IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 25, 2021, accepted November 18, 2021, date of publication November 22, 2021,
date of current version December 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129896

Reliability Optimization in Narrowband Device-to-Device Communication for 5G and Beyond-5G Networks

ALI NAUMAN[1], MUHAMMAD ALI JAMSHED[2], (Member, IEEE), YAZDAN AHMAD QADRI[1], RASHID ALI[3], (Member, IEEE), AND SUNG WON KIM[1]

[1]Wireless Information and Networking Laboratory, Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, Gyeongbuk 38541, Republic of Korea
[2]James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, U.K.
[3]School of Intelligent Mechatronics Engineering, Sejong University, Seoul 05006, Republic of Korea

Corresponding author: Sung Won Kim (swon@yu.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program IITP-2021-2016-0-00313, supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP); and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2021R1A6A1A03039493.

ABSTRACT The 5G and beyond-5G (B5G) is expected to be a key enabler for Internet-of-Everything (IoE). The narrowband Internet of Things (NB-IoT) is a low-power wide-area enabling technology introduced by the 3rd Generation Partnership in 5G. The objective of the NB-IoT is to enhance the mobile coverage area by increasing the number of repetitions of control and data packets between user equipment (UE) and the base station/evolved NodeB (BS/eNB). While these repetitions improve data delivery for delay-sensitive applications, they degrade the efficiency of the already resource-constrained IoT system by increasing the system overhead and energy consumption. Moreover, NB-IoT devices in the edge region of the cellular coverage area require more repetitions, which augment energy consumption. In this study, we investigate device-to-device (D2D) communication for NB-IoT delay-sensitive applications, such as healthcare-IoT services, to use two-hop communication instead of using a direct uplink. An optimization problem is formulated to achieve an optimal end-to-end delivery ratio (EDR). In addition, this study incorporates Q-Learning-based reinforcement learning (RL) for the selection of an optimal cellular relay, which assists NB-IoT UE in uploading sensitive data to BS/eNB. The proposed RL-intelligent-D2D (RL-ID2D) communication methodology selects the optimum relay with a maximum EDR, which ultimately augments energy efficiency.

INDEX TERMS Device-to-device (D2D) communication, machine learning (ML), narrowband Internet-of-Things (NB-IoT), reinforcement learning (RL).

I. INTRODUCTION

Internet of Things (IoT) is a key enabler of the smart city concept and is playing a major role in revolutionizing the future wireless communications and applications, such as Industry 5.0, connected autonomous healthcare, and smart transportation. IoT is an integration of physical and cyber world, where a collections of smart objects capable of sensing, and actuation are able to self-configure, process data, and are inter-operable to form a network [1]. As defined by the 3rd generation partnership project (3GPP), the IoT falls into the category of massive machine-type communication (mMTC) in beyond-5G (B5G) networks [2]. The basic network operation of IoT device is to transmit data in either uplink, downlink, or in both directions. Usually, IoT devices upload the sensed data to the sink node for processing, based on which instructions for actuation are transmitted in downlink direction. Critical IoT applications such as healthcare, industrial automation, and autonomous cars require periodic checking and timely delivery of information in uplink direction. Such sensitive services require higher data rates and the need to maintain a reliable uplink transmission

The associate editor coordinating the review of this manuscript and approving it for publication was Chih-Min Yu.

is of prime importance [3]. Currently, 23 billion devices are connected to the Internet, and this number is expected to increase to about 75 billion by 2025 [1]. Therefore, due to massive connectivity and ultra reliability requirement, there is a dire need of enabling technology for IoT devices integrated within the core cellular network with extended coverage and improved spectral efficiency. A common approach for realizing this reliability is the repetitive transmission of control and data packets. The narrowband-Internet of Things (NB-IoT), which is a standard of the 3GPP introduced in Release 13, has improved spectral efficiency and extended coverage [4]. NB-IoT is a low-power wide area (LPWA) networking technology that supports mMTC. NB-IoT can be deployed in long-term evolution advanced (LTE-A), or global system for mobile (GSM) communication networks to share the spectrum and reuse the same hardware to reduce the deployment costs. The NB-IoT operates on a single physical resource block (PRB) that has a 180-kHz bandwidth for both uplink and downlink within the cellular spectrum [5]. The extended reliability and coverage offered by the NB-IoT is due to the repetitive retransmissions, i.e., 128 re-transmissions for uplink and 2048 re-transmissions for downlink [6]. However, the repeated transmission degrades the spectral efficiency [7]; narrow-band re-transmissions and massive transmission time interval (TTI) bundling results in an increased time and energy resource consumption [8]. In addition, for environments with obstructive barriers, there is an additional penetration loss of 20 dB. Therefore, an efficient uplink transmission methodology with a high end-to-end delivery ratio (EDR) is required for IoT applications. Device-to-device (D2D) communication provides a promising solution to enhance network performance and enables devices to communicate directly without the intervention of a cellular network [9]. The 3GPP standardized D2D communication for the first time within Long-Term Evolution Advanced (LTE-A) in Release 12 [10] and further discussions are still in process in Release 17 [11], where it is also termed as Proximity-based Service (ProSe) [12]. In order to increase reliability, D2D communication provides an efficient mechanism to aid the NB-IoT user equipment (UE) to upload the data to base station/evolved NodeB (BS/eNB) [13].

A. MOTIVATION

5G and B5G are expected to be the enabling technologies for IoT. LTE-A is standardized to support B5G as non-standalone architecture. In the context of data delivery for IoT devices, quality-of-service (QoS) is of prime importance. Integration of D2D communication and artificial intelligence is considered an important piece of B5G and 6G jigsaw puzzle [9].

B. CONTRIBUTION

This work introduces a reinforcement-learning (RL) based intelligent approach RL-intelligent-D2D (RL-ID2D) relay selection methodology for D2D communication to ensure a high EDR for NB-IoT devices.

RL-ID2D reduces the additional overheads by working as a distributed system at the NB-IoT UE, which models the relay selection problem in a two-step Q-Learning (QL) framework [14]. The QL algorithm is an effective RL based machine learning (ML) method for dynamic scenarios, and it converges quickly in dynamic, independent, and randomly distributed traffic [14]. QL uses a matrix of evaluation scores based on the rewards of each successful action. The proposed scheme selects the best relay node for D2D communication to enhance the energy efficiency by maximizing the EDR. The contributions of this study are summarized as follows:

- *Narrowband D2D:* This article adapts the D2D communication as a routing extension for relaying NB-IoT UE data to the BS/eNB in order to maximize the EDR.
- *OptPRS:* This article formulates the D2D relay selection problem as an optimization problem to maximize the EDR. In order to solve this problem, we propose a simple yet effective solution called the optimum potential relay set (optPRS). The proposed optPRS forms a set of potential relays by comparing the essential parameters with very low complexity.
- *Intelligent-D2D:* Furthermore, to ensure the maximization of the EDR in dynamic environment, this article models the relay selection as an MDP problem and utilizes RL based ML to solve the problem. The proposed "RL-ID2D" mechanism selects the relay efficiently from the PRS to maximize the EDR and ultimately improves energy efficiency.
- *Performance Evaluation:* To validate the performance of the proposed intelligent scheme, simulation results are presented. A comparison of RL-ID2D with the opportunistic model [15] and the deterministic model [16] is presented, which shows that the RL-ID2D improves the EDR.

1) PAPER ORGANIZATION

The remainder of this paper is organized as follows: Section II presents detail of the related research work. Section III explains the system model and defines the essential parameters, and Section IV formulates the problem. Section V provides the optPRS solution and shows how RL-ID2D can enable NB-IoT UE to select relay nodes intelligently in detail. Section VI discusses the performance evaluation and simulation results, and Section VII provides application areas and future research directions. Section VIII concludes this article.

II. RELATED WORK

Reliability is a desirable requirement in 5G-enabled networks, and it corresponds to an increased data delivery ratio, especially for healthcare applications. Currently, a large number of studies have been conducted to improve the performance of D2D communication [17]. However, few studies have investigated ways of improving the reliability of D2D communication in NB-IoT-based networks. *Militano et al.* proposed coalition-based multi-hop D2D communication in

TABLE 1. Symbols used throughout the paper.

Symbol	Meaning	Symbol	Meaning
α	Step size or learning rate	β	SINR threshold
γ	Discount factor	σ	Exploration and exploitation threshold
g	Channel gain	μ	Path loss exponent
P_t	Transmission power	P_r	Received power
BS/eNB	Base Station/evolved NodeB	CUE	Cellular user equipment
L	Total data packets	δ_{k_n}	Relay binary allocating indicator
E	Total data packets received	k_n	State (CUE relay)
π	Reinforcement learning policy	$r_k(t)$	Reward of relay k at time t
a_t	Action of agent at time t	$Q(k_n(t), a_t)$	Action value function of state k at time t
ϵ	Exploration and exploitation ratio	R	Total number of UEs
$P^{NB-IoT}_{t_{(UE\rightarrow k_n)}}$	Transmission power of NB-IoT UE	ζ	Threshold of transmission power of NB-IoT UE
$P^{NB-IoT}_{t_{(UE\rightarrow k_n)(max)}}$	Maximum transmission power of NB-IoT UE	$P^{NB-IoT}_{t_{(UE\rightarrow k_n)(min)}}$	Minimum transmission power of NB-IoT UE
$PRS(K)$	Potential relay set	N	Total number of relays in PRS $n \in \{1, 2, ..., N\}$
t	Discrete time step	$k_n(t)$	State k_n (CUE relay) at time t where $n \in \{1, 2, ..., N\}$
$k^{'}$	State k_n at time $(t+1)$	$\mathcal{P}$	Transitional probability

5G networks [18]. In the proposed mechanism, a coalition is formed by a centralized eNodeB unit. The devices in its close proximity form a cooperative chain to upload the data to eNodeB.

The authors formulated an optimization problem to maximize the throughput and coalition time. The authors in [19] investigated a secure and trustworthy relay node selection for D2D communication in an NB-IoT network. This work is an extension of the work proposed in [18]. The authors proposed that the eNodeB should form an information matrix in order to maintain the record of the users and reliable relay nodes. Moreover, the nodes in the cooperative chain refer to the trustworthy connecting relay link to be stored in the information matrix. However, the focus of this study is to identify malicious relay devices and improve the data rate.

The reliability of a wireless network can be improved by using broadcasting-based *opportunistic routing* protocols. Opportunistic routing protocols broadcast messages to potential relays, which select the forwarder by co-ordination amongst themselves [15]. In [15], a UE transmits a packet to relaying nodes in an active duty cycle period if it finds an opportunity. However, if it fails in that attempt, it re-transmits in the next available opportunity to another node from the relaying group. The re-transmissions expire after a threshold time interval if the UE fails to successfully transmit the data. However, it is evident that the associated overhead is significant with this approach, and it leads to increased delay and energy consumption. This approach is not suitable for delay-sensitive applications such as healthcare, defense, and industrial automation. In [16], the authors proposed a deterministic approach instead of an opportunistic model for relay selection. The proposed approach selects the relay for D2D communication at the BS, which eliminates the additional delay present in the opportunistic model to wait for cellular UE (CUE) to operate as a relay. However, to select the relay in a deterministic manner, the NB-IoT UE must transmit a pilot signal every time it has data to upload to the eNB/BS. The CUEs that qualify and are available for D2D communication transmit the pilot signal to the eNB/BS to select the best relay. The eNB/BS selects the best candidate for relaying the data after ranking the relays in decreasing order on the basis of channel gain and residual power. This incorporates additional processing and delay, thereby increasing energy consumption.

To the best of the authors' knowledge, this article is the first to investigate the selection of a relay for D2D communication by modelling it as a MDP problem and solving it using RL.

III. SYSTEM MODEL

A. SYSTEM MODEL AND DEFINITIONS

In this study a two-tier network model is considered, which includes CUEs and NB-IoT UEs. The network model considers the scenario of a smart city, where the NB-IoT UEs have critical data to transmit. Timely and reliable data transmission is of utmost importance. The *R* UEs (including CUEs and NB-IoT UEs) are distributed randomly, and can directly communicate with the BS/eNB, which is placed at the center of the cell. However, it is expected that the NB-IoT UE should transmit the uplink data to the BS/eNB in a two-hop manner by exploiting the CUE as a relay in a D2D communication. The uplink bandwidth is subdivided into *C* sub-channels. Each NB-IoT UE and CUE shares the same uplink resources to communicate with BS/eNB. Any CUE, under conditions that qualify for it to act as a relay (explained in the next subsection), can assist a NB-IoT UE with uploading the data to BS/eNB, exploiting D2D communication. Table 1 presents the list of symbols used throughout the paper.

1) CHANNEL MODEL

The channel propagation model between the transmitter and receiver considered in this study is the Rayleigh channel, and the channel gain is normally distributed. Rayleigh channel correspond to urban environment, where transmitted signal experiences multipath fading.

2) MOBILITY MODEL

The mobility model used in the system model is random waypoint model (RWP). The BS/eNB is assumed to be at the center of the cell and the UEs are assumed to be distributed randomly around the BS/eNB. The UEs are identical and independently distributed and each UE is moving with

TABLE 2. Simulation parameters.

Symbol	Value
No. of UEs R	50-80
Max. communication range	130 meter
SINR threshold (β)	13 dB
Path-loss exponent (μ)	3.5
Channel model	Rayleigh
Mobility model	Random Waypoint
Velocity interval	[0.2 2.5] m/s
Walk interval	[1 2] s
Pause interval	[1 2] s
Noise power density	-174 dBm/Hz
Max. transmission power of NB-IoT UE (ζ)	18 dB
Max. transmission power of CUE	23 dB
No. of packets (L) with size	1000 packets of 32 bytes
Exploration and exploitation constant (σ)	$0 < \sigma \leq 1$
Discount factor (γ)	$0 \leq \gamma \leq 1$
Learning rate or step-size (α)	$\alpha \in (0, 1]$
Simulation iterations	1000

random velocity. The pause time which show how long the UE is static at one point is also random. The direction in which the UE is moving is also random. The NB-IoT UE is considered to be static, while CUEs are mobile. The lower and upper bound for the random values is shown in Table 2.

3) SIGNAL-TO-INTERFERENCE-NOISE-RATIO (SINR)

The communication links, that is, the D2D and cellular link uses the same network resources, therefore,

cross-tier interference among all the communication links is avoided assuming that a unique sub-carrier is allocated to each UE. Furthermore, the interference is minimized as the transmitting power for D2D link is lower than that of CUE direct links assuming the distance between the D2D pair is small. As per the 3GPP specifications, the SINR for the LTE-A link is calculated by the UE internally and reported to eNB during uplink transmission to determine the link quality of each UE. The SINR is calculated from the reference signal received quality (RSRQ), which is determined by the reference signal received power (RSRP) [20]:

$$RSRQ = N_{PRB} \times \frac{RSRP}{RSSI} \tag{1}$$

where $RSSI$ is the received signal strength indicator (RSSI) and N_{PRB} is the number of physical resource blocks. The SINR is then measured as

$$SINR = \frac{12.RSRQ}{x} \quad where \quad x = \frac{RE}{RB} \tag{2}$$

where RE indicates the resource element and RB indicates the resource block. The RE is composed of one sub-carrier, whereas RB is contains 12 sub-carriers. The 12 in eq.(2) indicates the 12 sub-carriers of RB over which the RSSI is measured. Whereas, the RSRP is measured over a single RE. In LTE-A, the channel quality indicator (CQI) based on SINR is calculated by UE and exchanged with eNB every 2–10 ms [20].

B. DEFINITIONS

1) PACKET DELIVERY RATIO

The PDR is a key performance metric for evaluating the reliability. The PDR is defined as the ratio of the number of packets that are transmitted at the transmitter to the number of packets received at the receiver end [21]. The following expression defines the PDR as

$$PDR = \frac{\sum_{e=0}^{E}(E_e)}{\sum_{l=1}^{L}(L_l)} \tag{3}$$

where L is the total number of packets transmitted and E is the total number of packets received.

2) POTENTIAL RELAY SET

The potential relay set (PRS) for NB-IoT UE is the N number of CUEs that are within the range of NB-IoT UE for D2D communication and assist NB-IoT UE to forward the packet to the eNB. The PRS is also denoted by the state-space $K = \{k_1, k_2, k_n, \ldots, k_N\}$ of the environment, where $n \in \{1, 2, \ldots, N\}$.

3) END-TO-END DELIVERY RATIO

The EDR is a performance metric in multi-hop transmissions. It is the product of PDR from NB-IoT UE to CUE relay and from relay CUE to BS/eNB. To achieve 100% EDR, both ratios should be 1. The EDR is calculated using the following expression:

$$EDR = \prod_{n=1}^{N} (PDR^{(k_n)}_{UE \to CUE_{k_n}} . PDR^{(k_n)}_{CUE_{k_n} \to BS/eNB}) \tag{4}$$

IV. PROBLEM FORMULATION

We formulate an EDR maximization optimization problem for the uplink of NB-IoT systems and provide an effective and simple solution to solve the problem. Mathematically, the problem is formulated as follows:

$$\begin{aligned}
&\text{maximize } EDR^{l(k_n)}_{UE \to CUE_{k_n} \to BS/eNB} \\
&\text{subject to } C_1 : P^{NB-IoT}_{t_{(UE \to k_n)}} \leq \zeta \\
&\qquad C_2 : \delta_{k_n} \in \{0, 1\} \\
&\qquad k_n \in \{k_1, k_2, \cdots, k_N\}, \quad l \in \{1, 2, \cdots, L\} \\
&\qquad \delta_{k_n} = \begin{cases} 1, & SINR^{(k_n)}_{k_n \to BS/eNB} \geq \beta \\ 0, & otherwise \end{cases} \\
&\qquad P^{NB-IoT}_{t_{(UE \to k_n)(min)}} \leq \zeta \leq P^{NB-IoT}_{t_{(UE \to k_n)(max)}}
\end{aligned} \tag{5}$$

In the optimization problem (5), the objective function is related to the total uplink EDR over k_N relays for L packets over time t. The constraint C_1 shows the required transmission power for the NB-IoT UE needed to transmit the data to the CUE relay. It also indicates that the CUE relay is bounded to be present in the feasible transmission power range ζ of the NB-IoT UE. The constraint C_2 reflects a binary allocating indicator δ_{k_n}, that is, δ_{k_n} is 1 if the CQI of the CUE relay k_n with BS/eNB is above the threshold β within the context of the SINR. When $\delta_{k_n} = 1$, CUE_{k_n} can be included in PRS. It also prevents the use of a relay with a poor CQI to minimize the delay in the uplink transmission.

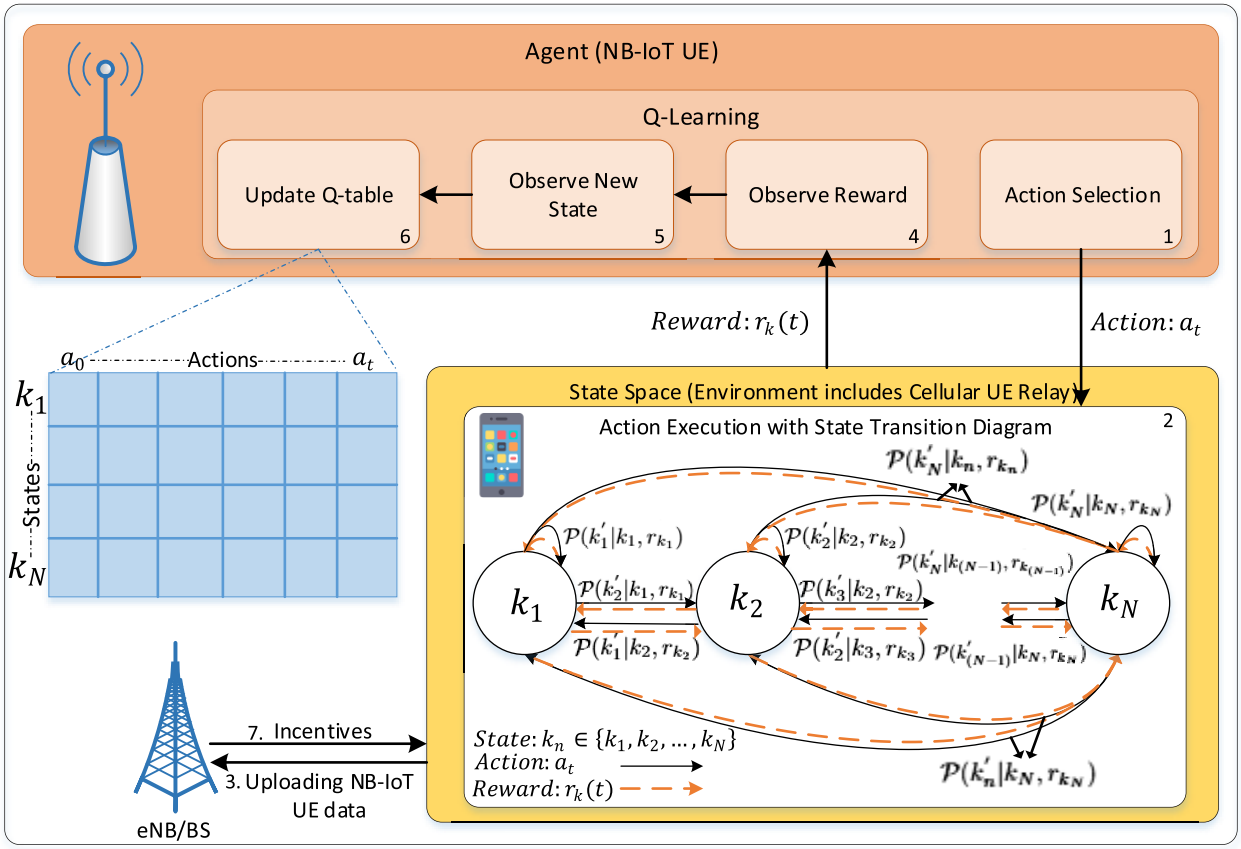


FIGURE 1. The agent–environment interaction in a Markov decision process using Q-Learning, and state-transition diagram in RL-ID2D with k_N states.

In order to solve the problem (5), several techniques are presented in the literature to solve for searching optimal combinations such as branch-and-bound and exhaustive searches. However, these techniques require all possible combinations to be searched to determine the optimal solution because these problems are usually NP-complete or even NP-hard [22]. To solve the problem for the optimal solution, we propose a two-step intelligent algorithm based on RL. Moreover, in practice, only the CQI and transmit power of the NB-IoT UE are unable to ensure a substantial EDR. Therefore, to ensure maximum EDR, we propose a RL-based algorithm that selects the relay based on its reward that reflects the history of practical performance.

V. REINFORCEMENT LEARNING ENABLED RELAY SELECTION PROPOSED SCHEME

This section presents the proposed two-step solution to find an optimal CUE relay. The first step refers to ''optPRS'' (Algorithm 1) and provides insight into our proposed solution to solve the problem (5). When the NB-IoT UE has critical data to upload, it broadcasts a pilot signal to R CUEs. The CUEs analyze the received pilot signal by determining their SINR threshold β and channel gain g with BS/eNB. The CUEs also compare the transmission power of the NB-IoT UE threshold ζ by analyzing the received power of the pilot signal. If the CUE satisfies the threshold values, it qualifies as a relay; otherwise, it withdraws itself. The NB-IoT UE forms a *PRS* by receiving the response from CUEs that are available for relaying the data. The *PRS* forms the state-space for RL-ID2D. Although the parameters considered in Algorithm 1 are sufficiently technical to ensure a good EDR, in dynamic environments, it is difficult to guarantee a good EDR. To consider the practical performance in a dynamic scenario, we propose a solution based on QL in Algorithm 2. Before considering the proposed RL-ID2D, it is critical to understand the formulation of MDP, and the relation between Bellman's equation and QL optimality.

A. MARKOV DECISION PROCESS

The MDP forms the basis of RL, as RL is the learning methodology for sequential process [23]. Such problems can be modelled as MDP represented by the tuple $(\mathcal{K}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ as shown in the Fig. 1. Where $\mathcal{K}$ represents finite state-space K, $\mathcal{A}$ is a finite action space of agent (a set of possible actions), $\mathcal{P}$ is the transitional probability matrix which determines the probability of transition from current state $k_n(t)$ to next state $k_n(t+1)$, and the $\mathcal{R}$ represents the reward function which determines the reward for the agent while moving from one state to another ($r : \mathcal{K} \times \mathcal{A} \times \mathcal{K} \leftarrow \mathbb{R}$). If the dynamics of the environment are fully known, dynamic programming

Algorithm 1 Optimal Selection of PRS (optPRS)

INPUT: ($\beta, \zeta, P^{NB-IoT}_{t_{(UE\rightarrow k_n)}}, SINR$)
for ($r = 1, \ldots, R$) **do**
  if $SINR > \beta$ **then**
    if $P^{NB-IoT}_{t_{(UE\rightarrow k_n)}} \leq \zeta$ **then**
      Push $r(t)$ UE in a new array *PRS*
    else
      withdraw from selection process
    end if
  else
    withdraw from selection process
  end if
end for
OUTPUT: (PRS $K = \{k_1, k_2, \ldots, k_N\}$)

provides the optimal solution. However, in the wireless environment, the transitional probabilities are unknown [24]. The policy π is an sub-element of RL, which is to set the rules for an agent to select an action against the state of the environment. An RL agent at each time step t observes the state $k_n(t)$, takes an action a_t, receives the reward $r(k_n(t), a_t)$, and observes the new state $k_n(t+1)$. The goal of RL agent is to develop a optimal policy $\pi(a_t|k_n(t))$ to know the dynamics of the environment to take best action on a state for optimal solution [14], [25].

The main objective of policy π is to maximize the accumulative expected reward in the long run (return). The return can be expressed as:

$$G_t = \mathbb{E}\left[\sum_{m=0}^{\infty} \gamma^m r(k(t+m), a_{t+m})|k_1 = k(t)\right] \quad (6)$$

The γ is the discount factor to keep the reward bounded, where $0 \leq \gamma \leq 1$. The discount factor determines the present value of future rewards. When the value of γ is set to 0, the agent is more concerned about the immediate reward, that is, $r_k(t)$. As the value of γ approaches 1, the agent takes the future reward into consideration, which is the reward over the long run.

An agent determines the quality of a state $k_n(t)$ as good or bad using a function known as value function $V(k)$. The value function is measured using following expression:

$$V(k_n) = \mathbb{E}[G(t)|k_n(t) = k_n] \quad (7)$$

Similarly, to determine the best action a_t at a specific state $k_n(t)$, an action-value function Q is used, which is measured as:

$$Q(k_n(t), a_t) = \mathbb{E}[G(t)|k_n(t) = k_n, a_t = a] \quad (8)$$

The state-value $V(k)$ can be evaluated if Q and π are known using:

$$V(k_n) = \sum_{a\in\mathcal{A}} \pi(a|k_n) Q(k_n, a) \quad (9)$$

The V and Q can be related as:

$$V(k_n) = \mathbb{E}_{a\sim\pi_{(a|k_n)}}[Q(k_n, a)] \quad (10)$$

Our main objective is to determine optimal policy π^*, which can easily be derived from optimal Q^*. The value function $Q(k_n(t), a_t)$ quantifies state-action pair, i.e., it determines how good is it to take a specific action a_t at a specific state $k_n(t)$ following the optimal policy. The action-value $Q(k_n(t), a_t)$ can be rewritten using Bellman expectation function as [14], [25]:

$$Q(k_n(t), a_t) = r(k_n(t), a_t) + \gamma \sum_{k'\in K} \mathcal{P}_{kk'}(a) V(k') \quad (11)$$

where k' represents state k_n at time $(t+1)$. Therefore, the Bellman optimality equation for action-value Q^* is expressed as [14], [25]:

$$Q^*(k_n(t), a_t) = r(k_n(t), a_t) + \gamma \sum_{k'\in K} \mathcal{P}_{kk'}(a) \max_{a'} Q^*(k', a') \quad (12)$$

where, a' represents action taken by agent at time $(t+1)$. The optimal policy can easily be deduced from optimal action-value function Q^* by selecting the maximum action-value at each state. This methodology is known as action-value based learning, as the optimal policy is derived from the action-value function [14]:

$$\pi^*(k_n) = arg\max_{a\in\mathcal{A}} Q^*(k_n, a), \quad \forall k_n \in K \quad (13)$$

However, the dynamics (transitional probabilities) in reality environment are unknown. One of the effective RL algorithms to solve Bellman optimality equation is QL [25]. QL is the off-policy temporal difference RL methodology [14]. In QL, the agent interacts with the unknown environment in order to learn and optimize the performance of the system. Off-policy refers to the behavior of the agent, which directly optimizes the action-value Q independently from the policy [26]. This approach streamlines the algorithm and enables quick convergence. The policy determines and updates the action-value of the state-action pairs that are conducted in each iteration as a lookup table using Bellman Equation:

$$Q(k_n(t), a_t) \leftarrow Q(k_n(t), a_t) + \alpha[\triangle Q(k_n(t), a_t)] \quad (14)$$

$$\triangle Q(k_n(t), a_t) = [r_k(t) + \gamma \max_a Q(k_n(t+1), a_t) - Q(k_n(t), a_t)] \quad (15)$$

where α is the step size, which is also known as the learning rate and the value of $\alpha \in (0, 1]$. It can be seen from (14) that if α is set to 0, then the agent will not learn, and when α has a high value such as 0.9, then the agent will learn quickly. $\triangle Q(k_n(t), a_t)$ is the 1-step error estimation with respect to the optimal function Q^*. It improves the action-value Q_t one step closer to the desired optimal action-value Q^* by minimizing expected value of error estimation [27].

1) CONVERGENCE OF QL

The convergence of the QL is guaranteed which means the policy will become arbitrarily close to the optimal policy

over the period of time. The convergence depends upon the following two limits [14]:

- The learning rates must approach zero, but not too quickly. Formally, this requires that the sum of the learning rates must diverge, but the sum of their squares must converge.
- Each state-action pair must be visited infinitely often. This has a precise mathematical definition: each action must have a non-zero probability of being selected by the policy in every state, i.e., $\pi(a|k_n) > 0$ for all state-action pair. In practice, using an ϵ-greedy policy (where $\epsilon > 0$) ensures that this condition is satisfied.

2) ϵ-GREEDY POLICY

In order to maximize the reward, the agent should prefer the learned actions in the past, which provide effective rewards; this is known as exploitation. When the agent explores different actions randomly in search for better action selection, it is known as exploration. One of the best ways to balance exploration and exploitation is by using ϵ-greedy. The agent will explore with ϵ probability and exploit with $1 - \epsilon$. The ϵ-greedy prevents system from premature convergence. Moreover, it raises the probability for the selection of unexplored actions. During exploitation, the agent selects the action greedily using eq. (14) according to the following expression:

$$a_{greedy} = arg\max_{a} Q(k_n(t), a_t) \tag{16}$$

B. REINFORCEMENT LEARNING ENABLED RELAY SELECTION

This work proposes a dynamic relay selection method based on RL. The proposed approach learns the behavior of the CUE relay from the perspective of availability and the EDR associated with it. After the learning period, the proposed RL mechanism intelligently selects the best CUE relay that provides the maximum EDR. The qualitative metric for the selection of the relay is the maximum EDR, which varies with location and SINR. Thus, it is critical to select the optimum CUE relay to achieve the best EDR for reliability and minimum energy consumption. This section presents the modeling of the learning process as a QL policy for the best CUE relay selection.

1) Q-LEARNING FRAMEWORK

The QL comprises an agent (NB-IoT UE) that learns the behavior of the states (the CUE relays) in the environment, a policy (parameters that define the optimal CUE relay, that is, the one with maximum EDR), a reward, and the action-value function Q (accumulated reward). The policy determines the behavior and learning of an agent (NB-IoT UE) at a given time step. Fig. 1 shows the proposed QL-based D2D (RL-ID2D) communication model for NB-IoT with the elements. In this study, the following parameters of QL are considered:

a: POLICY

The main objective of policy π in QL is to maximize the accumulative reward, which is an action-value function of the Q-value using eq. (14). The Q-value reflects the effectiveness of the state (CUE relay) of an environment in terms of the EDR, and it is critical for the NB-IoT UE to select the CUE relay that maximizes the EDR.

b: STATE-SPACE

The environment in this problem is the N number of states (CUE relays), that is, $k_n \in \{k_1, k_2, k_3, \ldots, k_N\}$, at discrete time step $t \in \{0, 1, 2, 3, \ldots, \}$. At each time step t, the NB-IoT UE (agent) observes the state k_n, that is, $k_n \in K$, and takes an action a_t such that $a_t \in A$ according to the policy π. Mathematically a state $k_n(t)$ is expressed as follows:

$$k_n(t) = (P^{NB-IoT}_{t_{(UE\rightarrow k_n)}} \leq \zeta, \delta_{k_n} = 1) \in K \qquad \forall k_n \in PRS \tag{17}$$

c: ACTION AND REWARD

The action a_t in RL-ID2D is defined as the selection of relay UE $k_n(t)$ from PRS. The reward is the quantitative performance metric of action $a_t(k)$ on a particular state. In this study, the reward $r_k(t)$ is the reward for choosing the CUE relay k_n at time step t, which uploads the data of NB-IoT UE to the BS/eNB in a two-hop manner by exploiting D2D communication. Two values for reward are considered: 1 or 0. The reward is 1 when the selected CUE relay successfully uploads the data of NB-IoT UE to the BS/eNB, and an acknowledgment is received. Otherwise, the reward is 0. The a_t and $r_k(t)$ are related as expressed in the action-value function (Q-value eq. (14)). Mathematically a_t and $r_k(t)$ is expressed as follows:

$$a_t = (k_n(t) \in K) \tag{18}$$

$$r_k(t) = \begin{cases} r^+ = 1, & \sum_{l=1}^{L} EDR^{l(k_n)} = 1 \\ r^- = 0, & otherwise \end{cases} \tag{19}$$

d: TRANSITIONAL PROBABILITY

The transitional probability matrix $P_{kk`}$ is expressed as follows:

$$P_{kk`} = \begin{bmatrix} P_{11} & \cdots & P_{1t} \\ \vdots & \ddots & \vdots \\ P_{N1} & \cdots & P_{Nt} \end{bmatrix}$$

The rows represent states K_N, where $N = \{1, 2, \cdots, N\}$ and columns represents action a taken at time step t. In reality, the dynamics (transitional matrix) is unknown, as explained above. In QL, the transitional matrix is replaced by $Q-value$ function, which is shown in the Fig. 1.

e: Q-VALUE FUNCTION

The action at each time step is categorized as good or poor according to the reward it gains from the environment. However, the productivity of action in the given state over the long run is determined by the action-value function, which

is the Q-value. The Q-value of a state-action pair specifies the accumulated reward that an agent achieves at a particular state over the long run. It is possible that a state has a low immediate reward, but has a high accumulated reward. The Q-value in QL is calculated and updated using eq. (14).

f: INCENTIVE FOR CUE RELAY

It is necessary to determine why CUEs will allow the personal resources and device privacy to be shared with NB-IoT UE, as practically CUEs are selfish and unwilling to share communication resources. The answer is defined in the Smart Media Pricing (SMP) framework in [28]. The SMP proposes that the CUEs will price their resources used in relay transmission in terms of incentives from service providers, as shown in Fig. 1. The price for these incentives is paid by the NB-IoT UE or the service provider depending on whether the relaying transmission is an uplink or downlink, respectively.

2) RL-ID2D

Algorithm 2 Intelligent Device-to-Device Communication (RL-ID2D)

INPUT ($PRS, \alpha, \gamma, Q(k_n, a), \sigma$)
Initialize: ($Q(k_n, a)$, α =step size, σ, γ)
Select the first $k_n(t)$ randomly from PRS
for ($l = 1, \ldots, L$) **do**
  ϵ = random ([0 → 1])
  if $\epsilon \leq \sigma$ **then**
    choose $k_n(t)$ randomly from the PRS for exploration
    if transmission successful **then**
      $r_k(t) = 1$
    else
      $r_k(t) = 0$
    end if
  else
    choose $k_n(t)$ with highest Q-value using eq. (16)
    if transmission successful **then**
      $r_k(t) = 1$
    else
      $r_k(t) = 0$
    end if
  end if
  update $Q(k_n(t), a_t)$ using eq. (14)
  update $k_n \leftarrow k_n(t+1)$
end for

Algorithm 2 provides detailed insight into the proposed RL-ID2D. In the RL-ID2D, the NB-IoT UE inputs the information of PRS into the QL-based relay selection mechanism, which is explained as follows:

- The step size α such that $\alpha \in (0, 1]$, discount factor $0 \leq \gamma \leq 1$, and $\epsilon > 0$ are set.
- Initializes $Q(k_n(t), a_t)$ for all states $k_n \in K$ and action $a \in \mathcal{A}$ except for a terminal state, which is $Q(terminal) = 0$.
- Select the initial relay UE randomly as an initial state.
- For each time step t, ϵ is randomly generated between $\{0 \rightarrow 1\}$. If the value of $\epsilon \leq \sigma$, the NB-IoT UE explores to find the best relay UE by selecting the relay UE randomly from PRS. Otherwise, it exploits by selecting the relay UE that has the maximum Q-value from the Q-matrix using (16).
- The reward at each time step is observed, whether it is exploration or exploitation, and the reward matrix is updated. It should be noted that the reward is 1 for successful transmission of NB-IoT UE data to BS/eNB with good PDR; otherwise, the reward is 0.
- After the reward is recorded, RL-ID2D updates the $Q(k_n(t), a_t)$ using (14).
- After the learning period ends, the RL-ID2D eventually selects the best relay that maximizes the EDR with minimum overhead.

3) COMPLEXITY ANALYSIS OF RL-ID2D

The complexity of the algorithm can be determined by analyzing the steps of optPRS and RL-ID2D. The optPRS will work until R CUEs in the network and all the other steps are single-step operations. Therefore, the complexity of optPRS is $O(R)$. The steps in RL-ID2D are single-step operations with the exception of *step* 14, in which RL-ID2D has to search for the maximum $Q(k_n(t), a_t)$. The state-action space in our environment is a single-dimension array as the only action a is defined in this article. Therefore, we can simply write that the complexity of RL-ID2D in the worst-case scenario is $O(K)$. The total complexity of the two-step solution is $O(R) + O(K)$, which means that in the worst-case scenario, the complexity directly scales with the number of CUEs in the network.

VI. PERFORMANCE EVALUATION

To validate the performance and efficiency of our proposed scheme, we performed simulations and compared the results with the state-of-the-art opportunistic [15] and deterministic [16] schemes. Moreover, a comparison with random relay selection is also presented. The simulation results show that our proposed intelligent methodology successfully maximizes the EDR in the uplink of NB-IoT systems.

A. SIMULATION SETUP

The simulation setup consists of R users that are independent, and randomly distributed in a single cell. Channel conditions are dynamics based on SINR value and CUEs are not static. First, we segregate the R users into NB-IoT and CUE relays, and then select one of the NB-IoT users that tries to upload its data packets to the eNB/BS using the selected CUE relay. The PRS is updated based on CQI and availability of CUE using optPRS. The NB-IoT is deployed in in-band mode, in which a PRB of 180 kHz is allocated for NB-IoT UE data within cellular band. In this study, it is assumed that upon selection of CUE relay the PRB is randomly assigned to NB-IoT UE within CUE's frame. The maximum transmit power of the NB-IoT UE is set to 14-18 dBm, which is

standardized by 3GPP for NB-IoT to limit the interference with cellular UEs. The SINR threshold β is assumed to be 13 dB. The agent in RL model learns by interacting with the state-space of environment at discrete time steps, therefore, simulation results are presented over time steps (No. of iterations). The QL parameters (α, γ, σ) are considered to be variable, and simulations are carried out with different parameters. Table 2 shows different simulation parameters along with Q-learning parameters considered for simulation study.

B. SIMULATION RESULTS

Fig. 2(a) illustrates the impact of considering QL for the selection of the best relay in the context of EDR. Fig. 2(a) depicts the results with different values of (α, γ, σ). The increasing value of σ shows that the agent (NB-IoT UE) will explore the environment more in search of a better CUE relay. The curve shown in purple indicates that the agent learning at the rate $\alpha = 0.1$ with $\gamma = 0.3, \sigma = 0.7$ converges most quickly between 80 and 100 iterations. The downward spikes in the graph show the penalty received when the selected CUE relay fails to deliver the packet. The failure is caused due to the selection of CUE relay while exploration or exploitation, where the selected CUE UE either has the poor SINR or unavailable at the time of uplink transmission due to dynamic conditions. It is evident from the graph that as the exploration σ, the learning rate α, and discount factor γ increases, i.e., value approaches 1, the agent explores and learns more with interest in long term reward for better CUE relay selection. Therefore, the cumulative Q-value increases and the downward spike due to loss in EDR also decreases. The convergence of the graph also shows the convergence of policy toward optimal selection, that is, the CUE relay is available and provides a good EDR.

Fig. 2(b) shows a comparison of achieved EDR using RL-ID2D for different parameters of QL. The difference in the achieved EDR is explained by the exploration and exploitation dilemma as explained in section VA. Increasing the exploration by increasing the value of σ and increasing learning rate α, and discount factor γ allows the agent (NB-IoT UE) to learn more and to search for a better CUE relay, which is essential in a dynamic environment and yields a better-accumulated reward in the long-run. Increasing the exploitation allows the agent to take the action based on past experiences, which in this case is the Q-value. The exploitation focuses on maximizing the immediate reward and allows the agent to act greedily. The uncertainty in exploration is that the action which produces a better reward is unknown. However, it is better to explore non-greedy actions if there are many time steps ahead in which they may be subsequently exploited, which is reflected in the result, that is, increasing the exploration increases the EDR. Moreover, the result depicts that how changing QL parameter affects in achieving the near optimal EDR.

Fig. 2(c) depicts the adaptive nature of RL-ID2D in dynamic environment. The EDR achieved in dynamic channel on every iteration and adaptiveness of RL-ID2D is shown. The result is simulated on fixed 18 dB transmission power of Nb-IoT UE with $\sigma = 0.1, \alpha = 0.8, \gamma = 0.9$. The EDR at every iteration is shown by blue squared dot. In case of the transmission failure, the EDR drops to zero %, which can be seen as a gap circled at iteration number 190 in Fig. 2(c). The failure occurs when agent (NB-IoT) selects a state (CUE relay) while exploration or exploitation, which either has the poor SINR or is unavailable at the time due to dynamic conditions. A zoomed-in version is included within the Fig. 2(c) for better visual of iteration number 190. RL-ID2D quickly adapts and maintain the EDR closer to 98% in the very next iteration. Similar behavior can be seen in the later iterations. The quick adaptiveness is due to the Algorithm 1 which makes sure to update the PRS with available and eligible CUE relays. It can be seen that as the model converges at 100^{th} iteration, the proposed methodology also converges to select best cellular relay with optimal EDR. The EDR becomes stable more or less at 98% after system converges at 100^{th} iteration.

Fig. 2(d) provides a comparison of our proposed RL-ID2D with a randomly selected CUE relay for D2D communication in terms of the EDR achieved. The graph clearly shows a significant difference in the EDR achieved using both methods. Moreover, it also depicts that increasing R increases the EDR. This can be explained by the fact that increasing R increases the probability of availability of potential CUE to be used as a D2D relay, which in turn augments EDR. In addition, the rising behavior of the curves can be understood by the fact that increasing the transmission power of the NB-IoT UE strengthens the link between the NB-IoT UE and the CUE relay, which augments the $\text{PDR}_{UE \rightarrow CUE_k}$. The results show that our proposed RL-ID2D converges to achieve an EDR of approximately 98%.

The coverage area is one of the most important parameters in communication, and directly affects the data delivery. Fig. 2(e) shows the results of EDR with varying coverage areas and transmit power of NB-IoT UE $P_t = P^{NB-IoT}_{t_{(UE \rightarrow k)}}$ values. It also presents a comparison of random selection and RL-ID2D for a fixed number of users, $R = 50$. The result demonstrates that increasing the transmission range degrades the performance significantly when selecting the CUE relay randomly without considering the SINR and transmission power of the NB-IoT UE. While increasing the coverage area, RL-ID2D enables the NB-IoT UE to upload the data to BS/eNB with 96% EDR transmitting at 18 dBm and 90% EDR, even with a 5-dBm transmission power. This behavior is explained by the fact that the two-step RL-D2D ensures that even when $P^{NB-IoT}_{t_{(UE \rightarrow k)}}$ is low, it selects the CUE relay with optimal CQI, and QL further enhances the performance by exploration and exploitation. The opportunistic model claims that the EDR of their proposed solution approaches 98% at 8 dB transmission power. However, the D2D communication range considered in their performance evaluation is only 40 meters. On the other hand, it is evident from the Fig. 2(e), our proposed RL-ID2D outperforms by achieving 98% EDR

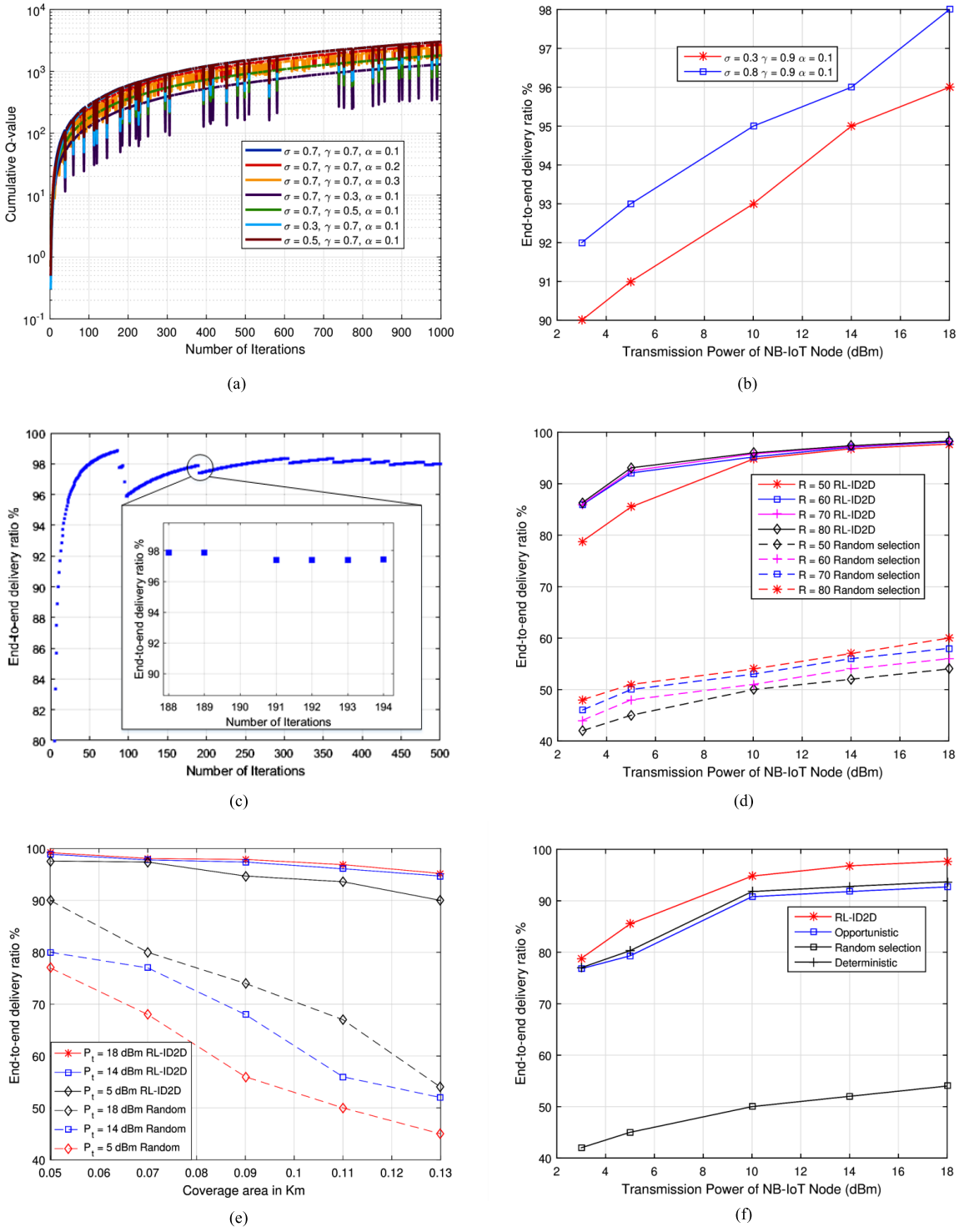


FIGURE 2. **(a) Convergence of learning estimate (cumulative *Q*-value) for varying QL parameters; (b) impact of considering QL for the selection of the best relay in the context of EDR; (c) adaptive behavior of RL-ID2D; (d) comparison of EDR versus the varying transmit powers of NB-IoT users for different fixed values of *R*; (e) comparison of EDR versus the coverage area in Km for different transmit power levels of the NB-IoT users over a fixed value of $R = 50$ users; (f) comparison of EDR with the varying transmit power values of NB-IoT users.**

at 70 meters of D2D communication range even with 5 dB transmission power.

Fig. 2(f) provides a comparison of RL-ID2D with state-of-the-art opportunistic [15] and deterministic [16] schemes. The D2D communication range considered in the given result is 130 meters. It can be seen that RL-ID2D outperforms other techniques, and it performs better than the opportunistic model because the opportunistic model offers the NB-IoT

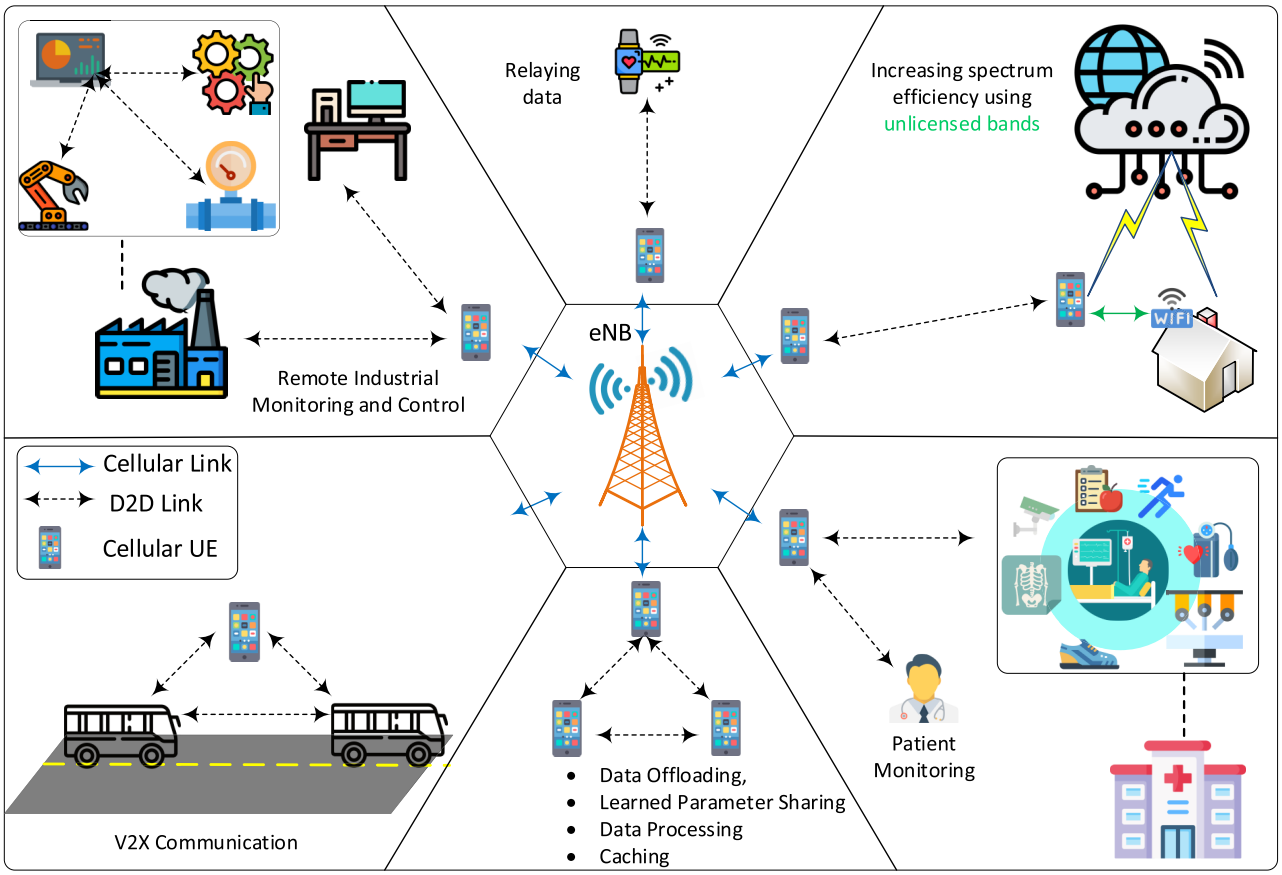


FIGURE 3. Use cases of D2D communication in B5G network.

UE a CUE relay in an opportunistic manner. The NB-IoT UE has to seize the opportunity to upload the data in a two-hop manner; if it fails, then NB-IoT UE has to wait for the next duty cycle. It also promotes dropping the packet after a certain threshold time. However, the deterministic model ensures the availability of the CUE relay by selecting the relay in a deterministic manner using BS/eNB. BS/eNB selects the relay after receiving the request from the NB-IoT UE to provide a CUE relay, and informs NB-IoT UE. This approach guarantees that the NB-IoT UE gets a CUE relay. However, the deterministic model augments the system overhead by increasing control signals. On the other hand, RL-ID2D improves the relay selection process by incorporating QL, which intelligently selects the optimal relay in dynamic environment. Moreover, step 1 'optPRS' guarantees the availability of the CUE relay by updating PRS periodically with CQI, as prescribed in the LTE-A standard [20].

VII. APPLICATION AREA AND FUTURE RESEARCH DIRECTION

D2D communication offers vast deployment scenarios and applications not only limited to 5G, but in B5G and 6G as well. Standalone and non-standalone development mode for B5G are standardized in Release 15-17. Non standalone development mode is based on LTE-A core network and offers back compatible with LTE-A. The standardizing working group is active to improve D2D communication in latest Releases 16 and 17, where it is termed as ProSe. The 3GPP has also introduced a new LPWA in Release 12-17 for mMTC termed as LTE for MTC (LTE-M), which has integrated features of LTE and an improved version of NB-IoT [29]. Therefore, this study is not only limited to NB-IoT and LTE-A, but can be modified for other enabling technologies as well. The application area of D2D in real-life is not only limited to relaying the data, but it includes data processing, data forwarding, cooperative learning, data caching, and enhancing spectrum efficiency as shown in the Fig. 3. Usually, D2D communication involves two tier network communication in all these scenario mentioned above. Zhang *et al.* in [9], presents a detailed vision of application area in B5G network, and explains how D2D communication can assist in data processing among edge devices. Moreover, the article presents the enabling technologies other than 5G along with specifications for D2D communication. The D2D standardization efforts by 3GPP till date in every release are very well presented in [12]. Cooperative communication and cooperative learning is considered to have prime role to support fully autonomous B5G and 6G network, which are well described in [30]. Therefore, investigating multi-agent RL (MARL) environment, where multiple devices interact

with each-other and share the already learned parameter using federated learning is an open issue, and future research should study MARL scenarios.

VIII. CONCLUSION

To increase the coverage area and reliability of the IoT system, the NB-IoT introduces an increased number of retransmissions for data and control packets, which degrades the overall energy efficiency and throughput of the system. This paper presents a two-step RL-based intelligent-D2D (RL-ID2D) communication mechanism for uploading the data of the NB-IoT UE to the BS/eNB in a two-hop manner. The first step ensures the optimal channel conditions and availability, whereas the second step of RL-ID2D ensures the selection of the relay with optimal practical performance. Simulation results show that RL-ID2D selects the cellular relay with the highest probability of availability with a good EDR.

REFERENCES

[1] A. Nauman, Y. A. Qadri, M. Amjad, Y. B. Zikria, M. K. Afzal, and S. W. Kim, "Multimedia Internet of Things: A comprehensive survey," *IEEE Access*, vol. 8, pp. 8202–8250, 2020.

[2] *Artificial Intelligence and Tactile Healthcare for Mitigating the Impact of COVID-19*. Accessed: Nov. 20, 2021. [Online]. Available: https://cmte.ieee.org/futuredirections/tech-policy-ethics/july-2021/artific%ial-intelligence-and-tactile-healthcare-for-mitigating-the-impact-of-covid-19/%

[3] H. Samani and R. Zhu, "Robotic automated external defibrillator ambulance for emergency medical service in smart cities," *IEEE Access*, vol. 4, pp. 268–283, 2016.

[4] J. Chen, K. Hu, Q. Wang, Y. Sun, Z. Shi, and S. He, "Narrowband Internet of Things: Implementations and applications," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2309–2314, Dec. 2017.

[5] S. Popli, R. K. Jha, and S. Jain, "A survey on energy efficient narrowband Internet of Things (NBIoT): Architecture, application and challenges," *IEEE Access*, vol. 7, pp. 16739–16776, 2019.

[6] C. B. Mwakwata, H. Malik, M. M. Alam, Y. L. Moullec, S. Parand, and S. Mumtaz, "Narrowband Internet of Things (NB-IoT): From physical (PHY) and media access control (MAC) layers perspectives," *Sensors*, vol. 19, no. 11, p. 2613, Jun. 2019.

[7] C. Yu, L. Yu, Y. Wu, Y. He, and Q. Lu, "Uplink scheduling and link adaptation for narrowband Internet of Things systems," *IEEE Access*, vol. 5, pp. 1724–1734, 2017.

[8] L. Ji, B. Han, M. Liu, and H. D. Schotten, "Applying device-to-device communication to enhance IoT services," *IEEE Commun. Standards Mag.*, vol. 1, no. 2, pp. 85–91, Jul. 2017.

[9] S. Zhang, J. Liu, H. Guo, M. Qi, and N. Kato, "Envisioning device-to-device communications in 6G," *IEEE Netw.*, vol. 34, no. 3, pp. 86–91, Jun. 2020.

[10] U. N. Kar and D. K. Sanyal, "A critical review of 3GPP standardization of device-to-device communication in cellular networks," *Social Netw. Comput. Sci.*, vol. 1, no. 1, pp. 1–18, Oct. 2019.

[11] *3GPP Releases 16, 17 & Beyond*, 5G-Amer., Jan. 2021. [Online]. Available: https://www.5gamericas.org/3gpp-releases-16-17-beyond/

[12] H. Wu, X. Gao, S. Xu, D. O. Wu, and P. Gong, "Proximate device discovery for D2D communication in LTE advanced: Challenges and approaches," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 140–147, Aug. 2020.

[13] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 4th Quart., 2014.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[15] Y. Li, K. Chi, H. Chen, Z. Wang, and Y. Zhu, "Narrowband Internet of Things systems with opportunistic D2D communication," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1474–1484, Jun. 2018.

[16] A. Nauman, M. A. Jamshed, Y. Ahmad, R. Ali, Y. B. Zikria, and S. W. Kim, "An intelligent deterministic D2D communication in narrow-band Internet of Things," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2019, pp. 2111–2115.

[17] A. Nauman, M. A. Jamshed, R. Ali, K. Cengiz, and S. W. Kim, "Reinforcement learning-enabled intelligent device-to-device (I-D2D) communication in narrowband Internet of Things (NB-IoT)," *Comput. Commun.*, vol. 176, pp. 13–22, Aug. 2021.

[18] L. Militano, A. Orsino, G. Araniti, A. Molinaro, and A. Iera, "A constrained coalition formation game for multihop D2D content uploading," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2012–2024, Mar. 2016.

[19] L. Militano, A. Orsino, G. Araniti, and A. Iera, "NB-IoT for D2D-enhanced content uploading with social trustworthiness in 5G systems," *Future Internet*, vol. 9, no. 3, p. 31, Jul. 2017.

[20] J. Parikh and A. Basu, "Effect of mobility on SINR in long term evolution systems," *ICTACT J. Commun. Technol.*, vol. 7, pp. 1239–1244, Mar. 2016.

[21] Z. Nain, A. Musaddiq, Y. A. Qadri, A. Nauman, M. K. Afzal, and S. W. Kim, "RIATA: A reinforcement learning-based intelligent routing update scheme for future generation IoT networks," *IEEE Access*, vol. 9, pp. 81161–81172, 2021.

[22] C.-Y. R. Huang, C.-Y. Lai, and K.-T. T. Cheng, "Fundamentals of algorithms," in *Electronic Design Automation*, L.-T. Wang, Y.-W. Chang, and K.-T. T. Cheng, Eds. Boston, MA, USA: Morgan Kaufmann, 2009, pp. 173–234.

[23] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2020.

[24] A. Nauman, Y. A. Qadri, R. Ali, and S. W. Kim, "Machine learning-enabled Internet of Things for medical informatics," in *Machine Learning, Big Data, and IoT for Medical Informatics* (Intelligent Data-Centric Systems), P. Kumar, Y. Kumar, and M. A. Tawhid, Eds. New York, NY, USA: Academic, 2021, pp. 111–126.

[25] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, May 1992.

[26] R. Ali, B. Kim, S. W. Kim, H. S. Kim, and F. Ishmanov, "(ReLBT): A reinforcement learning-enabled listen before talk mechanism for LTE-LAA and Wi-Fi coexistence in IoT," *Comput. Commun.*, vol. 150, pp. 498–505, Jan. 2020.

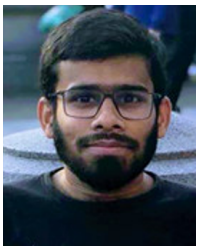
[27] F. S. Melo and M. I. Ribeiro, "Convergence of Q-learning with linear function approximation," in *Proc. Eur. Control Conf. (ECC)*, Jul. 2007, pp. 2671–2678.

[28] W. Wang and Q. Wang, "Price the QoE, not the data: SMP-economic resource allocation in wireless multimedia Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 74–79, Sep. 2018.

[29] *The 5G Evolution: 3GPP Releases 16–17*. Accessed: Sep. 7, 2021. [Online]. Available: https://www.wifi.org/download.php?file=/sites/default/files/private/%Economic_Value_of_Wi_-Fi_Highlights_202102_0.pdf

[30] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, Dec. 2020.

ALI NAUMAN received the B.E. degree in electrical (telecommunication) engineering from COMSATS University Islamabad, Pakistan, in 2013, and the M.S. degree in wireless communications from the Institute of Space Technology, Islamabad, Pakistan, in 2016. He is currently pursuing the doctoral degree with the Wireless Information Networking Laboratory (WINLab), Department of Information and Communication Engineering, Yeungnam University, Gyeongsangbuk-do, Republic of Korea. His research interests include artificial intelligence-enabled wireless networks for healthcare, multimedia, and Industry 5.0. His research interests also include radio resource management and allocation for 5G and beyond-5G (B5G) networks, routing protocols, the Internet-of-Everything (IoE), URLLC, Tactile Internet (TI), and artificial intelligence.



MUHAMMAD ALI JAMSHED (Member, IEEE) received the M.Sc. degree in wireless communications from the Institute of Space Technology, Islamabad, Pakistan, in 2016, and the Ph.D. degree from the University of Surrey, Guildford, U.K, in 2021. He is a Postdoctoral Research Assistant with the James Watt School of Engineering, University of Glasgow, U.K. He served briefly as a Wireless Research Engineer at BriteYellow Ltd., U.K., and then moved to the James Watt School of Engineering, University of Glasgow, as a Postdoctoral Research Assistant. His main research interests include EMF exposure reduction, low SAR antennas for mobile handsets, machine learning for wireless communication, backscatter communication, and wireless sensor networks. He was nominated for the Departmental Prize for Excellence in Research at the University of Surrey, in 2019. He is serving as a Reviewer for IEEE Wireless Communication Letters. Moreover, he served as a Reviewer, the TPC Chair, and the Session Chair for many well-known conferences, i.e., ICC, WCNC, VTC, GlobeCom, and other scientific workshops.

YAZDAN AHMAD QADRI received the bachelor's and master's degrees in electronics and communication engineering from LP University, India, in 2016. He is currently pursuing the Ph.D. degree with the Wireless Information Networking Laboratory (WINLab), Department of Information and Communication Engineering, Yeungnam University, Republic of Korea. His research interests include enabling technologies for Medicine 4.0, which include wireless body area networks and complementing technologies, such as ultra-reliable low-latency communication (URLLC) in 5G, Tactile Internet, and artificial intelligence.

RASHID ALI (Member, IEEE) received the B.S. degree in information technology from Gomal University, Pakistan, in 2007, the master's degree in computer science (advanced networks design) and the master's degree in informatics from University West, Sweden, in 2010 and 2013, respectively, and the Ph.D. Diploma degree in information and communication engineering from the Department of Information and Communication Engineering, Yeungnam University, South Korea, in February 2019. Between 2007 and 2009, he worked for Wateen Telecom Pvt. Ltd., Pakistan, as a WiMAX Engineer with the Operations and Research Department. From July 2013 to June 2014, he worked for COMSATS University Islamabad, Vehari, Pakistan, as a Lecturer. He has also served as a Postdoctoral Research Fellow at the Department of Information and Communications Engineering, Yeungnam University. Currently, he is working as an Assistant Professor with the School of Intelligent Mechatronics Engineering, Sejong University, South Korea. His research interests include next-generation wireless local area networks (IEEE 802.11 ax/ah), unlicensed wireless networks in 5G, the Internet of Things, performance evaluation of wireless networks, named-data/information-centric networking, reinforcement learning techniques for wireless networks, and federated reinforcement learning for next-generation WLANs.

SUNG WON KIM received the B.S. and M.S. degrees from the Department of Control and Instrumentation Engineering, Seoul National University, Republic of Korea, in 1990 and 1992, respectively, and the Ph.D. degree from the School of Electrical Engineering and Computer Sciences, Seoul National University, in August 2002. From January 1992 to August 2001, he was a Researcher at the Research and Development Center of LG Electronics, Republic of Korea. From August 2001 to August 2003, he was a Researcher at the Research and Development Center of AL Tech, Republic of Korea. From August 2003 to February 2005, he was a Postdoctoral Researcher at the Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA. In March 2005, he joined the Department of Information and Communication Engineering, Yeungnam University, Gyeongsangbuk-do, Republic of Korea, where he is currently a Professor. His research interests include resource management, wireless networks, mobile computing, performance evaluation, and machine learning.