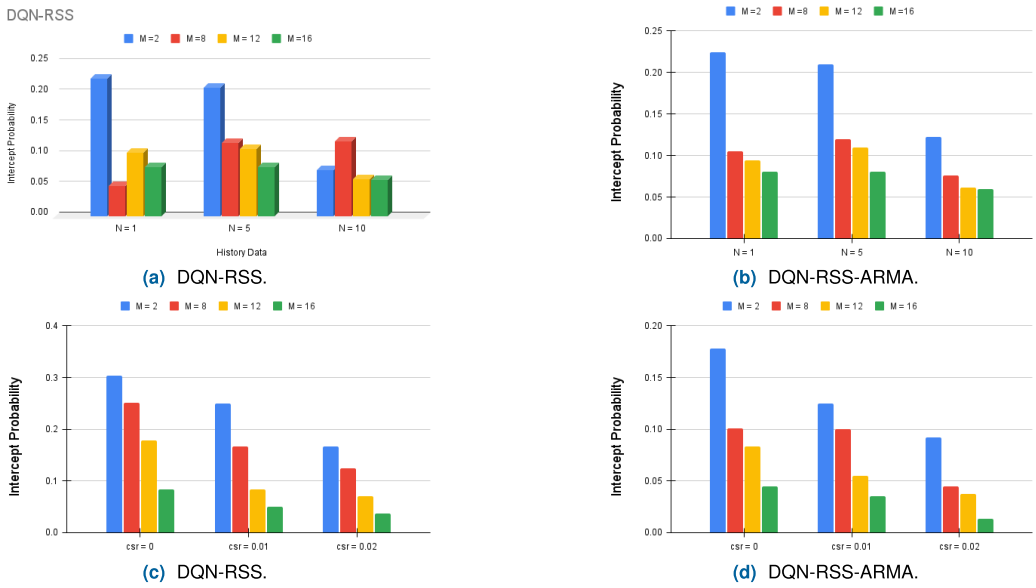
C. INTERCEPT PROBABILITY

Figure 10 illustrates a comparison of the intercept probability of the proposed DQN approaches (DQN-RSS & DQN-RSS-ARMA) for different M and varying N and csr. In addition, Fig. 10 demonstrates that the proposed framework



(a) Secrecy Capacity (b) Intercept Probability

FIGURE 9. Performance comparison of different RS approaches, including Random, ARMA, max-min, and the proposed DQN approaches (DQN-RSS & DQN-RSS-ARMA).



(a) DQN-RSS. (b) DQN-RSS-ARMA.

(c) DQN-RSS. (d) DQN-RSS-ARMA.

FIGURE 10. Comparison of the intercept probability of the proposed DQN approaches (DQN-RSS & DQN-RSS-ARMA) at (a,b) when N = [1], [5], [10] for different M (2,8,12,16) and (c,d) when csr = [0,0.01,0.02] for different M (2,8,12,16).

(DQN-RSS-ARMA) archives a lower intercept probability than the other frameworks when M or csr change. From Fig. 10(a) and Fig. 10(b), it can be seen that for a given number M, the intercept probability decreases as N increases, while for a large number of M (i.e., $M > 10$), the intercept probability remains almost the same regardless of the available past data. Thus, if a large number of relays (M) are available in the lane, there is no need to consider large data from outdated data. On the other hand, the intercept probability decreases when csr increases and the number of vehicular relays increases.

Fig. 11 shows the comparison of intercept probability performance between the proposed frameworks and the conventional ARMA given the change in available previous CSI data. This figure shows the achieved performance when only an outdated CSI is considered. It is observed that the proposed DQN-RSS-ARMA outperforms both ARMA and the DQN-RSS approaches. In particular, for a small number of relays (i.e., $M < 6$), ARMA outperforms the DQN-RSS approach, while this pattern is reversed for a large number of relays. It is worth noting that the proposed approaches behave identically as the number of relays increases. On the other hand, Fig. 11(b) demonstrates the achieved intercept probability when ($N = 10$). In this context, the ARMA approach exhibits the lowest intercept probability for a small number of relays, while the proposed approaches outperform the ARMA approach when M increases. In conclusion, for a small value of M, our proposed frameworks outperform ARMA by almost 30%, while for a higher value of M, ARMA performs almost 5% lower than the proposed frameworks.

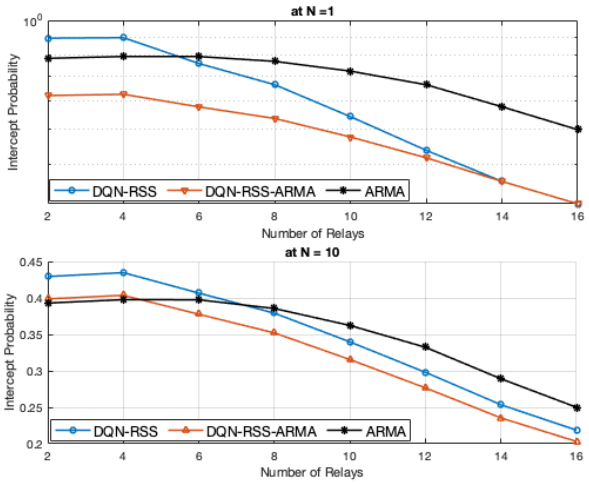


FIGURE 11. Performance comparison of intercept probability between the proposed frameworks and the conventional ARMA was given the change in available previous data.

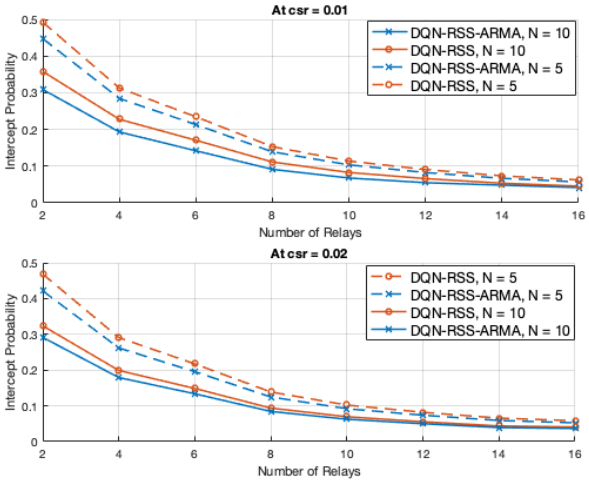


FIGURE 12. Performance comparison of intercept probability between the proposed frameworks given the change in *csr* and available previous data.

Fig. 12 shows the performance comparison of the intercept probability between the proposed frameworks given the change in *csr* and the available previous data. In Fig. 12(a), we observe the corresponding behavior at $csr = 0.01$ and $N = [5, 10]$ over the available relays, while Fig. 12(b) shows the performance when *csr* increases to 0.02. For a given *csr*, the intercept probability decreases as N or M increases. Also, for a given N, the intercept probability decreases as *csr* or M increases. It is worth noting that the proposed DQN-RSS-ARMA outperforms the DQN-RSS approach. Therefore, if we include the predicted CSI in the reward when training the agent, the agent will be rendered a more experienced decision at selecting the optimal relay.

VIII. CONCLUSION

The present contribution investigated the use of cooperative communication with adaptive RS for WVNs as well as proposed intelligent RS algorithms DQN-RSS and DQN-RSS-ARMA, to improve the security of the physical layer and select the optimal relay. In the proposed algorithm, we leveraged recent advances to formulate the opportunistic relaying optimization as an MDP model and transform the secure RS into a prediction and decision-making problem. A source node collects the CSI from the environment and then sends the integral system state to the DQN to derive the optimal RS policy. The effects of increasing the number of relays (M) and outdated CSI (N) were analyzed to select the optimal relay over the proposed schemes. The results confirm that the proposed DQN-RSS-ARMA algorithm outperforms all other RS schemes discussed in this paper in terms of lower intercept probability and higher secrecy capacity. In future work, we will consider blockchain ledgers to guarantee the reliability and authenticity of the involved relays over more complex channel models that are more realistic in real WVNs.

REFERENCES

[1] J. D. V. Sánchez, L. Urquiza-Aguiar, M. C. P. Paredes, and D. P. M. Osorio, "Survey on physical layer security for 5G wireless networks," *Ann. Telecommun.*, vol. 76, no. 3, pp. 155–174, Apr. 2021.

[2] N. Wang, P. Wang, A. Alipour-Fanid, L. Jiao, and K. Zeng, "Physical-layer security of 5G wireless networks for IoT: Challenges and opportunities," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8169–8181, Oct. 2019.

[3] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. Di Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[4] J. Zhang, T. Q. Duong, A. Marshall, and R. Woods, "Key generation from wireless channels: A review," *IEEE Access*, vol. 4, pp. 614–626, 2016.

[5] A. Sanenga, G. Mapunda, T. Jacob, L. Marata, B. Basutli, and J. Chuma, "An overview of key technologies in physical layer security," *Entropy*, vol. 22, no. 11, p. 1261, Nov. 2020.

[6] X. Liao, Z. Wu, Y. Zhang, and X. Jiang, "Buffer-aided relay selection for secure communication in two-hop wireless networks with limited packet lifetime," *Ad Hoc Netw.*, vol. 121, Oct. 2021, Art. no. 102580.

[7] J. Mo, M. Tao, and Y. Liu, "Relay placement for physical layer security: A secure connection perspective," *IEEE Commun. Lett.*, vol. 16, no. 6, pp. 878–881, Jun. 2012.

[8] A. Tukmanov, S. Boussakta, Z. Ding, and A. Jamalipour, "Outage performance analysis of imperfect-CSI-based selection cooperation in random networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2747–2757, Aug. 2014.

[9] Y. Su, L. Jiang, and C. He, "Joint relay selection and power allocation for full-duplex DF co-operative networks with outdated CSI," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 510–513, Mar. 2016.

[10] M. A. Jadoon and S. Kim, "Relay selection algorithm for wireless cooperative networks: A learning-based approach," *IET Commun.*, vol. 11, no. 7, pp. 1061–1066, May 2017.

[11] J. Lu, D. He, and Z. Wang, "Learning-assisted secure relay selection with outdated CSI for finite-state Markov channel," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–5.

[12] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[13] F. Jameel, S. Wyne, G. Kaddoum, and T. Q. Duong, "A comprehensive survey on cooperative relaying and jamming strategies for physical layer security," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2734–2771, 3rd Quart., 2019.

[14] X. Shang, H. Yin, Y. Wang, M. Li, and Y. Wang, "Secure multiuser scheduling for hybrid relay-assisted wireless powered cooperative communication networks with full-duplex destination-based jamming," *IEEE Access*, vol. 9, pp. 49774–49787, 2021.

[15] L. Sun, T. Zhang, L. Lu, and H. Niu, "Cooperative communications with relay selection in wireless sensor networks," *IEEE Trans. Consum. Electron.*, vol. 55, no. 2, pp. 513–517, May 2009.

[16] W. Fang, F. Liu, F. Yang, L. Shu, and S. Nishio, "Energy-efficient cooperative communication for data transmission in wireless sensor networks," *IEEE Trans. Consum. Electron.*, vol. 56, no. 4, pp. 2185–2192, Nov. 2010.

[17] I. Krikidis, J. Thompson, and S. McLaughlin, "Relay selection issues for amplify-and-forward cooperative systems with interference," in *Proc. IEEE WCNC*, Budapest, Hungary, Apr. 2009, pp. 1–6.

[18] V. N. Q. Bao, N. Linh-Trung, and M. Debbah, "Relay selection schemes for dual-hop networks under security constraints with multiple eavesdroppers," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6076–6085, Dec. 2013.

[19] N.-P. Nguyen, T. Q. Duong, H. Q. Ngo, Z. Hadzi-Velkov, and L. Shu, "Secure 5G wireless communications: A joint relay selection and wireless power transfer approach," *IEEE Access*, vol. 4, pp. 3349–3359, 2016.

[20] N. Zhang, N. Cheng, N. Lu, X. Zhang, J. W. Mark, and X. Shen, "Partner selection and incentive mechanism for physical layer security," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4265–4276, Aug. 2015.

[21] L. Wang, Y. Cai, Y. Zou, W. Yang, and L. Hanzo, "Joint relay and jammer selection improves the physical layer security in the face of CSI feedback delays," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6259–6274, Aug. 2016.

[22] S. Yadav and P. K. Upadhyay, "Impact of outdated channel estimates on opportunistic two-way ANC-based relaying with three-phase transmissions," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5750–5766, Dec. 2015.

[23] C. Wu, X. Yi, Y. Zhu, W. Wang, L. You, and X. Gao, "Channel prediction in high-mobility massive MIMO: From spatio-temporal autoregression to deep learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1915–1930, Jul. 2021.

[24] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5G wireless communications: A deep learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 227–236, Jan. 2020.

[25] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[26] A. Masmoudi and T. Le-Ngoc, "A maximum-likelihood channel estimator for self-interference cancelation in full-duplex systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5122–5132, Jul. 2016.

[27] J. Ma, S. Zhang, H. Li, N. Zhao, and A. Nallanathan, "Iterative LMMSE individual channel estimation over relay networks with multiple antennas," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 423–435, Jan. 2018.

[28] M. Bhuyan, K. K. Sarma, and N. E. Mastorakis, "Nonlinear mobile link adaptation using modified FLNN and channel sounder arrangement," *IEEE Access*, vol. 5, pp. 10390–10402, 2017.

[29] S. Kashyap, C. Mollén, E. Björnson, and E. G. Larsson, "Performance analysis of (TDD) massive MIMO with Kalman channel prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 3554–3558.

[30] Z. Chen, X. Gao, W. Liu, D. Yu, and G. Yue, "Channel prediction for millimeter-wave V2V communication using autoregressive models," in *Proc. 13th Int. Symp. Antennas, Propag. EM Theory (ISAPE)*, Zhuhai, China, Dec. 2021, pp. 1–3.

[31] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran, and Q. E. Ul Haq, "Machine learning techniques for 5G and beyond," *IEEE Access*, vol. 9, pp. 23472–23488, 2021.

[32] X. Wang and F. Liu, "Data-driven relay selection for physical-layer security: A decision tree approach," *IEEE Access*, vol. 8, pp. 12105–12116, 2020.

[33] A. K. Kamboj, P. Jindal, and P. Verma, "Intelligent physical layer secure relay selection for wireless cooperative networks with multiple eavesdroppers," *Wireless Pers. Commun.*, vol. 120, no. 3, pp. 2449–2472, Oct. 2021.

[34] Y. Su, X. Lu, Y. Zhao, L. Huang, and X. Du, "Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks," *IEEE Sensors J.*, vol. 19, no. 20, pp. 9561–9569, Oct. 2019.

[35] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[36] Y. Su, M. Liwang, Z. Gao, L. Huang, X. Du, and M. Guizani, "Optimal cooperative relaying and power control for IoUT networks with reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 791–801, Jan. 2021.

[37] C. Huang, G. Chen, and Y. Gong, "Delay-constrained buffer-aided relay selection in the Internet of Things with decision-assisted reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 10198–10208, Jun. 2021.

[38] X. Wang, T. Jin, L. Hu, and Z. Qian, "Energy-efficient power allocation and Q-learning-based relay selection for relay-aided D2D communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6452–6462, Jun. 2020.

[39] Y. Geng, E. Liu, R. Wang, and Y. Liu, "Hierarchical reinforcement learning for relay selection and power optimization in two-hop cooperative relay network," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 171–184, Jan. 2022.

[40] C. Huang, G. Chen, Y. Gong, P. Xu, Z. Han, and J. A. Chambers, "Buffer-aided relay selection for cooperative hybrid NOMA/OMA networks with asynchronous deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2514–2525, Aug. 2021.

[41] H. Zhang, S. Chong, X. Zhang, and N. Lin, "A deep reinforcement learning based D2D relay selection and power level allocation in mmWave vehicular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 416–419, Mar. 2020.

[42] W. Viriyasitavat, M. Boban, H. Tsai, and A. Vasilakos, "Vehicular communications: Survey and challenges of channel and propagation models," *IEEE Veh. Technol. Mag.*, vol. 10, no. 2, pp. 55–66, Jun. 2015.

[43] D. Sun and Y. Li, "A channel prediction scheme with channel matrix doubling and temporal-spatial smoothing," *Wireless Pers. Commun.*, vol. 122, no. 3, pp. 2045–2055, Feb. 2022.

[44] L. Liu, H. Feng, T. Yang, and B. Hu, "MIMO-OFDM wireless channel prediction by exploiting spatial–temporal correlation," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 310–319, Jan. 2014.

[45] M. Torabi and D. Haccoun, "Capacity analysis of opportunistic relaying in cooperative systems with outdated channel information," *IEEE Commun. Lett.*, vol. 14, no. 12, pp. 1137–1139, Dec. 2010.

[46] J. M. Hamamreh, H. M. Furqan, and H. Arslan, "Classifications and applications of physical layer security techniques for confidentiality: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1773–1828, 2nd Quart., 2019.

[47] Y. Zou, X. Wang, and W. Shen, "Optimal relay selection for physical-layer security in cooperative wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2099–2111, Oct. 2013.

[48] T. M. Moerland et al., "Model-based reinforcement learning: A survey," *Found. Trends Mach. Learn.*, vol. 16, no. 1, pp. 1–118, Jan. 2023.

[49] Q. Wei, D. Liu, and G. Shi, "A novel dual iterative Q-learning method for optimal battery management in smart residential environments," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2509–2518, Apr. 2015.

[50] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, Nov. 2018.