IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 24, 2021, accepted March 8, 2021, date of publication March 10, 2021, date of current version March 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065314

Priority-Based Joint Resource Allocation With Deep Q-Learning for Heterogeneous NOMA Systems

SIFAT REZWAN, (Student Member, IEEE), AND WOOYEOL CHOI, (Member, IEEE)
Department of Computer Engineering, Chosun University, Gwangju 61452, Republic of Korea

Corresponding author: Wooyeol Choi (wyc@chosun.ac.kr)

This research was supported by research fund from Chosun University, 2018.

ABSTRACT For heterogeneous demands in fifth-generation (5G) new radio (NR), a massive machine type communication (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable and low-latency communication (URLLC) services have been introduced. To ensure these quality-of-service (QoS) requirements, non-orthogonal multiple access (NOMA) has been introduced in which multiple devices can be served from the same frequency by manipulating the power domain and successive interference cancellation (SIC) technique. To maximize the efficiency of NOMA systems, an optimal resource allocation, such as power allocation and channel assignment, is a key issue that needs to be solved. Although many researchers have proposed multiple solutions, there have been no studies addressing the 5G QoS requirements and three services that coexist in the same network. In this paper, we formulate an optimal power allocation scheme under Karush–Kuhn–Tucker (KKT) optimality conditions incorporating different NOMA constraints to maximize the channel sum-rate and system fairness. We then propose a priority-based channel assignment with a deep *Q*-learning algorithm to maintain the 5G QoS requirements and increase the network performance. Finally, We conduct extensive simulations with respect to different system parameters and can confirm that the proposed scheme performs better than other existing schemes.

INDEX TERMS Deep Q-learning, Internet of Things, joint resource allocation, non-orthogonal multiple access (NOMA).

I. INTRODUCTION

With the rapid increase in the popularity of the Internet of Things (IoT) and cloud computing, the demand for highly reliable data rates and massive connectivity is increasing day by day for wireless communication networks [1]. IoT can provide connections among many types of smart devices, such as mobile devices, smart sensors, and all kind of robots, using radio or wireless access networks to build a massive Eco-system [2]. To fulfill these demands, the 3rd Generation Partnership Project (3GPP) introduced the fifth generation (5G) wireless network that provides three major services [3]. These major services include massive machine type communication (mMTC) that allows massive connectivity for IoT devices, enhanced mobile broadband (eMBB) that provides a high data rate for mobile platforms, and ultra-reliable and low-latency communication (URLLC) that ensures reliability and low latency for highly sensitive and crucial applications [4]–[6]. These services are categorized in terms of their quality-of-service (QoS), where URLLC has a strict QoS policy for high reliability and low latency, eMBB service has a moderate QoS policy, and mMTC has no specific QoS policy except for massive connectivity [7].

These types of QoS policies are extremely difficult to fulfill with the traditional orthogonal multiple access (OMA) due to limited spectrum resources, great transmission losses, and long queuing delays [8], [9]. To maintain these diverse QoS requirements many potential technologies have been introduced into 5G communication network [10]. Among them, non-orthogonal multiple access (NOMA) is gaining popularity because it can support massive connectivity with limited resources, highly reliable transmissions, low transmission delays, and high spectral efficiency [11]–[13]. The key feature of NOMA is that multiple devices can be served

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

from the same radio resource block (RRB), such as time, frequency, and codes, simultaneously utilizing the power domain [14], [15]. NOMA applies superposition coding to combine signals of multiple devices at the transmitter and successive interference cancellation (SIC) at the receiver to differentiate the signals of multiple devices manipulating the power domain [16], [17]. This not only mitigates the multiple access interference, but also increases the spectral efficiency and device fairness [18]. Thus, NOMA can easily maintain strict QoS policies for eMBB, mMTC, and URLLC services. By contrast, with conventional OMA, only one device can be served from each RRB at a time to avoid multiple access interference which is insufficient to support high data rates and massive connectivity [19].

However, there are some major challenges when it comes to resource allocation in the NOMA system, which includes power allocation and channel assignment. One of the major challenges is that joint power allocation and channel assignment involve a mixed-integer program which is a non-deterministic polynomial-time hard (NP-hard) problem [20]–[22]. For example, all possible combinations of channel assignment and power allocation are required to reach an optimal solution which make the system complicated and requires extremely high computational power [23], [24]. When it comes to multi-carrier NOMA the system becomes more complex.

Another problem in multi-carrier NOMA is the channel sum-rate fairness as an increase in the system sum-rate, does not necessarily increase the sum-rate of each channel. The Poor sum-rate of any channel can decrease the performance of the devices assigned to that channel [25]. Moreover, perfect signal decoding using SIC and fulfilling the QoS requirements of 5G services also depends on the power allocation and channel assignment [26]. An imperfect SIC and an inappropriate channel assignment can easily decrease the overall performance of the system. Therefore, in this paper, we investigate the power allocation and channel assignment jointly to overcome the challenges of the downlink NOMA system under various criteria.

A. RELATED WORKS

Optimal resource allocation, such as power allocation and channel assignment, is the key to increase the overall system performance and fulfill the QoS requirements of the 5G network. Many researchers have proposed many approaches to obtain optimal solutions with different performance objectives [27], [28]. The most common objectives are to maximize the overall sum-rate of the system and fulfill the minimum data rate.

Ali *et al.* [27] proposed a power allocation technique with a user grouping scheme for a single-carrier NOMA system to maximize the sum-rate using Lagrange equations under Karush–Kuhn–Tucker (KKT) conditions. The authors have derived the Lagrange equations to obtain an optimal power allocation scheme while considering total power limitation, minimum data rate requirement, and SIC constraints under Karush–Kuhn–Tucker (KKT) conditions. Shao *et al.* [29] derived a dynamic device clustering technique and an optimal power allocation solution using the Nash bargaining solution (NBS) for NOMA system based on the number of devices and channel gains. However, only single-carrier NOMA system for IoT devices is considered. In [7], Shahini *et al.* proposed priority-based URLLC and mMTC device grouping with fixed power allocation scheme. However, no authors considered the presence of URLLC, eMBB, and mMTC services in 5G networks. Parida and Das [30] solved only the non-convex power allocation problem using the difference of two convex functions (DC) programming to maximize the sum-rate of orthogonal frequency division multiple access (OFDMA)-based NOMA system. In another paper [31], Hojeij *et al.* used the water-filling algorithm for resource allocation to obtain the highest sum-rate possible. However, no optimality was provided for the obtained solution.

Nevertheless, the system sum-rate increases when it comes to multi-carrier NOMA. In [1], Zhu *et al.* derived an near-optimal power allocation solution considering two users per channel and iteratively assigned channel to the users. They also considered the minimum data rate constraints for each user while maximizing the sum-rate. However, authors did not consider different services of the 5G network. Choi [28] used convex optimization to approximate the maximization problem for the minimum data rate requirement of users. Ning *et al.* [32] adopted a heuristic approach to solve the power allocation and channel assignment problem of the NOMA system for vehicular ad-hoc networks.

In addition to conventional convex optimization, many researchers explored the machine learning and artificial intelligence sectors to optimize the resource allocation problem of the NOMA system. In [33], Xiao *et al.* proposed fast and dynamic reinforcement learning (RL) based power allocation to maximize sum-rate and spectral efficiency of a multiple-input multiple-output (MIMO) NOMA system in presence of smart jamming. The authors initially formulated the anti-jamming transmission game and derived the Stackelberg equilibrium of the game. Q-learning-based power allocation is then used to allocate power to users against jamming attacks. He *et al.* [34] proposed a joint power allocation and channel assignment for the NOMA system using deep reinforcement learning (DRL). They used the derived near-optimal power allocation from [1] considering two users per channel and performed channel assignment using DRL algorithm consisting an attention-based neural network. The authors then used a DRL algorithm consisting an attention-based neural network to perform channel assignment while maximizing the overall sum-rate and minimum data rate for user fairness. An actor-critic (A2C) RL algorithm was used in [35] to obtain the optimal policy for resource allocation and user scheduling in HetNets with a hybrid energy supply. The actor parameterizes the policy using the Gaussian distribution to take stochastic actions, and the critic evaluates the value function and helps the actor learn the optimal policy.

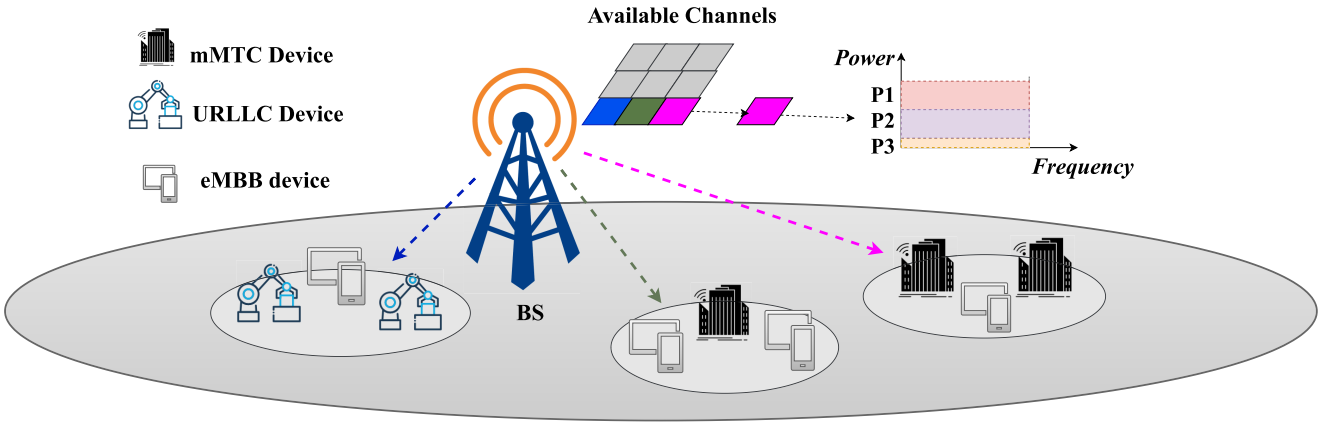

FIGURE 1. Simple multi-carrier NOMA system.

In summary, many researchers found many optimal and near-optimal power allocation solutions for a single-carrier only. Most researchers focused on increasing the overall sum-rate while maintaining a minimum data rate for fairness. However, an increase in the overall sum-rate does not ensure an increase in the sum-rate of each channel. Furthermore, the sum-rate of a device is directly connected with the sum-rate of the channel. Some researchers have also found the optimal and near-optimal solutions for both power allocation and channel assignment problems while considering only eMBB services of 5G network and have not done it for more than two devices per channel. Nevertheless, achieving optimal and near-optimal solutions using conventional methods are very computationally complex. Some researchers have adopted neural networks (NN) to replace the complex methods owing to their low complexity.

B. CONTRIBUTIONS

In this paper, we investigate resource allocation schemes to maximize the performance of multi-carrier NOMA system under multiple performance metrics. We propose a priority-based joint resource allocation scheme with DQL for heterogeneous NOMA system considering the key constraints and services of 5G networks. The contributions of this paper are described as follows:

- We formulate an optimal power allocation scheme that maximizes the overall system efficiency for any given channel assignment using Lagrange multipliers under KKT optimality conditions and incorporates different constraints of NOMA.
- We propose a priority-based channel assignment scheme using deep Q-learning (DQL) to maximize the performance and fairness of multi-carrier NOMA. We prioritize the devices present in the 5G network based on the QoS requirement and categorize them based on URLLC, eMBB, and mMTC services. The agent of the DQL explores the 5G network environment and learns the prioritization and channel assignment to achieve an optimal policy. We use an autoencoder architecture for the policy network, followed by a long short-term memory (LSTM) network.
- We consider different constraints of the NOMA system, including the total power budget of the base station (BS), the minimum data rate requirement of each device, the QoS policies of different services of the 5G network, and the sum-rate maximization with channel fairness constraints.
- We consider maximizing sum-rate (MSR), maximizing channel sum-rate (MCSR), and maintaining the 5G QoS policies as our main objectives.
- Finally, we analyze and compare the proposed schemes in different scenarios with the conventional OMA system.

The remainder of this paper is organized as follows. Section II introduces the problem statement of the NOMA system. The power allocation solution derivation and proposed priority-based channel assignment scheme are discussed in Sections III, and IV, respectively. The simulation results are then analyzed in Section V and some concluding remarks are given in Section VI.

II. PROBLEM STATEMENT

In this section, we discuss the fundamentals of multi-carrier NOMA. We also briefly describe the system model and derive different equations based on the constraints of NOMA system and the objectives of our proposed solution.

A. MULTI-CARRIER NOMA

With NOMA, multiple devices can be served using the same RRB utilizing the power domain for both uplink and downlink transmissions. We consider a simple downlink multi-carrier NOMA system where the BS serves different types of devices at the same time over the wireless channels. Fig. 1 shows, a scenario of 5G network consisting of three different devices. The BS assigns one channel to every three devices, where the signals of the three devices are multiplexed at different power levels. Therefore, the devices receive their desire signals along with the signals of other two devices of that channel as noise or interference. The unwanted signals

will act as noise if the power level of the desired signal is high; otherwise the unwanted signals will act as interference. To decode the desired signal, each device uses SIC technology. SIC decodes the signal with the highest power and subtracts that signal from the main signal until it decodes the desire signal. The perfect SIC depends on the channel state information (CSI) such as signal-to-noise and interference ratio (SINR) [36], and the SINR depends on the channel assignment and power allocation. In this case, the data rate for each device for its channel can be calculated using (1).

$$R_i^k = \log_2\left(1 + \frac{P_i^k \Gamma_i^k}{\sum_{j=1}^{i-1} P_j^k \Gamma_i^k + 1}\right), \quad k, i = 1, 2, 3, \qquad (1)$$

where Γ is the channel to noise ratio (CNR) for the assigned channel k and P is the assigned power. The details of (1) are given in Section II-B.

B. SYSTEM MODEL

We consider a micro-cell of a 5G network consisting of 5G enabled devices with a base station (BS). We also consider the downlink of single-input and single-output (SISO) NOMA system as shown in Fig. 2, where the total number of devices is N and the number of channels is K. There are three types of devices that require three different services of 5G network: eMBB devices $UE_1, UE_2, \ldots, UE_e$; URLLC devices $UL_1, UL_2, \ldots, UL_l$; and mMTC devices $MC_1, MC_2, \ldots, MC_m$. We also consider that the total available bandwidth (BW_t) is divided into all channels having channel bandwidth (BW_{ch}) of 180 kHz. The maximum number of devices per channel is n, which ranges from $2 \leq n \leq N$, and the total number of channels is $K = ceil(N/n)$.

We consider perfect CSI to develop the proposed scheme. However, for a practical wireless environment, we also consider an imperfect CSI to evaluate the proposed scheme. Let us assume that the k^{th} channel is assigned to n devices, where the power allocated to the n^{th} device is P_n and the desired signal of the n^{th} device is x_n. After combining the signals of the n devices, the BS transmits them over the k^{th} channel which can be represented as follows:

$$X^k = \sum_{i=1}^{n} \sqrt{P_i^k} x_i, \quad i = 1, 2, \ldots, n \qquad (2)$$

At the device end, the transmitted signal reaches with path loss component and additive white Gaussian noise (AWGN), which can be represented as

$$y^k = \sum_{i=1}^{n} \sqrt{P_i^k} h_i^k x_i + w^k, \quad i = 1, 2, \ldots, n, \qquad (3)$$

where h_i^k is the channel gain of the i^{th} device and w^k denotes the AWGN with thermal noise power variance, σ_k. After receiving the signal, the receiver uses the SIC technique to decode its signal. Perfect SIC depends on the SINR of the device on the channel that it has been using for communication. Let us consider the CNR of the n^{th} device for

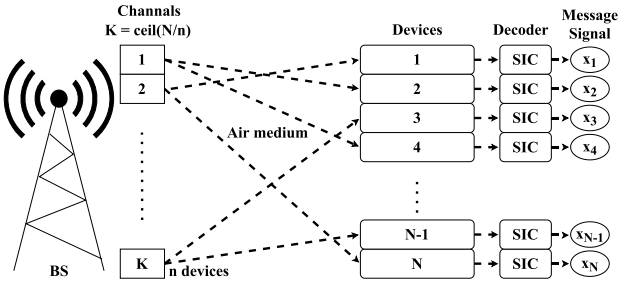

FIGURE 2. System architecture of multi-carrier SISO-NOMA system.

k^{th} channel is

$$\Gamma_n^k = \frac{|\, h_i \,|^2}{\sigma_k}. \qquad (4)$$

We know from the earlier discussion that different power levels are allocated to the devices of a channel. As per NOMA, the highest power is allocated to the device with the lowest CNR and vice versa. For example, for devices having $\Gamma_1^k > \Gamma_2^k > \ldots > \Gamma_n^k$ CNR are assigned with power $P_1^k < P_2^k < \ldots < P_n^k$, respectively. Therefore, the SINR and the data rate for each device of a specific channel can represented as (5) and (1), respectively.

$$\gamma_i^k = \frac{P_i^k \Gamma_i^k}{\sum_{j=1}^{i-1} P_j^k \Gamma_i^k + 1}, \quad i = 1, 2, \ldots, n. \qquad (5)$$

To perform perfect SIC, the BS allocate power to each device above certain threshold level P_{th} as shown in (6). For example, the device with low CNR must have higher power than the sum of other high CNR devices' power for perfect completion of the SIC technique.

$$\left(P_i^k - \left(\sum_{j=1}^{i-1} P_j^k\right)\right) \Gamma_d^k \geq P_{th},$$
$$i = 1, 2, \ldots, (n-1),$$
$$d = n, \ldots, 2, 1,$$
$$k = 1, 2, \ldots, K. \qquad (6)$$

C. PROBLEM FORMULATION

We consider each device has a set of channels $\Gamma_N = \{\Gamma_N^1, \Gamma_N^2, \ldots, \Gamma_N^k\}$ for channel assignment and range of power from $P_N \in [0.01, 0.99] \times P_T$ where P_T is the total power budget per channel for power allocation. In this paper, we focus on the sum-rate as the key performance indicator for the optimization of channel assignment and power allocation in the NOMA system which can be represented as

$$R_{\text{sum}} = \sum_{1}^{K} \sum_{i=1}^{n} \log_2\left(1 + \frac{P_i^k \Gamma_i^k}{\sum_{j=1}^{i-1} P_j^k \Gamma_i^k + 1}\right),$$
$$i = 1, 2, \ldots, n,$$
$$k = 1, 2, \ldots, K. \qquad (7)$$

We also consider the minimum data rate requirement of all devices which can be expressed as

$$\log_2\left(1+\frac{P_i^k\Gamma_i^k}{\sum_{j=1}^{i-1}P_j^k\Gamma_i^k+1}\right)\geq R_i^k,\quad i=1,2,\ldots,n,\quad k=1,2,\ldots,K. \tag{8}$$

The sum of the power per device in a channel must less or equal than P_T, and can be written as

$$\sum_{i=1}^{n}P_i^k\leq P_{\mathrm{T}},\quad k=1,2,\ldots,K. \tag{9}$$

In this paper, we derive an optimal power allocation scheme and propose a priority-based channel assignment with a deep Q-learning algorithm for maintaining the QoS policies of the 5G services, MSR, and MCSR to ensure fairness among the devices and the increase in system performance. As DQL requires power allocation to evaluate the channel assignment and train the DNN, we first derive a power allocation solution for any given channels, and then we build the DQL framework for priority-based channel assignment to obtain an optimal solution for the NOMA system.

III. POWER ALLOCATION

In this section, we derive the optimal power allocation for any given channel while considering different constraints of NOMA to increase the maximum sum-rates and system efficiency. The power allocation solution is derived based on the power allocation solution in [27]. We consider sorting the devices in descending order based on their distances from BS. As our main target is to maximize the sum-rates, we can represent (7) as a maximizing convex function for a given channel k considering (6), (8), and (9), which can be formulated as follows:

$$\begin{aligned}\underset{P_i^k}{\text{maximize}}\ &\sum_{i=1}^{n}\log_2\left(1+\frac{P_i^k\Gamma_i^k}{\sum_{j=1}^{i-1}P_j^k\Gamma_i^k+1}\right)\\ \text{subject to}\ &\log_2\left(1+\frac{P_i^k\Gamma_i^k}{\sum_{j=1}^{i-1}P_j^k\Gamma_i^k+1}\right)\geq R_i^k,\\ &\sum_{i=1}^{n}P_i^k\leq P_{\mathrm{T}},\\ &\left(P_i^k-\left(\sum_{j=1}^{i-1}P_j^k\right)\right)\Gamma_d^k\geq P_{th},\\ &\forall i=1,2,\ldots,n;\ d=n,\ldots,2,1.\end{aligned} \tag{10}$$

The convex problem (10) can also be expressed in Lagrangian form as

$$\begin{aligned}&\mathscr{L}(P,\tau,\nu,\psi)\\ &=\sum_{i=1}^{n}\log_2\left(1+\frac{P_i^k\Gamma_i^k}{\sum_{j=1}^{i-1}P_j^k\Gamma_i^k+1}\right)\\ &=\tau^k\left(P_T-\sum_{i=1}^{n}P_i^k\right)\\ &\quad+\sum_{i=1}^{n}\nu_i^k\left\{P_i^k\Gamma_i^k-\left(\sum_{j=1}^{i-1}P_j^k\Gamma_i^k-1\right)\times\left(\phi_i^k-1\right)\right\}\\ &\quad+\sum_{i=2}^{n}\psi_i^k\left(P_i^k\Gamma_d^k-\sum_{l=1}^{i}P_l^k\Gamma_d^k-P_{Th}\right),\end{aligned} \tag{11}$$

where τ, ν, and ψ are the Lagrange multipliers, $\forall i=1,2,\ldots,n$, and $\phi_i^k=2^{\frac{R_i^k}{KBW_{ch}}}$. Taking the derivatives of (11) with respect to P_i, τ, ν, and ψ, multiple KKT conditions can be found. For n-device NOMA, there are $2n$ Lagrange multipliers resulting in 2^{2n} combinations. For example, for $n=2,3,4,\ldots,8$, the number of combinations are $16,64,256,\ldots,65536$, respectively. However, checking all types of combinations is not computationally feasible. After solving only n equations according to [37] for 2, 3, 4-device NOMA, $2,4,8$ combinations are found that satisfy the KKT conditions, respectively. Therefore, the closed-form solution of the power allocation for n-device NOMA for a given channel k is near-optimal and can be written as

$$\begin{aligned}P_x&=\frac{P_T}{2^{(n-1)}}+\frac{(x-1)P_{th}}{2^{(x-1)}\Gamma_{(x-1)}}-\left(\sum_{i=x}^{n-1}\frac{P_{th}}{2^i\Gamma_i}\right),\\ P_j&=\frac{P_T}{2^{(n-q-2)}}+\frac{P_{th}}{2\Gamma_{(j-1)}}-\left(\sum_{i=j}^{n-1}\frac{P_{th}}{2\Gamma_i}\right),\end{aligned} \tag{12}$$

where $x=1,2$, $j=3,4,\ldots,n$, $q=0,1,\ldots,(n-3)$, and devices have $\Gamma_1^k>\Gamma_2^k>\ldots>\Gamma_n^k$ CNR with power $P_1^k<P_2^k<\ldots<P_n^k$, respectively.

IV. PRIORITY-BASED CHANNEL ASSIGNMENT

In this section, we propose a priority-based channel assignment scheme using deep Q-learning. First, we formulate the channel assignment problem based on the priority, MSR, and MCSR, and then model the channel assignment problem as a reinforcement task and introduce an autoencoder followed by an LSTM network to create the DQL framework. Finally, we use the near-optimal power allocation solution and train the DNN for validation.

A. PRIORITY-BASED CHANNEL ASSIGNMENT

The 5G wireless network provides three different services with different QoS requirements, such as URLLC service has highest QoS requirements, eMBB service has average QoS requirements, and mMTC service has least QoS requirements. We prioritize the devices in the network based on the services they are using and their QoS requirements where the URLLC devices have the highest priority, the eMBB devices have the second-highest priority and the mMTC devices are the least priority devices. The BS sorts the URLLC, eMBB, and mMTC devices in descending order based on their distances from BS. Subsequently, the BS assigns URLLC

devices to the channels with highest gain first, then assigns the eMBB devices and mMTC devices accordingly to the channels available as shown in Fig. 3. This figure shows an illustration of priority-based channel assignment for 3-device NOMA where 4 URLLC, 5 eMBB, and 3 mMTC devices are present. However, assigning channels is subject to the CNR of each device with the BS.

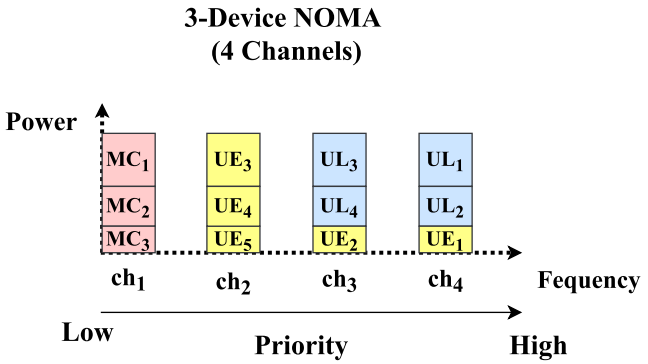

FIGURE 3. Proposed priority-based sample channel assignment for 3-device NOMA system for 12 active devices.

Another main requirement of the optimization of the channel assignment is to maximize the channel and overall sum-rates. The BS have $\binom{N}{n}$ combinations for each channel k to check for maximize the sum-rate. Therefore, the total combination in general is $\sum_{i=1}^{K}\binom{N-(n\times i)}{n}$ for MCSR. When it comes to priority, the low priority devices cannot replace the high priority devices in a channel. However, high or equal priority devices can replace the equal or low priority devices in any given channel. The maximization process incorporating with the priority scheme is computationally complex since the BS has to check all the possible combinations of the device. To reduce the computational complexity, we propose a DQL framework to assign channels to the devices while maintaining the priority and maximizing the sum-rates.

B. DEEP Q-LEARNING FRAMEWORK

In this subsection, we propose a DQL framework and train it to optimize the priority-based channel assignment problem. The deep Q-learning algorithm generally consists of an agent with a deep neural network (DNN) and an environment. The agent interacts with the environment and decides which action to take. The BS acts as an agent and interacts with the environment consisting of URLLC, eMBB, and mMTC devices' information. Initially, the agent starts exploring the environment to collect the channel information of every device. At each time step t, based on the present state s_t of the agent in the environment, the agent predicts an action a_t using the DNN to assign a channel. In return, the agent receives an immediate reward r_t and the next state s_{t+1} from the environment as shown in Fig. 4. The agent receives a good reward r_t if it performs a good channel assignment. By predicting actions, the agent learns about the environment and achieves an optimal channel assignment policy π_c.

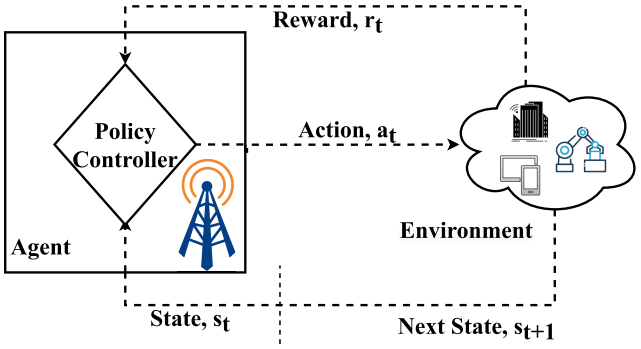

FIGURE 4. Simple Q-learning.

This optimal policy is learned at each time step t by the DNN. The agent updates and improves the policy π_c by repeating the channel assignment process for multiple episodes. One episode terminates when there are no channels left for assignment.

We define the state, action, and reward for use in the proposed DNN as follows:

- **State**: We consider the channel information for each device as the states of the environment. There are N devices having K channel preferences. Therefore, the state space has $N \times K$ elements and can be represented as $S = \{\Gamma_1^1, \Gamma_1^2, \Gamma_1^3, \ldots, \Gamma_1^K, \Gamma_2^K, \Gamma_3^K, \ldots, \Gamma_N^K\}$.
- **Action**: The main action of the agent is to assign channels to the devices which belong to the action space A. At each episode for a set of S, the agent has to take $N \in A$ actions while maintaining one action per K elements from S. For $2, 3, \ldots, n$-device NOMA, the agent can take one action $2, 3, \ldots, n$-times, respectively.
- **Reward**: Whenever the agent completes taking N actions, the agent gets a reward r_t^l for each action. For each correct action, the agent gets a positive reward r_i and when the agent takes correct n actions, the agent gets the sum-rate of that channel as a reward for the taken actions. For example, let us assume a 3-device NOMA. The agent has to assign 3 devices per channel. In this case, when the agent successfully selects an appropriate channel based on priority for a device, the agent gets a positive reward r_i (i.e., 10). If the agent can select the same appropriate channel for 3 devices, the agent gets the sum-rate calculated by (1) as a reward for its 3 actions. The reward function can be defined as

$$r_t^l = \begin{cases} \sum_{i=1}^{n} R_i^k & \text{if } a_p^k = n \\ 0 < r_i & \\ < \sum_{i=1}^{n} R_i^k \text{ for each } a_t^l & \text{if } a_p^k < n \\ 0 & \text{if } a_p^k = 0 \end{cases}, \tag{13}$$

where a_p^k is the number of appropriate action a_t^l taken per channel k and $\forall l = 1, 2, \ldots, N \in A$. Here, we consider maximizing the sum-rate for each channel which results in increased performance and fairness of the whole system.

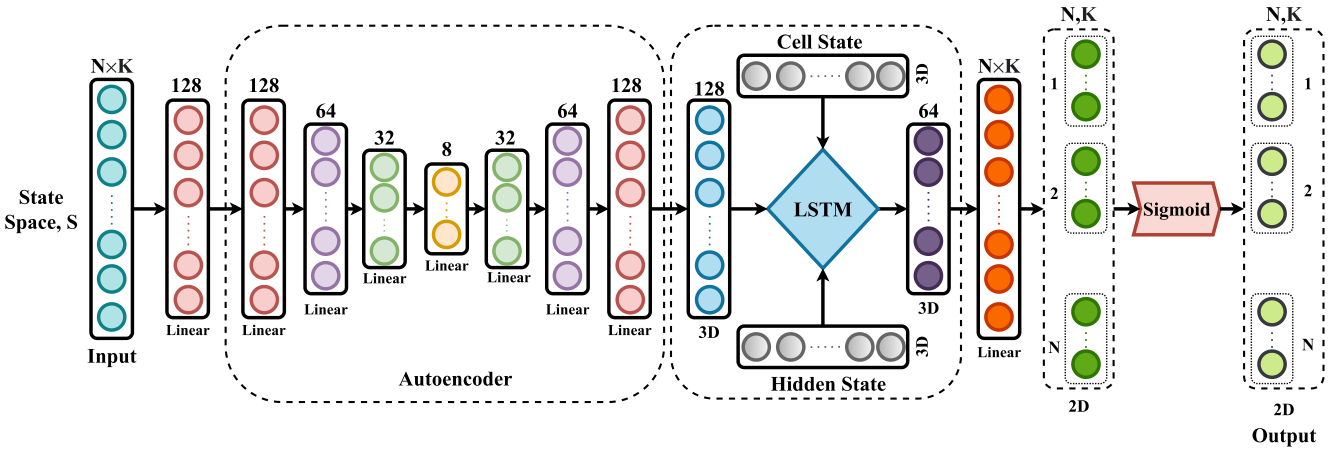

FIGURE 5. Proposed DNN structure.

With the state, action, and reward, we propose the deep neural network (DNN) structure shown in Fig. 5 as the policy controller for channel assignment. The DNN replaces the Q-table and estimates the Q-values for each state-action pair of the environment. Eventually, the DNN approximates the optimal policy for channel assignment. The proposed DNN has two parts, an autoencoder model and an LSTM model. The main goal of the DNN is to derive probabilities for each device-channel pair for each state space, which can be expressed as $Q(S, A)$. These probabilities are the Q-values for DQL.

1) AUTOENCODER

An autoencoder is a feed-forward neural network where the number of inputs is same as the number of output neurons. It compresses the input into a lower-dimensional code and then reconstructs the input data from the code at the output. The autoencoder can easily handle raw input data without any fancy processing or labeling. Therefore, the autoencoder is considered as a part of the unsupervised learning technique [38] and can generate their labels from the training data. The autoencoder has three main parts named an encoder, code, and decoder as shown in Fig. 6. Both the encoder and decoder are fully connected neural networks. The encoder starts with an input layer having 2^n neurons followed by multiple hidden layers having 2^{n-h} neurons, where h is the position of the layer. The number of neurons per hidden layer continues to decrease till the code part of the autoencoder. In this paper, we use 2^3 neurons for the code layer. The decoder part is the mirror image of the encoder ending with an output layer. This type of structure is known as *stacked autoencoder* as the layers are stacked one after another, like a sandwich. Moreover, we use ReLU as an activation function for each layer in the autoencoder.

2) LONG SHORT-TERM MEMORY

Long short-term memory (LSTM) is an evolved form of recurrent neural network (RNN). LSTMs are a special type

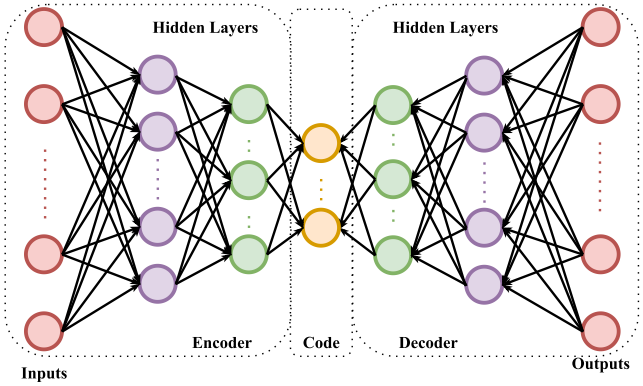

FIGURE 6. Autoencoder architecture.

of RNN that can learn long-term dependencies and remember previous information for future usage. The LSTM network has a chain structure composed of multiple LSTM cells. We use three LSTM cells to build our LSTM network. The structure of a single LSTM cell is shown in Fig. 7 [39]. An LSTM cell has three input and two output parameters. The cell and hidden states are the common parameters between inputs and outputs. The other parameter is the current input. The LSTM cell also contains three sigmoid layers and two tanh layers involving some linear transformations as shown in Fig. 7. Initially, random cell and hidden states are given along with the input for the first LSTM cell. Then the two outputs (hidden state, cell state) become the three inputs of the next cell as shown in Fig. 7.

In this paper, we use an autoencoder having input and output size of 128 and code size 8 followed by an LSTM network having 128 input size, 64 hidden state size, and 3 recurrent layers. Finally, the output of the LSTM is passed through a linear layer and a sigmoid layer to obtain the probabilities of the preferred channels for each device. The state space S is given as the input of our policy network. Initially, the input is first embedded with dimension 128. It then passes through the policy network to generate the channel assigning probabilities, as shown in Fig. 5.

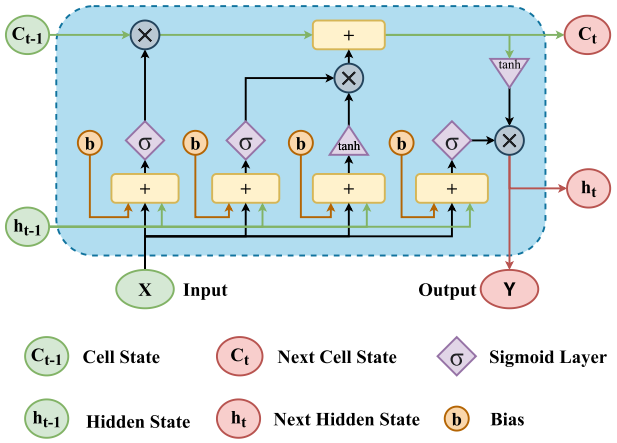

FIGURE 7. An LSTM cell.

C. TRAINING

The proposed DNN is trained gradually with a set of training data $T_{data} = \{S^1, S^2, \ldots, S^{ins}\}$ per episode. For each state space S, the device-channel pairs are selected using ϵ-greedy policy according to the output probabilities from the DNN. An episode terminates when all state spaces are passed through the DNN. The policy to take action for each device per state space can be expressed as

$$a_i^l = \begin{cases} \text{argmax } Q(S^i, A_i^l) & \text{if } \epsilon < \epsilon_{th}; \text{ where } \epsilon_{th} \in (0, 1] \\ \text{random action } [1, K] & \text{otherwise} \end{cases}, \quad \forall l = 1, 2, \ldots, N \in A, \quad \forall i = 1, 2, \ldots, ins. \tag{14}$$

After taking the actions using (14), the agent gets the rewards according to (13) and the next state space S^{i+1}.

To train the DNN, we calculate the loss and optimize the parameters of the DNN performing back-propagation. To calculate the loss, we approximate the optimal Q^*-values for each device-channel pair of S^{i+1} from a different DNN called the target DNN [40]. The target DNN is identical to the policy DNN and initialized by the parameters of the policy DNN. The next state space S^{i+1} is given as an input to the target DNN and from the outputs the optimal Q^*-values are chosen greedily by the agent. Because assigning the channel is a classification problem, we use the categorical cross-entropy loss function to calculate the loss between the optimal Q^*-values and normal Q-values [41]. After calculating the loss, we optimize the policy DNN using the Adam optimizer [42]. To estimate the optimal Q^*-values correctly, we periodically update the target DNN with the parameters of the policy DNN after certain episodes.

For a more stable convergence of the optimal policy, we introduce the experience replay memory (ERM) to the DQL [43]. Initially, the agent explores the environment and saves current states, actions, rewards, and next states (S^i, A_i, r_i, S^{i+1}) as a tuple in the ERM. Subsequently, the agent takes a mini-batch of tuples from the ERM and trains the policy DNN. The ERM continues to be updated for each training data. Fig. 8 and Algorithm 1 summarize the proposed DQL framework and the working flow.

Algorithm 1 Proposed Deep Q-Learning Algorithm

```
Initialize policy and target DQL network with random parameters (p and p').
Initialize experience replay memory (ERM).
Initialize ϵ.
for each episode do
    for each instance do
        for each device do
            Select an channel and add to action space A_i for present state space S^i based on ϵ.
        end for
        Observe the immediate rewards r_i and next state space S^{i+1}.
        Insert (S^i, A_i, r_t, S^{i+1}) in ERM.
        Create a mini-batch with random sample of (S^i, A_i, r_t, S^{i+1}) from ERM.
        for each tuple in mini-batch do
            Obtain Q-values using policy DNN.
            Approximate Q*-values using target DNN.
            Calculate the loss using Q an Q*-values.
            Optimize the parameters p of the policy DNN using Adam optimizer.
        end for
    end for
    p' ← p after certain number of episodes.
end for
```

V. SIMULATION ANALYSIS

In this section, we perform multiple simulations to analyze the performance of the proposed DQL algorithm for priority-based channel assignment and compare the proposed priority-based joint resource allocation (priority-JRA) with the joint resource allocation (JRA) method and dynamic power allocation with fixed channels (DPA-FC) method proposed in [1] and [27], respectively. Moreover, we compare the priority-JRA NOMA system with the conventional OMA system. Finally, we also analyze the system complexity and system convergence varying different parameters.

A. SIMULATION ENVIRONMENT

For the simulation environment, we consider a 5G micro-cell where 24 devices are randomly and uniformly distributed. We only consider three types of devices, URLLC, eMBB, and mMTC devices. We model the channel gain h_i^k of the k^{th} channel for each device based on the Rayleigh fading model, where the path loss exponent, $\eta = 3$. Then we calculate the CNR of each channel for each device using (4) where $\sigma_k = \frac{BW_t \times N_0}{k}$ for $\forall k = 1, 2, \ldots, K$ with $BW_t = 5$MHz and $N_0 = -172$ dBm/Hz.

To analyze the performance, simulation parameters similar to [1], [27] are used as given in Table 1. The parameters of proposed DNN such as weights and biases are initialized

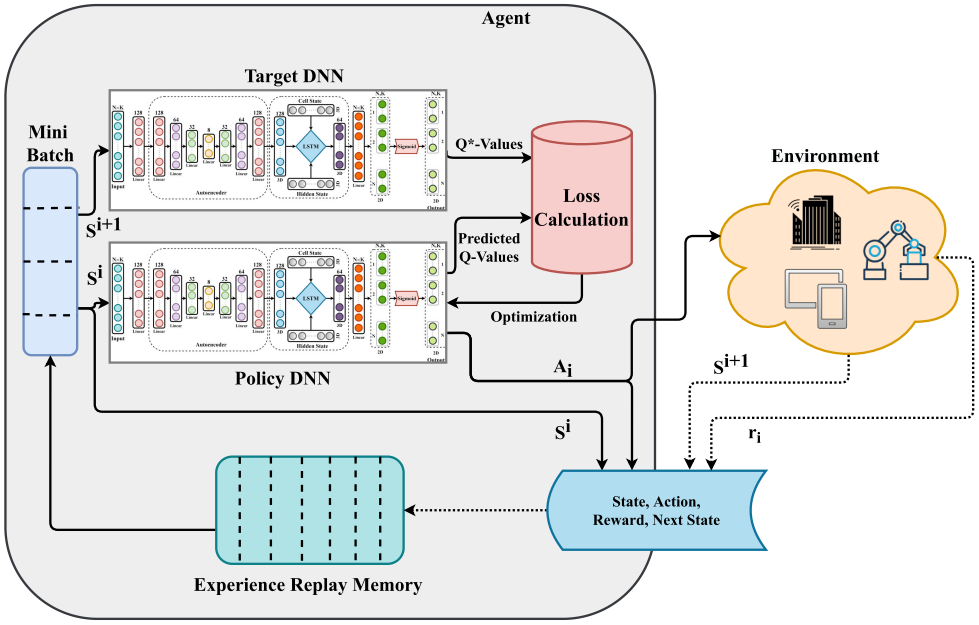

FIGURE 8. Proposed DQL framework.

TABLE 1. Simulation parameters.

Parameter	Value	Parameter	Value
BW_t	5 MHz	BW_{ch}	180 kHz
P_{T}	$2-18$ W	Learning rate	0.01
N	24	Batch size	24
n	$2,3,4$	Circuit power	20 dBm
K	$12,8,6$	Number of episodes	200
T_{data}	5000 instances	R_0	2 bps/Hz

randomly and uniformly. The input size of the DNN is $N \times K$ and the embedded size is 128. We generate 5000 instances for training and 1000 instances for validation data-set randomly for each episode. Each instance consists of $N \times K$ user-channel information.

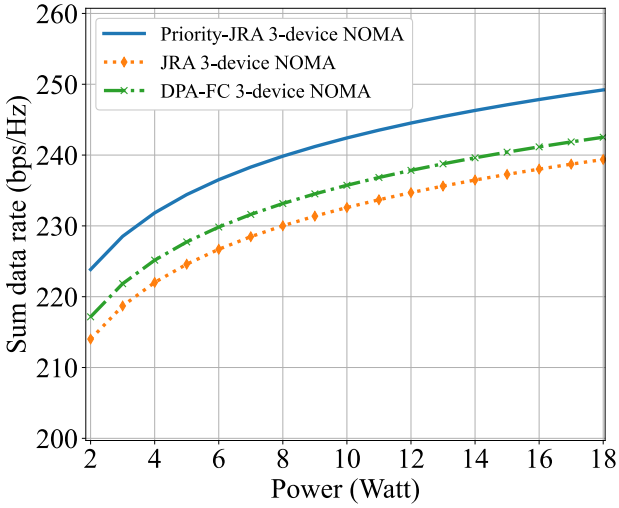

FIGURE 9. Sum-rate of 3-device NOMA system.

B. PERFORMANCE ANALYSIS

In this subsection, we compare the proposed priority-JRA with JRA and DPA-FC in terms of system sum-rate, sum-rate per channel, and energy-efficiency varying power, number of users, and location.

Fig. 9 shows the sum-rate versus the BS power comparison among priority-JRA, JRA, DPA-FC 3-device NOMA system. It is also evident from the figure that the proposed scheme outperforms the other two methods. In the JRA method, the power allocation solution is derived first, and the channels are then assigned using a matching algorithm [1]. By contrast, in the DPA-FC method, power allocation is done dynamically based on the channel response between the device and the BS while assigning fixed channels to the devices [27]. Hence, we can conclude that the priority-based channel assignment technique is more efficient than the JRA, and DPA-FC methods. From Fig. 9, we can also observe that the sum-rate is shown in bps/Hz which also reinforces the spectral efficiency of the system. Moreover, due to the converging nature of (7), the graph saturates when the BS power is extremely large.

Sum-rate for each channel comparison among priority-JRA, JRA, and DPA-FC for 3-device NOMA is shown in Fig. 10. It is evident from the figure that the proposed priority-JRA achieves the highest sum-rate in most of the channels while maintaining the proposed priority scheme. In few channels, the sum-rate is low because of the trade-off between the priority scheme and the maximum sum-rate.

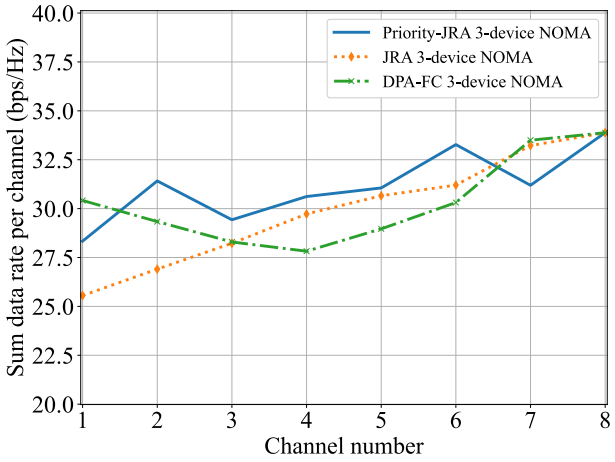

FIGURE 10. Sum-rate per channel of 3-device NOMA system where the channel number, $K = 8$.

Our main target is to fulfill the QoS requirements of the 5G services while achieving the maximum possible sum-rate.

Fig. 11 shows the sum-rate achieved by the three schemes for the 2, 3, 4-device NOMA system. For every NOMA system, the proposed priority-JRA achieves the highest sum-rate compared to the other methods. Moreover, we can also observe that the sum-rate decreases when the number of devices per channel increases. This is due to the increase in system complexity and the division of the same amount of power into more devices.

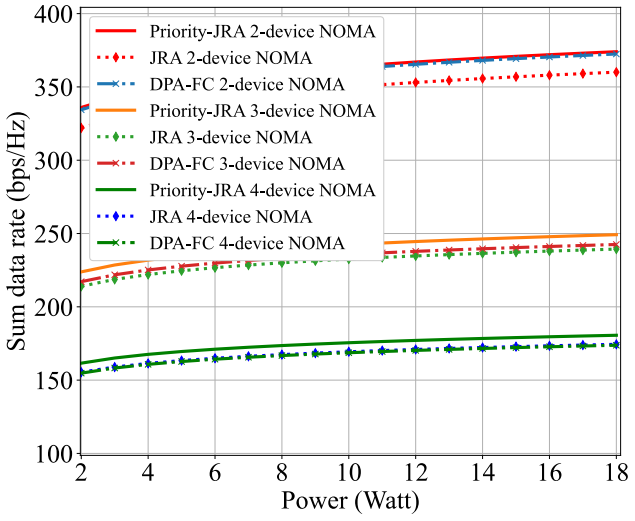

FIGURE 11. Sum-rate of 2, 3, 4-device NOMA systems.

In Fig. 12, we compare the conventional OMA system with priority-JRA along with JRA and DPA-FC NOMA systems in terms of the sum-rate with respect to power and number of users, respectively for the 3-device NOMA system. The sum-rate shown in the figure also represents the spectral efficiency of the system. It is clear that all NOMA systems outperform the traditional OMA system in terms of both the sum-rate and spectral efficiency. Moreover, we can also

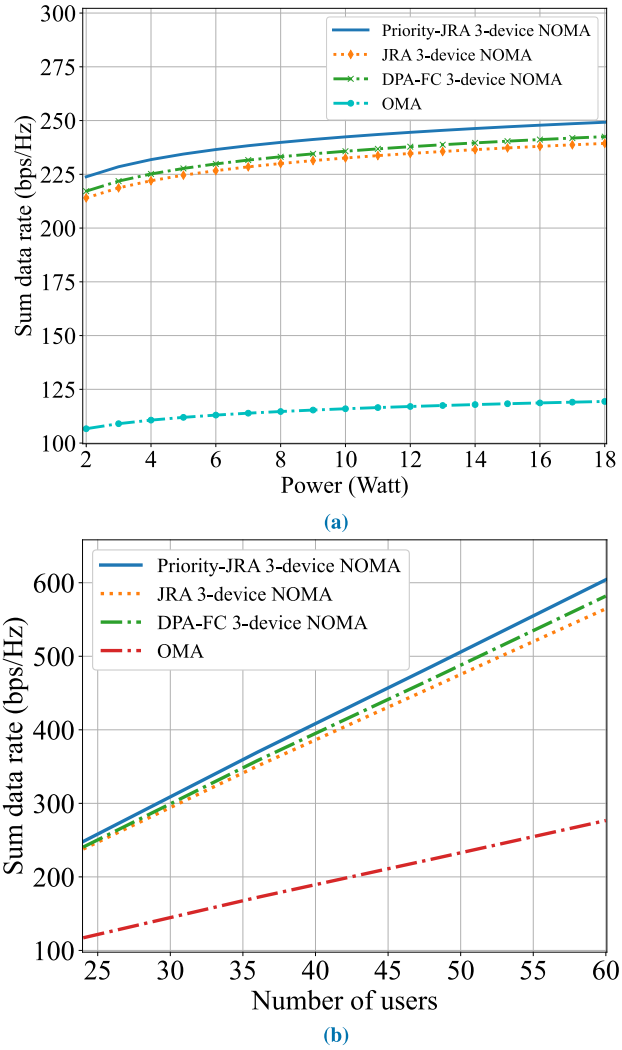

FIGURE 12. Sum-rate of 3-device NOMA system and OMA system with respect to (a) power and (b) number of users.

conclude from the Fig. 12 that the proposed priority-JRA outperforms all the other methods for any given power and number of users.

In Fig. 13, we compare the energy-efficiency of the OMA system with different methods of the NOMA system with respect to number of users and power, respectively. Energy-efficiency of a system represents the number of sent bits per joule of energy. The graph shows that the energy-efficiency decreases as the power increases because the energy efficiency is inversely proportional to power. We can conclude from the figure that the NOMA system is more energy-efficient than the conventional OMA system in any scenario. Moreover, from Fig. 13, we can also observe that the proposed priority-JRA is the most energy-efficient method for channel assignment among all for any given power and number of users. We calculated the energy efficiency graph using the BS power and circuit power for each method [1].

Moreover, Fig. 14 shows the sum-rate comparison among priority-JRA, JRA, DPA-FC 3-device NOMA system for

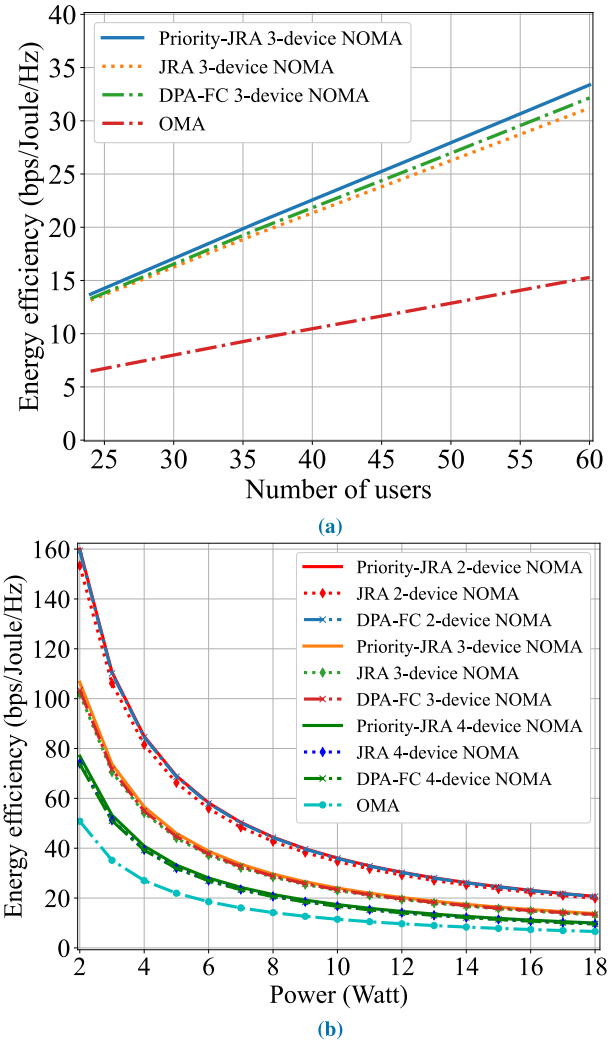

FIGURE 13. Energy-efficiency of (a) 3-device NOMA system and OMA system with respect to number of users and (b) 2, 3,and 4-device NOMA systems and OMA system with respect to power.

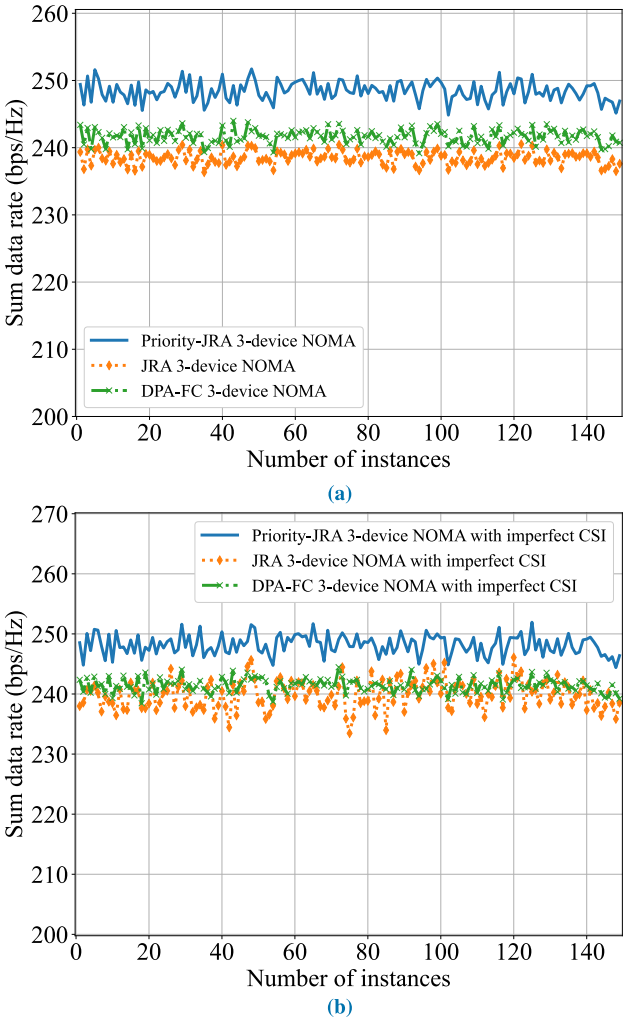

FIGURE 14. Sum-rate of 3-device NOMA systems for multiple validating instances considering (a) perfect and (b) imperfect CSI.

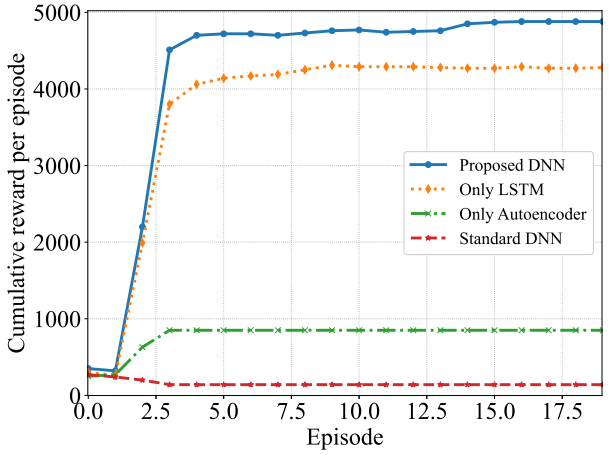

FIGURE 15. Channel assignment policy convergence for different DNN structures.

different user-data instances considering perfect and imperfect CSI. As mentioned earlier, we generate 5000 and 1000 instances consisting of $N \times K$ user-channel information per instance for training and testing the proposed priority-JRA scheme, respectively. In every instance, the positions of the users are randomly and uniformly generated within the transmission range of the BS. From Fig. 14a, it is evident that the proposed priority-JRA achieves the highest sum-rate for any given positions of the users. By contrast, we consider $\pm 30\%$ CSI error to evaluate the performance of the aforementioned systems in Fig. 14b. It is noticeable from Fig. 14b that the performance of the proposed priority-JRA remains almost unchanged compared to the JRA, DPA-FC schemes.

C. COMPLEXITY AND PARAMETER ANALYSIS

The proposed priority-JRA scheme contains a DNN network. To visualize the efficiency of the proposed DNN network, we derive and analyze the time complexity. The proposed DNN can be divided into three main elements for complexity analysis, which are an auto-encoder, an LSTM, and two linear layers as shown in Fig. 5.

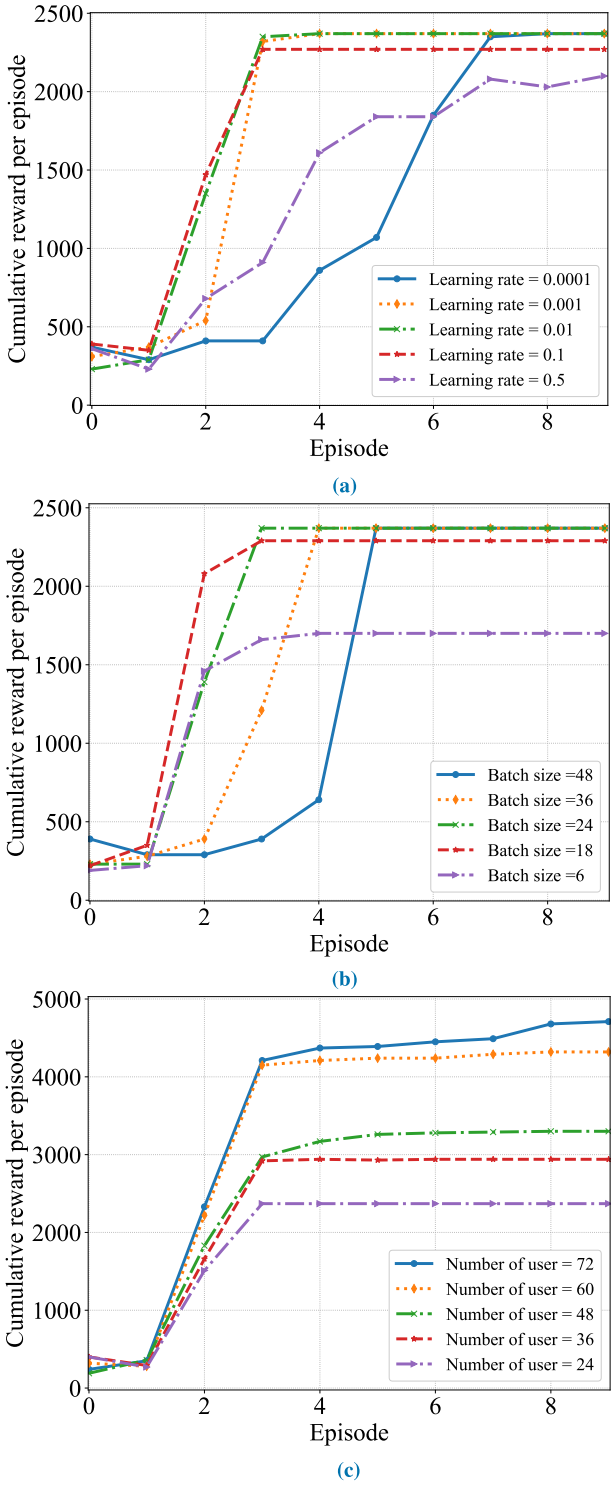

FIGURE 16. Channel assignment policy convergence for different (a) learning rate, (b) batch sizes, and (c) number of user.

The proposed DNN has an input of (NK) and two linear layers of size $d_e = 128$. The time complexity can be written as $O(2Id_e^2(NK))$, where I refers to the kernel size. The auto-encoder has one code layer and two identical encoder and decoder layers. According to [44], the time complexity of the auto-encoder can be written as

$$\begin{aligned} &O(2Id_e^2(1+\frac{1}{2}+\frac{1}{4}+\frac{1}{32})(NK)) \\ &= O(\frac{55}{16}Id_e^2(NK)) \\ &\simeq O(3Id_e^2(NK)) \end{aligned} \tag{15}$$

For the LSTM the time complexity can be calculated as $O(I)$. Therefore, the overall time complexity of the proposed DNN can be written as

$$\begin{aligned} &O(3Id_e^2(NK)) + O(2Id_e^2(NK)) + O(I) \\ &= O(5Id_e^2(NK)) + O(I) \end{aligned} \tag{16}$$

By contrast, for the JRA scheme, the time complexity can be calculated as $O((\frac{I^2-I}{2})\binom{N}{n}^2)$, which includes all $\binom{N}{n}$ combinations for each channel k. Therefore, the complexity of the priority-JRA is much lower. However, DPA-FC scheme has the lowest complexity and it does not outperform the priority-JRA scheme.

To justify our proposed DNN structure, we compare it with multiple DNN structures such as standard fully-connected DNN, only LSTM, and only autoencoder in Fig. 15 for 72-devices and at a learning rate 0.01 and batch size of 24. It is evident from Fig. 15 that the proposed DNN structure achieves maximum cumulative reward and converges faster among all. Furthermore, Fig. 16a shows the effect of different learning rates on the proposed DNN for 24-devices and a batch size of 24. As shown in Fig. 16a, the proposed DNN cannot learn the optimal channel assignment policy for learning rates of 0.5, 0.1, and 0.001. However, for learning rates 0.01 and 0.001, the proposed DNN reached the optimal solution quickly in the same episode. Therefore, we can use any one of them. Fig. 16b shows the effect of different batch sizes on the proposed DNN for 24-devices and a learning rate of 0.01. As shown in Fig. 16b, the batch size should be greater than or equal to 24 to achieve optimality. However, a larger batch size refers to more room for exploration and slow convergence. Lastly, Fig. 16c represents the convergence of the proposed DNN for different number of users at a learning rate of 0.01 and batch size 24. The converging graphs of Fig. 16c signify the high scalability and stability of the proposed DNN for increasing number of users under the BS. Finally, we can ensure from the analysis that the proposed scheme can achieve a near-optimal performance with low complexity and high efficiency.

VI. CONCLUSION

In this paper, we propose a priority-based resource allocation scheme with deep Q-learning to fulfill the QoS requirements of the 5G services, such as URLLC, eMBB, and mMTC services, while maximizing the system performance and fairness of the multi-carrier NOMA system. We consider SISO-NOMA system architecture to derived the power allocation and the channel assignment problems into optimization problems. To resolve these problems,

we first formulated a near-optimal power allocation solution using Lagrange multipliers under KKT optimality conditions while incorporating different constraints of the NOMA system. Then with the derived power allocation solution, we formulated priority-based channel assignment with deep Q-learning utilizing an autoencoder and LSTM in the DNN model. After that we compared the proposed scheme with JRA and DPA-FC schemes and proved that the proposed priority-JRA performs better than other schemes under different conditions. We plan to extend our proposed solutions considering MIMO-NOMA with beamforming in future works, where the BS with multiple antennas will assign each channel to multiple devices using beamforming utilizing a machine learning algorithm. Finally, we can conclude that our proposed priority-JRA method is less complex than other optimal exhaustive search based solutions while achieving near-optimal solution.

REFERENCES

[1] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[2] K. Wang, Y. Zhou, Z. Liu, Z. Shao, X. Luo, and Y. Yang, "Online task scheduling and resource allocation for intelligent NOMA-based industrial Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 803–815, May 2020.

[3] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[4] S. Hu, B. Yu, C. Qian, Y. Xiao, Q. Xiong, C. Sun, and Y. Gao, "Nonorthogonal interleave-grid multiple access scheme for industrial Internet of Things in 5G network," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5436–5446, Dec. 2018.

[5] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[6] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[7] A. Shahini and N. Ansari, "NOMA aided narrowband IoT for machine type communications with user clustering," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7183–7191, Aug. 2019.

[8] X. Liu, X. B. Zhai, W. Lu, and C. Wu, "QoS-guarantee resource allocation for multibeam satellite industrial Internet of Things with NOMA," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2052–2061, Mar. 2021.

[9] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, and H. Alves, "Network slicing for URLLC and eMBB with max-matching diversity channel allocation," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 658–661, Mar. 2020.

[10] G. Gui, H. Sari, and E. Biglieri, "A new definition of fairness for non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1267–1271, Jul. 2019.

[11] X. Yan, K. An, T. Liang, G. Zheng, Z. Ding, S. Chatzinotas, and Y. Liu, "The application of power-domain non-orthogonal multiple access in satellite communication networks," *IEEE Access*, vol. 7, pp. 63531–63539, 2019.

[12] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[13] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous IoT with imperfect SIC," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2885–2894, Apr. 2019.

[14] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[15] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[16] A. Celik, M.-C. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M.-S. Alouini, "Distributed user clustering and resource allocation for imperfect NOMA in heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7211–7227, Oct. 2019.

[17] S. Rezwan, S. Shin, and W. Choi, "Efficient user clustering and reinforcement learning based power allocation for NOMA systems," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju, South Korea, Oct. 2020, pp. 143–147.

[18] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[19] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[20] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 583–596, Feb. 2014.

[21] S. Zhang, B. Di, L. Song, and Y. Li, "Radio resource allocation for non-orthogonal multiple access (NOMA) relay network using matching game," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[22] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[23] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[24] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[25] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[26] J. G. Andrews and T. H. Meng, "Optimum power control for successive interference cancellation with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 375–383, Mar. 2003.

[27] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[28] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[29] X. Shao, C. Yang, D. Chen, N. Zhao, and F. R. Yu, "Dynamic IoT device clustering and energy management with hybrid NOMA systems," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4622–4630, Oct. 2018.

[30] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1026–1031.

[31] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–6.

[32] Z. Ning, X. Wang, J. J. P. C. Rodrigues, and F. Xia, "Joint computation offloading, power allocation, and channel assignment for 5G-enabled traffic management systems," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3058–3067, May 2019.

[33] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[34] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.
[35] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.
[36] Z. Wei, D. W. K. Ng, J. Yuan, and H.-M. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.
[37] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 3rd ed. Hoboken, NJ, USA: Wiley, 2008.
[38] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, 2014, pp. 1–14.
[39] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, Jul. 2019.
[40] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 1999, pp. 1057–1063.
[41] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 8792–8802.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3nd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980
[43] S. Zhang and R. S. Sutton, "A deeper look at experience replay," *Comput. Res. Repository (CoRR)*, vol. abs/1712.01275, 2017. [Online]. Available: http://arxiv.org/abs/1712.01275
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Comput. Res. Repository (CoRR)*, vol. abs/1706.03762, 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762

SIFAT REZWAN (Student Member, IEEE) received the B.S. degree from the Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh, in 2018. He is currently pursuing the M.Sc. degree with the Smart Networking Laboratory, Chosun University, South Korea, under the supervision of Prof. W. Choi. From 2019 to 2020, he was with Grameenphone Ltd., where he was involved in transmission network operations and automation and received the Top Performer Award in 2020. His research interests include multiple access for 5G and beyond 5G networks, deep learning-based resource optimization, and heterogeneous wireless networks.

WOOYEOL CHOI (Member, IEEE) received the B.S. degree from the Department of Computer Science and Engineering, Pusan National University, Busan, South Korea, in 2008, and the M.S. and Ph.D. degrees from the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2010 and 2015, respectively. He was a Senior Research Scientist with the Korea Institute of Ocean Science and Technology (KIOST), Ansan, South Korea, in 2015 to 2017, and a Senior Researcher, Korea Aerospace Research Institute (KARI), Daejeon, South Korea, in 2017 to 2018. He is currently an Assistant Professor with the Department of Computer Engineering, Chosun University, Gwangju. His research interests include cross-layer protocol design, deep learning-based resource optimization, and experiment-driven evaluation of wireless networks.