DRL-Assisted Dynamic Subconnected Hybrid Precoding for Multi-Layer THz mMIMO-NOMA System

Md. Shahjalal, *Member, IEEE,* Md. Habibur Rahman, *Student Member, IEEE,* Md Morshed Alam, *Graduate Student Member, IEEE,* Mostafa Zaman Chowdhury, *Senior Member, IEEE,* and Yeong Min Jang, *Member, IEEE*

***Abstract*—Massive multiple-input multiple-output (mMIMO) techniques can be combined with the non-orthogonal multiple access (NOMA) scheme in terahertz (THz) communication to achieve multiplexing gains and satisfy the ultra-high capacity and massive connectivity requirements. However, the development of a near-optimal solution for energy and spectral efficiency problems in a dynamic wireless cellular environment remains challenging. In this paper, a cooperative THz mMIMO-NOMA enabled base station is established to optimize the power consumption and maximize the spectral efficiency. A multi-layer mMIMO antenna architecture is used to perform dynamic sub-connected hybrid precoding in each layer. The fuzzy c-means clustering algorithm is used to group densely located users into clusters to efficiently use the power coefficients. To optimize the power distribution constraints and coordination of the hybrid precoding structure, a multi-agent deep reinforcement learning algorithm is developed, which operates in a distributive manner. Each base station layer involves an agent that trains a deep Q-network, and optimal actions are executed by sharing exchangeable network parameters among layers. The simulation results indicate that the proposed scheme is able to learn the trade-off between maximization of the energy efficiency and overall system capacity.**

***Index Terms*—Deep reinforcement learning (DRL), hybrid precoding, massive multiple-input multiple-output (mMIMO), non-orthogonal multiple access (NOMA), Terahertz (THz).**

I. Introduction

ULTRA-massive interconnectivity with rapid growth of wireless data rate requirements is one of the major challenges in future wireless networks. As a result, new spectra with promising features such as ultra-broad bandwidth is necessary. In addition to the millimeter-wave (24–300 GHz) band, attention must be focused on the terahertz (THz) band, which is associated with much higher bandwidth. In accordance with the recommendation of the International Telecommunication Union, the frequencies between 275 GHz and 3 THz are reserved for sixth-generation THz wireless systems. The THz band can support a transmission rate of tens of Gbps while enabling ultra-low-latency communications with improved directionality, confidentiality, and strong anti-interference ability [1]. Therefore, high data rate short-range broadband THz wireless communication is feasible. In a THz cellular network, a data rate of 18.3 Gbps was achieved with 99.999% reliability for a Matern hardcore point process-based virtual reality users [2]. However, the THz band, which typically has a multi-GHz bandwidth, incurs an extremely high propagation loss that decreases the communication range [3]. The multiple-input multiple-output (MIMO) technology can provide a high beamforming gain to compensate the path loss. A THz base station (BS) must be implement with large antenna arrays (i.e. more than 500 antenna elements) referred to massive MIMO (mMIMO). Large antenna elements can be integrated in a physically limited space because of the sub-millimeter wavelength. Thus, a THz mMIMO BS can provide multiplexing gain by supporting multiple data streams, thereby enhancing the spectral efficiency. The system performance can be enhanced using non-orthogonal multiple access (NOMA) technologies, which can support an increased number of users and channel capacity by simultaneously using the same time and frequency domains [4]. The successive interference cancellation (SIC) technique inherent to NOMA can help eliminate the interference from strong users, thereby increasing the system throughput [5], [6]. In case of mMIMO antenna system, the transmitting beams are restricted by the number of transmitting antennas. An intra-beam superposition coding can be applied supporting multiple users enabling SIC when mMIMO and NOMA are used cooperatively [6]. Therefore, the cooperative use of NOMA and mMIMO in THz communications has been recommended to enhance the spectral efficiency, energy efficiency (EE), and connectivity at a large scale.

In particular, certain researchers [7] presented analytical and numerical solutions for the power allocation problem in a cooperative THz MIMO-NOMA system to maximize the minimum achievable rate. Based on cooperative simultaneous wireless information and power transfer, a THz MIMO-NOMA system was proposed in [8] to enhance the wireless connectivity, resource management, scalability, reliability, and

M. Shahjalal is with the Department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh, Dhaka, 1207, Bangladesh e-mail: (md.shahjalal@ulab.edu.bd).

M. H. Rahman was with the Department of Electronics Engineering, Kookmin University, Seoul 02707, South Korea, and is currently with the Department of Electrical and Computer Engineering, Virginia Tech, VA, USA e-mail: (rahman.habibur@ieee.org).

M. M. Alam is with the School of Electrical and Data Engineering, University of Technology Sydney NSW, Australia e-mail: (mmorshed@ieee.org).

M. Z. Chowdhury is with the Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna, 9203 Bangladesh email: (mzceee@ieee.org).

Y. M. Jang is with the Department of Electronics Engineering, Kookmin University, Seoul, 02707 South Korea e-mail: (yjang@kookmin.ac.kr).

Manuscript received XXX, XX, 2015; revised XXX, XX, 2015.

user fairness. To enhance the channel conditions, a spatial tuning technique was developed in [9] for a THz ultra-massive MIMO-NOMA configuration. Moreover, in the THz MIMO BS, to avoid lots of hardware complexity, hybrid precoding (HP) technique is being researched recently and can achieve satisfactory spectral efficiency compared to traditional digital precoding. In HP frameworks, the signal processing is divided into low-dimensional digital baseband precoding and high-dimensional analog radio frequency (RF) precoding. The networking performance and power consumption depend on the connection dimensions of the RF chain and mMIMO transmitting antennas. HP architectures can typically be divided into 1) fully connected HP (FCHP) and 2) sub-connected HP (SCHP). The critical difference is that in the FCHP, each RF chain is connected to every antenna, whereas in the SCHP, an array of antennas is connected to only one RF chain. Therefore, FCHP achieves higher spectral efficiency and consumes more power than that of SCHP. An appropriate precoder can be used to provide highly directional beams to mitigate multiuser interference [10]. However, none of these techniques can adaptively and dynamically control the circuit connections, which is necessary for dynamically varying the network channel conditions.

Furthermore, to perform resource optimization in THz mMIMO-NOMA networks, developing a suitable user clustering technique is a challenging research area. Only a few studies on MIMO-NOMA systems have considered the user pairing when clustering the users for a limited number of active users [11]-[13]. A joint user clustering and power allocation algorithm was proposed in [14], where the main target was to elevate the sum-rate by selecting two best users in a cluster. However, this system introduced polynomial computational complexities in user clustering. An interference aware graph based clustering approach was proposed in [15] for two types of users: cellular and device-to-device. A channel graph was constructed by the BS for cellular users by measuring the channel correlation between two users. Although the user clustering for power optimization and resource management in low-frequency band networks has been extensively studied, the corresponding strategies for THz mMIMO-NOMA networks are limited.

The maximization of the energy or spectral efficiency of a cellular network is typically a nondeterministic polynomial (NP)-hard and nonconvex problem, and thus, an optimal solution is difficult to obtain. Although several sub-optimal solutions are available to address these problems, these solutions are based on a centralized approach that involves a non-negligible delay and are thus not suitable for a dynamic wireless environment, which is commonly encountered in practice. Recently, deep reinforcement learning (DRL) has been used to develop promising solutions to address various NP-hard problems in the communications and networking domains. DRL represents a combination of deep learning and reinforcement learning techniques, aimed at performing several practical decision-making tasks with a large state-action space. Certain researchers [16] proposed a reconfigurable THz MIMO-NOMA framework to intelligently coordinate beamforming between access points and reconfigurable intelligent surfaces

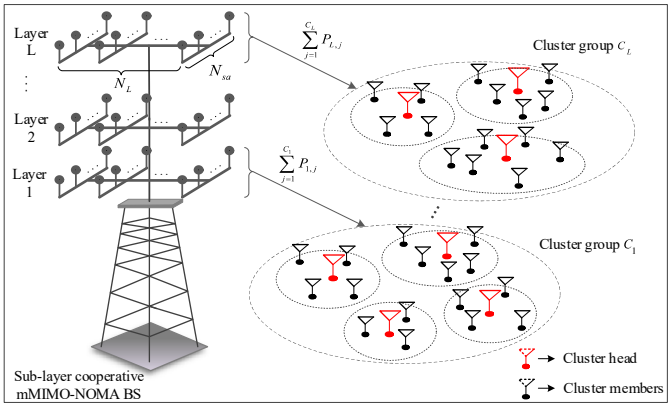

Fig. 1. Proposed multi-layer mMIMO-NOMA downlink network architecture.

(RISs) by using a multi-agent DRL algorithm. Moreover [17], a dynamic downlink-beamforming coordination system consisting of multiple BSs with a single user was proposed using a distributed DRL algorithm to maximize the achievable rate. A hybrid beamforming scheme [18] was developed for multihop RIS-assisted THz communication networks. Specifically, a DRL-based algorithm was developed for the multihop environment to enhance the coverage range by solving a NP-hard beamforming problem. Furthermore, a dynamic power allocation scheme was proposed [19] based on a distributively executed DRL technique for practical wireless scenarios.

This paper proposes a multi-agent DRL-based THz mMIMO-NOMA system, in which the mMIMO antennas are grouped into multiple layers, as shown in Fig. 1. For simplicity, the number of antennas per layer is kept equal. The distinct layers of the antenna are subject to support categorized users based on the requirements of their channel gain. Therefore, the antenna layers are incorporated with a dynamic SCHP (DSCHP) scheme to obtain an adaptive antenna configuration (see Fig. 2) that can satisfy the dynamic wireless channel requirements. The conventional FCHP scheme in mMIMO provides full array gain compared to SCHP while consuming more energy. Hence, the proposed DSCHP structure is advantageous in reducing power consumption while supporting users with moderate channel gain. Achieving this goal, a multi-agent DRL framework is developed where each layer corresponds to an agent. The agents determine the beamforming parameters by playing the best action with a fair reward by exchanging information between the layers. Fuzzy c-means (FCM) clustering algorithm is adopted for partitioning the users supported by NOMA-enabled beamforming. The layers can hold multiple clusters by subdividing the antennas into subarrays. The simulation results described in Section V show the proposed system achieves better performance in terms of energy efficiency while meeting the required channel gain. The key contributions of this manuscript are elaborated as follows

- A multi-layer mMIMO antenna system for THz communications is proposed which incorporates the DSCHP scheme in each antenna layer. Therefore, the antennas in each layer are divided into multiple subarrays, with each subarray responsible for one user cluster.

- The users of distinct channel responses are grouped into multiple clusters using Fuzzy c-means (FCM) clustering. The results show the clustering performance based on the proposed system has a higher fuzzy partitioning coefficient (FPC).
- A multi-agent DRL framework is presented where a deep Q-network (DQN), known as the training DQN, is centrally trained using the shared experiences stored in the replay memory of all the agents (i.e., layer) to decrease the computational overhead and memory usage. Each agent also involves another DQN named the target DQN, whose parameters are updated periodically and executed distributively in each BS layer.
- The proposed cellular environment was simulated in the Python environment. Based on the DRL framework, a near-optimal solution is achieved for the system as the spectral efficiency converges to a higher value compared to other baseline schemes. Moreover, the EE performance shows a better result compared to a scheme with full array gain.

The remaining paper is organized as follows. Section II describes the system model with the proposed DSCHP structure and channel properties. Section III describes the FCM-based clustering technique for the DSCHP system. Section IV describes the proposed DRL framework for the DSCHP-based THz mMIMO-NOMA system. Section V presents the simulation results and their discussion. Finally, the conclusion is drawn in Section VI.

II. System Model

We consider a multi-layer architecture of a single-cell downlink mMIMO-NOMA enabled BS for 6G THz communications. As depicted in Fig. 1, the BS has N antennas divided into L layers, with each layer containing $N_l = N/L$ antennas. Multiple layers are introduced to exploit the different combinations of HP schemes. Therefore, each layer consists of subarrays of antennas, with each subarray containing N_{sa} antennas. The HP scheme is applied to the mMIMO transmitting subarray antennas in each layer, where $N_{rf,l}$ is the number of RF chains connected in lth layer. NOMA is incorporated to satisfy the massive connectivity requirements. NOMA is implemented with each transmitted beam to support multiple users by using the same time and frequency resources. In this mMIMO-NOMA system, all the single antenna users are served in the form of clusters supported by the NOMA-enabled transmitted beams. The users are divided into C clusters, with each beam dedicated to one cluster. A group of clusters, $C_l = \frac{N_l}{N_{sa}}$ is supported by the lth layer, and $K_{l,c}$ is the number of users in the cth cluster of the C_lth group, with $c = 1, ..., C$ and $l = 1, ..., L$. Without loss of generality, the number of clusters supported by the lth layer is set equal to the number of RF chains in that layer to ensure the multiplexing gain.

A. DSCHP Modelling

In the FCHP system, each RF chain is connected to every subarray antenna through finite-resolution phase shifters to achieve a full array gain by all the RF chains. In contrast,

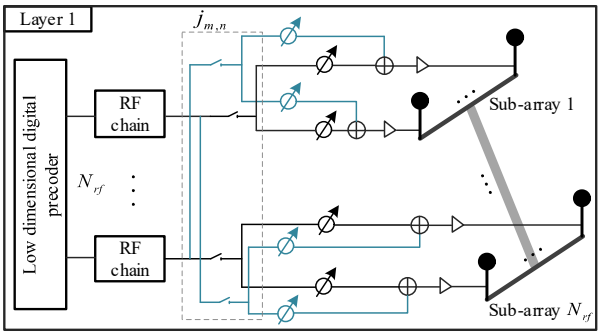

Fig. 2. Connection diagram of the lth layer DSCHP structure.

in the SCHP system, a finite set of transmitting antennas are connected to each RF chain in $N_{rf,l}$ by fewer phase shifters. Therefore, a low-complexity circuit arrangement is required for the SCHP system, and it consumes less power than the FCHP system. However, the SCHP system exhibits inferior network performances, such as the spectral efficiency, compared to the FCHP. Considering these limitations, we establish a multi-layer architecture in which each layer incorporates the DSCHP scheme following the received signal strength of the user. Fig. 2 shows the DSCHP structure with a junction network introduced between the RF chain and subarray antennas in the lth layer. The objective is to decrease the power consumption by using fewer connected RF chains and phase shifters while satisfying the user requirements.

The signal received at time t for the ith user in the cth cluster of the lth layer can be represented as

$$y_{l,c}^{(i)}(t) = h_{l,c}^{H(i)} A_l \sum_{c=1}^{C} D_{l,c} X_{l,c}(t) + \eta_{l,c}^{(i)}, \tag{1}$$

where $X_{l,c}(t)$ represents the superposed signal for all $K_{l,c}$ users in the cth cluster of the lth layer, defined as

$$X_{l,c}(t) = \sum_{i=1}^{K_{l,c}} \sqrt{\rho_{l,c}^{(i)} P_{l,c}} x_{l,c}^{(i)}(t), \tag{2}$$

where $x_{l,c}^{(i)}$ represents the signal of the ith user in the cth cluster of the lth layer, $[x_{l,c}^{(1)}, x_{l,c}^{(2)}, ..., x_{l,c}^{(K_{l,c})}]$, at time t. $P_{l,c}$ represents the set of transmit powers at each subarray of the lth layer, and $\rho_{l,c}^{(i)}$ denotes the power coefficient of the ith user in the cth cluster of the lth layer, $[\rho_{l,c}^{(1)}, \rho_{l,c}^{(2)}, ..., \rho_{l,c}^{(K_{l,c})}]$, subject to the following conditions:

$$\sum_{c=1}^{C} P_{l,c} \leq P_l, \quad for\ l = 1, ..., L, \tag{3}$$

and

$$\sum_{i=1}^{K_{l,c}} \rho_{l,c}^{(i)} = 1, \quad for\ l = 1, ..., L\ and\ c = 1, ..., C, \tag{4}$$

where P_l is the transmit power of the lth layer.

In (1), $D_{l,c}$ represents the digital precoding vector of size $N_{rf,l} \times 1$ for the lth layer, A_l represents the analog precoding matrix of size $N_l \times N_{rf,l}$ such that $\|A_l D_{l,c}\|_2 = 1$, and $\eta_{l,c}$

is a complex Gaussian thermal noise vector following the independent and identically distributed $C\mathcal{N}(0,\, \sigma_b^2)$. For the FCHP structure, A_l can be represented as

$$A_l^{(f)} = \left[a_{l,1}^{(f)},\, a_{l,2}^{(f)},\, ...,\, a_{l,N_{rf,l}}^{(f)}\right], \tag{5}$$

where the matrix elements $a_{l,n}^{(f)} \in \mathbb{R}^{N_l \times 1}$ for $n = 1,\, 2,\, ...,\, N_{rf,l}$ exhibit the same amplitude $N_l^{-\frac{1}{2}}$ with different phases (phase shifting is performed after the digital precoding). For the SCHP structure, each RF chain is connected to a smaller number of antenna subarrays, N_{sa}. Generally, N_{sa} is an integer and calculated as $N_{sa} = \frac{N_l}{N_{rf,l}}$. Therefore, A_l for the sub-connected structure is

$$A_l^{(s)} = \begin{bmatrix} a_{l,1}^{(s)} & 0 & \cdots & 0 \\ 0 & a_{l,2}^{(s)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{l,N_{rf,l}}^{(s)} \end{bmatrix}, \tag{6}$$

where the matrix elements $a_{l,n}^{(s)} \in \mathbb{R}^{N_{sa} \times 1}$ for $n = 1,\, 2,\, ...,\, N_{rf,l}$ exhibit the same amplitude $N_{sa}^{-\frac{1}{2}}$.

A junction network $J_l \in \mathbb{R}^{N_{rf,l} \times N_{rf,l} \times N_{rf,l}}$ is designed to achieve the dynamic switching of each RF chain to the phase shifters in each lth layer.

$$J_l = \begin{bmatrix} j_{1,1} & j_{1,2} & \cdots & j_{1,N_{rf,l}} \\ j_{2,1} & j_{2,2} & & j_{2,N_{rf,l}} \\ \vdots & & \ddots & \vdots \\ j_{N_{rf,l},1} & j_{N_{rf,l},2} & \cdots & j_{N_{rf,l},N_{rf,l}} \end{bmatrix}, \tag{7}$$

where $j_{m,n} \in \mathbb{R}^{N_{rf,l} \times 1}$ is a column vector, where $m = 1,\, 2,\, ...,\, N_{rf,l}$. Here, $j_{m,n}$ is either an all-one vector or a zero vector if it connects or disconnects the nth RF chain to the mth subarray antennas, respectively.

B. Channel Properties

After being precoded by the dedicated baseband hybrid precoder, the superposed signal is transmitted by the antenna through the high-frequency THz wireless channel to the users located in the cth cluster. This channel matrix has dimensions $L \times C \times N_l \times K_{l,c}$ and can be defined as

$$H = [H_1,\, H_2, \cdots, H_l, \cdots, H_L], \tag{8}$$

where

$$H_l = \begin{bmatrix} h_{l,1}^{(1)} & h_{l,1}^{(2)} & \cdots & h_{l,1}^{(K_{l,c})} \\ h_{l,2}^{(1)} & h_{l,2}^{(2)} & & h_{l,2}^{(K_{l,c})} \\ \vdots & & \ddots & \vdots \\ h_{l,c}^{(1)} & h_{l,c}^{(2)} & \cdots & h_{l,c}^{(K_{l,c})} \end{bmatrix}. \tag{9}$$

Therefore, the $N_l \times 1$ channel vector $h_{l,c}^{(i)}$ of the ith user in the cth cluster of the lth layer is formulated as

$$h_{l,c}^{(i)}(f,d) = \sqrt{N_l \Omega_{l,c}(f,d)} G_r G_t \alpha(\phi_{l,c}^{(i)}, \theta_{l,c}^{(i)}), \tag{10}$$

where $\Omega_{l,c}$ denotes the line-of-sight (LoS) path loss that depends on the THz frequency f and distance d between the BS and user; G_r and G_t are the receive and transmit antenna gains, respectively; and α is the steering vector of the $N_l \times 1$ layer. $\phi_{l,c}^{(i)}$ and $\theta_{l,c}^{(i)}$ are the azimuth and elevation angles of departure for the ith user in the cth cluster of the lth layer, respectively. The complex gain of the LoS path loss term, $\Omega_{l,c}$, can be defined as

$$\Omega_{l,c}(f,d)[dB] = 20\log_{10}\left(\frac{V}{4\pi f d}\right) + 10d\left(\kappa(f) + j\frac{4\pi f}{V}\right)\log_{10} e, \tag{11}$$

where $\kappa(f)$ represents the frequency-dependent molecular absorption coefficient, and V is the speed of light. $\kappa(f)$ is computed as a sum of the absorption contributions from the isotopes of gases in a medium.

Because the mMIMO-NOMA BS is subdivided into multiple layers, the interlayer interference must be considered in addition to the intercluster interference. Therefore, the received signal in (1) can be modeled by considering the desired and interfering signals as

$$y_{l,c}^{(i)}(t) = h_{l,c}^{H(i)} A_l D_{l,c} \sqrt{\rho_{l,c}^{(i)} P_{l,c}} x_{l,c}^{(i)}(t) + I_{IC} + I_{ILC} + \eta_{l,c}^{(i)}, \tag{12}$$

where I_{IC} and I_{ILC} represent the intra-cluster and inter-layer interference, respectively, defined as

$$I_{IC} = h_{l,c}^{H(i)} A_l D_{l,c} \sum_{j=1}^{K_{l,c}-1} \sqrt{\rho_{l,c}^{(j)} P_{l,c}} x_{l,c}^{(j)}(t) \tag{13}$$

and

$$I_{ILC} = h_{l,c}^{H(i)} A_l \sum_{k=1}^{L} \sum_{j=1, j\neq c}^{C-1} D_{k,j} \sqrt{P_{k,c}} x_{k,c}^{(j)}(t). \tag{14}$$

According to (12), the signal-to-interference-plus-noise ratio (SINR) for the ith user in the cth cluster of the lth layer can be expressed as

$$\zeta_{l,c}^{(i)} = \frac{\rho_{l,c}^{(i)} P_{l,c} \left\| h_{l,c}^{H(i)} A_l D_{l,c} \right\|_2^2}{\lambda_{l,c}^{(i)}}, \tag{15}$$

where

$$\lambda_{l,c}^{(i)} = \sum_{k=1}^{L} \sum_{j=1, j\neq c}^{C-1} P_{l,j} \left\| h_{l,c}^{H(i)} A_l D_{l,j} \right\|_2^2 + \left\| h_{l,c}^{H(i)} A_l D_{l,c} \right\|_2^2 \sum_{j=1}^{K_{l,c}-1} \rho_{l,c}^{(j)} P_{l,c} + \eta_{l,c}^{(i)}. \tag{16}$$

The achievable rate for the ith user in the cth cluster of the lth layer can be represented using the THz channel capacity model [1]:

$$\Gamma_{l,c}^{(i)} = \frac{\xi}{N} \log_2(1 + \zeta_{l,c}^{(i)}), \tag{17}$$

where ξ is the bandwidth employed in the lth layer of the THz BS. The achievable sum rate at the lth layer can be expressed as

$$\Gamma_l^{sum} = \sum_{c=1}^{C} \sum_{i=1}^{K_{l,c}} \Gamma_{l,c}^{(i)}. \tag{18}$$

C. Problem Definition

The proposed system aims to reduce the power consumption by dynamically switching the connections of the RF chain. This section describes the power consumption profile and formulation of the utility function for the system. The total power consumption, P_{con}, at the mMIMO-NOMA BS is the sum of the power consumed by the circuitry and transmitted signals:

$$P_{con} = P_T + \sum_{l=1}^{L} \left(N_{rf,l} P_{rf} + \left(P_{sw} + N_{sa} P_{ph} \right) N_{sw,l} \right) + \left(P_{amp} + P_{com} \right) N + P_{bb}, \tag{19}$$

where $P_T = \sum_{l=1}^{L} \sum_{c=1}^{C} P_{l,c}$ is the total transmitted power; P_{rf}, P_{sw}, P_{ph}, P_{amp}, and P_{com} is the power consumption for each RF chain, switch, phase shifter, power amplifier, and combiner, respectively; P_{bb} is the baseband power consumption, and $N_{sw,l}$ is the number of closed switches in lth layer.

The spectral efficiency is defined as the achievable sum rate, as indicated in (18). The EE is formulated as the ratio of the spectral efficiency to the total power consumption [1]. Therefore, the system utility function can be expressed as

$$\max_{P_{l,c}, \rho_{l,c}} \frac{\Gamma_{sum}}{P_{con}} \tag{20a}$$

$$s.t.\ C1: \sum_{l=1}^{L} P_l \leq P_T, \tag{20b}$$

$$C2: \sum_{i=1}^{K_{l,c}} \rho_{l,c}^{(i)} = 1, \quad \forall\, c, l, \tag{20c}$$

$$C3: \sum_{c=1}^{C_l} \Gamma_c^{sum} \geq \Gamma_{th,l}, \quad \forall\, l, \tag{20d}$$

$$C4:\ P_{l,c} \geq 0,\ \rho_{l,c} \in [0,\ 1], \quad \forall\, c, l, \tag{20e}$$

where Γ_c^{sum} is the achievable sum rate of the users in the cth cluster, and $\Gamma_{th,l}$ is the minimum required sum rate for the lth layer. Constraint C1 ensures that the maximum total transmit power is P_T; C2 ensures that the sum of the power coefficients of all users in a cluster is 1; C3 is the sum rate constraint that ensures that the achievable sum rate is greater than $\Gamma_{th,l}$; and C4 is the inherent constraint of $P_{l,c}$ and $\rho_{l,c}$, which ensures that the power allocated to each cluster is positive, and the power coefficient is a positive fraction ranging between 0 and 1.

Algorithm 1 FCM clustering algorithm in lth layer

Input: Data point matrix Q
Output: S and v_c

1: Initialize $r \geq 1$ and $C \geq 2$;
2: Partitioning of S with Q;
3: Compute v_c and $\mu_{i,c}$;
4: **if** $J_r(S,\, Q) \leq \epsilon$ **then**
5: take final output;
6: **else**
7: update v_c and $\mu_{i,c}$
8: **end if**

III. FCM Clustering for the Multi-Layer mMIMO System

The FCM is an unsupervised clustering algorithm for feature analysis, aimed at classifying the users into several clusters. First, the algorithm initializes the number of clusters and fuzzy exponent. Next, a membership function is assigned to each user to determine a fractional relation with a cluster. The cluster head is computed repeatedly by updating the membership coefficient of each data point to minimize the objective function at a certain threshold [20]. Details of the FCM clustering algorithm are presented as Algorithm 2.

The initial operation of the FCM algorithm is to design fuzzy partitioning (S, Q), where S is generally known as fuzzy matrix $[\mu_{11}, ..., \mu_{K_{l,c},C}]$. Here μ_{ic} is termed as the membership function of the cth cluster satisfying $0 \leq \mu_{ic} \leq 1$, $i = 1, ..., K_{l,c}$, and $c = 1, ..., C$. And the Q is a data point matrix of elements q_i where $i = 1, ..., K_{l,c}$. The objective of the FCM algorithm is to identify the fuzzy matrix and a set of mean of the ith points in the cth cluster i.e. the centroid. We can define the objective function as

$$\min_{(S,\, Q)} \left(J_r(S,\, Q) = \sum_{c=1}^{C} \sum_{i=1}^{K_{l,c}} \mu_{i,c}^r \|q_i - v_c\|^2 \right), \tag{21}$$

where $r \geq 1$ denotes the fuzzy exponent, v_c represents the centroid of the cth cluster. With a view to converging the algorithm up to the minimum error ϵ, the membership function and the centroids is updated in every iteration upon each of the clusters. Those are given as

$$\mu_{i,c} = \frac{1}{\sum_{k=1,\, k \neq c}^{C} \left(\frac{d_{i,c}}{d_{i,k}} \right)^{\frac{2}{r-1}}}, \quad for\ i = 1, ..., K_{l,c}, \tag{22}$$

and

$$v_c = \frac{\sum_{i=1}^{K_{l,c}} q_i \mu_{i,c}^r}{\sum_{i=1}^{K_{l,c}} \mu_{i,c}^r}, \quad \forall\, c, l \tag{23}$$

where $d_{i,c}^2 = \|q_i - v_c\|^2$ and $d_{i,k}^2 = \|q_i - v_k\|^2$ are the distance between the ith user to the cth cluster and kth cluster, respectively.

IV. Multi-Agent DRL-Based mMIMO-NOMA System

A. Overview of DRL

Among value-based reinforcement learning techniques, Q-learning is a widely used and efficient algorithm to address

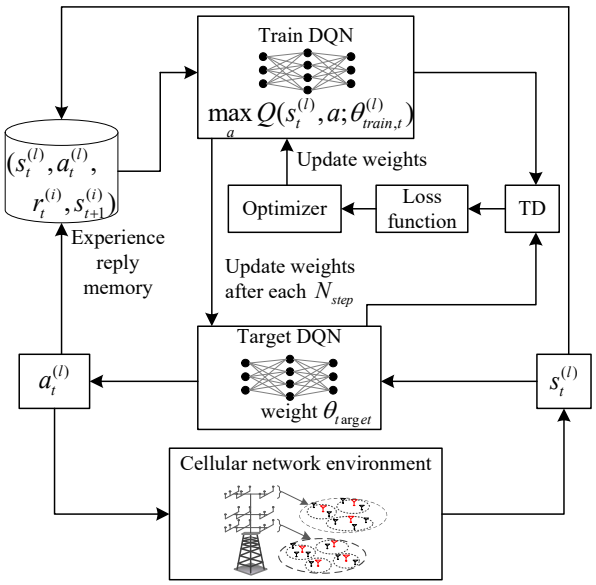

Fig. 3. Multi-agent DRL framework for the proposed mMIMO-NOMA system.

Markov decision process problems [21]. Q-learning represents the mathematical formulation of a decision-making problem to be solved by a decision maker or an agent, defined as the quintuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $\mathcal{S}$ is the set of different states, $s \in \mathcal{S}$; $\mathcal{A}$ is the action space, $a \in \mathcal{A}$; $\mathcal{R}$ is the set of rewards, $r \in \mathcal{R} \rightarrow \mathbb{R}$; $\mathcal{T}$ is the state transition probability, $\mathcal{T}(s_t, a_t, s_{t+1})$; and γ represents a discount factor. An agent follows a policy $\pi(a_t|s_t)$ to execute an action a_t while remaining in state s_t at time t. The agent immediately receives a reward r_t and transits to the next state s_{t+1} according to the transition probability $\mathcal{T}$. The probability of transition to the next state depends only on the current state and action played and not on any previous states or actions, i.e., $Pr(s_{t+1}|s_t, a_t)$.

In Q-learning, the agent seeks to achieve the final goal by considering the future cumulative reward instead of only the immediate reward. A certain policy π of a couple of action and state is associated with an action value function named the Q-value function. Therefore, the agent can achieve the optimal Q-value function by following an optimal policy $\pi^*(a|s_t)$ involving the best action $a_t = a$. The optimal Q-value function can be expressed as

$$Q^*(s_t, a_t) = \max_{\pi} \mathbb{E}\left[R_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} r_{t+\tau} | s_t, a, \pi^*\right], \quad (24)$$

where $\gamma \in (0, 1]$ adds the discounted future rewards to the system. The Q-value function can be represented by the Bellman equation [22], and its convergence can be achieved by iteratively applying the Q-learning algorithm as

$$Q^*(s_t, a_t) = \mathcal{R}(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{T}(s_{t+1}|s_t, a_t) \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}). \quad (25)$$

The algorithm constructs a table of Q values based on the Q-value function. The values are randomly initialized and later updated considering the best action that can be taken in the future state, reflected by the temporal difference (TD). $TD_t(s_t, a_t)$ can be considered an intrinsic reward obtained by the difference between $r_t + \gamma \max_a Q(s_{t+1}, a)$ and $Q(s_t, a_t)$, where r_t is the reward obtained by taking action a_t in state s_t. Artificial intelligence is reinforced with higher values of $TD_t(s_t, a_t)$. Therefore,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha TD_t(s_t, a_t), \quad (26)$$

where $\alpha \in (0, 1]$ is the learning rate and

$$TD_t(s_t, a_t) = r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t). \quad (27)$$

If the action and state spaces are extremely large, as in the case of the proposed system, the classical Q-learning fails because the Q-table cannot be feasibly stored. This problem can be addressed by applying a deep neural network, that is, DQN, which estimates the Q-function. In the proposed DRL framework, two DQNs (actor and critic networks) are used to estimate the action-state value function. The actor DQN constructs a policy according to the observed states and produces an action. The critic DQN evaluates the current policy based on the rewards. The policy can be represented as $\pi(\theta|s, a)$, where θ denotes the weight parameter, which is a real-valued vector. The Q-value function is estimated by the critic DQN, and its policy parameters are updated by the actor DQN [21]. The following gradient rule is applied to update the weight parameter

$$\theta \leftarrow \theta - \alpha \Delta_{\theta} \mathbb{L}(\theta), \quad (28)$$

where Δ_{θ} applies the gradient of loss function $\mathbb{L}(\theta)$. The loss function is computed as the difference between the target and training Q-value functions. Therefore, the loss function is expressed as

$$\mathbb{L}(\theta) = \frac{1}{2\mathcal{M}_b} \sum_{(s_t, a_t, r_t, s_{t+1}) \in \mathcal{D}} \left(TD_t(s_t, a_t)\right)^2, \quad (29)$$

where $\mathcal{D}$ is the mini-batch of $\mathcal{M}_b$ experiences.

B. Proposed Multi-Agent DRL Scheme

In the proposed multi-layer DSCHP system, each layer determines its own SCHP structure and downlink parameters. Therefore, this problem can be formulated as a multi-agent DRL where each layer of the BS acts as an independent agent. We adopt a distributed scheme for the proposed multi-agent DRL. In this approach, the DQN are executed distributively at the antennas groups in each layer. Each agent l holds a copy of target DQN parameter at time t. The proposed model has a train DQN which is trained centrally using the shared experiences from all the agents that reduces the memory and computational overhead of the layers. The illustration of the distributed multi-agent DRL for the proposed DSCHP system is shown in Fig. 3. The agent l executes an action $a_t^{(l)}$ at

Algorithm 2 Proposed multi-agent DRL algorithm for the mMIMO-NOMA scheme

Initialize $counter = 0$, $\mathcal{M}$, and $step = 0$;
Determine the number of layers L;
Initialize train network parameters $\theta_{train}^{(l)}$ randomly for $l = 1, ..., L$;
Set target network parameters $\theta_{target} \leftarrow \theta_{train}^{(l)}$ for $l = 1, ..., L$;
repeat
  Observe the state space $s_t^{(l)}$ at time t for $l = 1, ..., L$;
  repeat
    repeat
      Take action according to ϵ-greedy policy $a_t^{(l)} = arg\max_a q(s_t^{(l)}, a; \theta_{target})$ for $l = 1, ..., L$;
      Receive an immediate reward $r_t^{(l)}$ and reach to a new state $s_{t+1}^{(l)}$ for $l = 1, ..., L$;
      Store new experience $(s_t^{(l)}, a_t^{(l)}, r_t^{(l)}, s_{t+1}^{(l)})$ in its experience pool for $l = 1, ..., L$;
      $counter \leftarrow counter + 1$;
    until $counter == \mathcal{M}$
    A mini-batch $\mathcal{D}$ of size $\mathcal{M}_b$ is sampled from the experience pool, where $\mathcal{M}_b \in \mathcal{M}$;
    Updates the parameters $\theta_{train}^{(l)}$ using the gradient decent optimizer in (27) for $l = 1, ..., L$;
    $step \leftarrow step + 1$;
  until $step == N_{step}$
  Update θ_{target} after each N_{step};
until convergence

time t based on its current state $s_t^{(l)}$, which is obtained by following its updated current policy π_l. The DQN helps to determine the policy π_l from the past experience that maximize the expected future reward. An experience reply memory is formed with a fixed size to store new sets of experiences $(s_t^{(l)}, a_t^{(l)}, r_t^{(l)}, s_{t+1}^{(l)})$ from each agents l. During the training stage, the agent selects a mini-batch $\mathcal{D}^{(l)}$ of $\mathcal{M}_b$ experiences from the reply memory. Once the agent take $\mathcal{D}^{(l)}$, the training DQN updates its parameters to minimize the loss in (27) using an optimizer such as stochastic gradient descent optimizer. The trained DQN then broadcast its latest updated parameters θ_{train} after each N_{step} to update θ_{target}.

1) Beamformer Approximation: A simplified method was proposed in [17] to address the complex valued beamformer problem, where it was decomposed into the transmit power and normalized beamformer. Therefore, inspiring by this technique, the beamformer of the sub-array is represented by the transmit power coefficient and the direction of the beam. Assume that $\mathcal{P}$ is the set of available power levels which is achieved by distributing total transmit power into the subarrays and $K_{l,c}$ users by applying the conditions from (3) and (4). Therefore,

$$\mathcal{P} = \{\mathcal{P}_{l,1}, ...\mathcal{P}_{l,c}\} \tag{30}$$

$$\mathcal{P}_{l,c} = \{\rho_{l,c}^{(1)} P_{l,c}, \rho_{l,c}^{(2)} P_{l,c}, ...\rho_{l,c}^{(K_{l,c})} P_{l,c}\}. \tag{31}$$

A beamforming codebook matrix $\mathcal{C}$ with dimensions of $N_{sa} \times Q$ is designed, in which each column represents a beam directional code satisfying $Q \geq N_{sa}$. The set of beam directional vectors are represented as $\mathcal{D}$ of Q directional codes $d_q \in \mathbb{R}^{N_{sa} \times 1}$, which covers directions in $[0, 2\pi)$ and can be expressed as

$$\mathcal{D} = \{d_0, d_1, ...d_{Q-1}\}. \tag{32}$$

Therefore, at time t, an agent l in the lth layer of the BS can determine the beamformer for the cth subarray by selecting the required power levels and codes from the defined sets.

2) State Space: An agent in the BS layer l has states $s_t^{(l)}$ that are the representative features of the connected channels with the associated clustered users in a given time slot t. The lth layer of the BS obtains the received signal strength and total interference-plus-noise power from each ith user in the cth cluster, at time t i.e., $\rho_{l,c}^{(i)}(t-1)P_{l,c}\left\|h_{l,c}^{H(i)}(t)\mathcal{D}_{l,c}^{(i)}(t-1)\right\|_2^2$ and $I_{IC}^{(i)}(t) + I_{ILC}^{(i)}(t) + \eta_{l,c}^{(i)}$, respectively. Next, evaluates the equivalent channel gain, the channel SINR $\zeta_{l,c}^{(i)}(\mathcal{D}(t-1))$, $\zeta_{l,c}^{(i)}(\mathcal{D}(t-2))$, and achievable rate $\Gamma_{l,c}^{(i)}(\mathcal{D}(t-1))$, $\Gamma_{l,c}^{(i)}(\mathcal{D}(t-2))$ for the lth layer. These parameters are the five constituents of the state space observations $s_t^{(l)}$. Subsequently, based on the action taken at time $t-1$, there exist two member elements for the transmit power $\rho_{l,c}^{(i)}(t-1)P_{l,c}$ and selected normalized beamformer from set $\mathcal{D}$. Another $N_{sw,l}$ member element for the $s_t^{(l)}$ is identified by evaluating the junction configuration matrix $J_l(t-1)$ status at $t-1$. Therefore, $7U_l + N_{sw,l}$ input ports of the DQN are occupied by the local parameters where $U_l = C_l K_{l,c}$ is the number of users in the lth layer.

In addition, we consider several other input features based on the status of the interferers and the iterfered neighbors. These are the shared information between layers (i.e., agents) evaluated to perform an action and the resultant effect on the common objective, which is to maximize the EE. Four input ports are occupied by the interference information from the interferers (interfering layer, $l^{'} \neq l$): i) the interferer $l^{'}$, ii) total interference power pertaining to layer $l^{'}$, iii) normalized beamformer used by the $l^{'}$, and iv) utility function for the $l^{'}$, $\Gamma_{l^{'},c}^{(i)}(\mathcal{D}(t-1))$. Therefore, $4(L-1)$ members exists in the state space $s_t^{(l)}$ corresponding to the interferer information. Similarly, the agent l receives the feedback regarding the effect of its action on the interfered layers, specifically, i) the amount of interference power induced to the neighboring layers, ii) neighbors' received channel gain, and iii) influence on the utility function of the interfered neighbors. The sum yields a total of $7U_l + N_{sw,l} + 7(L-1)$ members in state space $s_t^{(l)}$.

3) Action Space: The action space of the agent determines the number of output ports of the DQN. The power coefficient $\rho_{l,c}$ for each user in the cluster is used to discretize $P_{l,c}$ into $K_{l,c}$ levels. Following the conditions applied in (3) and (4), the total number of power levels for the lth layer is $\sum_{c=1}^{C} \mathcal{P}_{l,c}$. Because the BS take a discrete power value and a specified beamformer code, $\mathcal{D}\sum_{c=1}^{C} \mathcal{P}_{l,c}$ member elements are determined for action space $a_t^{(l)}$. Additionally, actions are executed to control the junction switches at time t. Because the switch has two phases (ON and OFF), $2N_{sw,l}$ additional member elements exist for $a_t^{(l)}$. The total number of output ports in action space $a_t^{(l)}$ is $\mathcal{D}\sum_{c=1}^{C} \mathcal{P}_{l,c} + 2N_{sw,l}$.

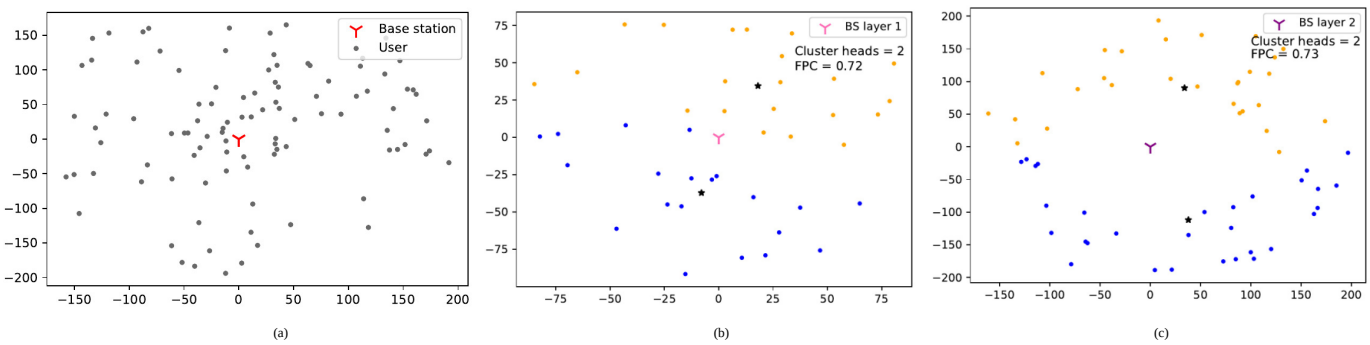

Fig. 4. An example of the BS network configurations, (a) BS with 100 random users, (b) classified users in BS layer 1, and (c) classified users in BS layer 2

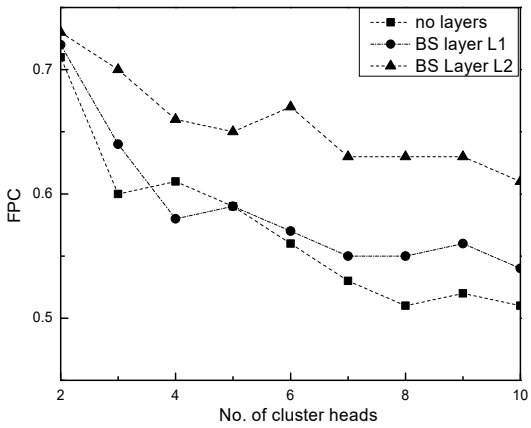

Fig. 5. FCM clustering performance considering multiple cluster heads for 100 users.

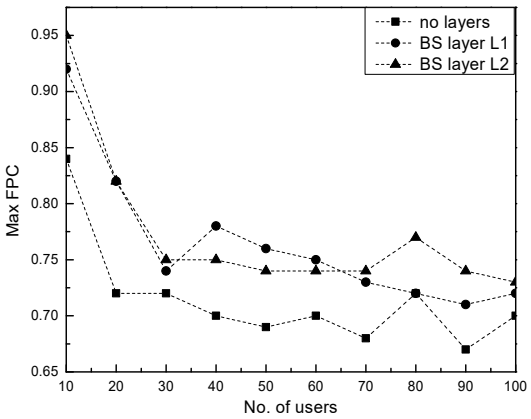

Fig. 6. FCM clustering performance for varying no. of users.

4) Reward Function: As described in Section IV(A), the agent observes the future cumulative rewards when transiting the next state by executing the best action. Therefore, we define the reward function $r_{t+1}^{(l)}$, which indicates the effect of executed action $a_t^{(l)}$ on the optimization of the network objective specified in (20). Assume that to maximize the network objective, agent l selects the best beamformer that results in a higher transmit power in an advantageous direction. Alongside, full array connection to the switches in $J_l(t-1)$ will be required to achieve full array gain at the transmitter. However, full array connection will cause additional power consumption as P_{sw} and P_{ph} that consequently affects the network objective. In addition, while the agent l performs such advantageous actions to augment the system utility, it also instigate higher interference power to its neighboring layers. Therefore, the reward function at time $t+1$ is defined as

$$r_{t+1}^{(l)} = System_{utility}^{(l)} - System_{penalty}^{(l)}, \tag{33}$$

where

$$System_{penalty}^{(l)} = \sum_{l',l' \neq l} log(1 + \zeta_{l',c}^{(i)} \mathcal{D}(t-1)) - System_{utility}^{(l')}, \tag{34}$$

The $System_{utility}^{(l)}$ in (33) refers to the maximization of the spectral efficiency with the highest possible minimization of the connection between the RF chain and consecutive subarrays, represented in (20). The second term, $System_{penalty}^{(l)}$ is the penalty for agent l, which refers to the cumulative loss of the achievable rate of the interfered neighbors $l^{'}$.

V. Simulation Results

The performance of the proposed multi-agent DRL-based DSCHP scheme is evaluated in MIMO-NOMA system. Specifically, we analyze the FCM clustering performance with a variation of number of users and cluster heads. Additionally, the performance of the multi-agent DRL technique has been evaluated with a comparative analysis. First, the DQN network performance is examined by evaluating the training loss for both agents in both the BS layers. Second, the average achievable rate is assessed and compared with the random and greedy approaches. Third, the EE of the proposed DSCHP is compared with that of a scheme involving a full array gain.

In the simulation, we consider a mMIMO BS whose antennas are equally divided into two layers L_1 and L_2. In each layer, a dynamic HP scheme is established according to the configuration shown in Fig. 2 and the antennas group is divided into two subarrays. Hence, two RF chain is connected to the sub-arrays and the junction matrix takes 4 switches. The radius of the cellular BS is considered to be 200 m. The users are randomly located within the cell radius. Firstly, the users

TABLE I
DQN HYPER-PARAMETERS

Parameter	Value
Hidden layer size	64×32
Input Layer size	39
Output layer size	16
Activation function	ReLU
Optimizer	RMSprop
Experience pool memory	500
Mini-batch size	32
Step size	100
Learning rate	0.0005
Reward decay	0.5

u are classified according to the region and being supported by the two layers as

$$u = U_{l1} \quad if\ U_{dist} < 100, \tag{35a}$$

$$u = U_{l2} \quad if\ 100 \leq U_{dist} \leq 200, \tag{35b}$$

where U_{l1} and U_{l2} are the number of users for L_1 and L_2, respectively and U_{dist} is the Euclidean distance of the users from the BS. The antenna groups in L_2 supports the distant users with lower channel gain than that of L_1. Therefore, the SCHP structure will be different for both the layers. For instances, L_1 will intent to minimize the antenna array gain for its nearer users and require comparatively reduced number of RF chains and phase shifters to be connected to the antenna sub-arrays.

In Fig. 4, an example of the bi-layer BS network configuration is shown where 100 users are randomly distributed within a radius of 200 m. Initially, we consider 100 users to effectuate the simulation of FCM clustering on nearly overlapped users. We have distinguished the users in L_1 and L_2 based on (32) which are shown in Figs. 4(b) and (c), respectively. The FCM clustering is applied to users in both the BS layers to allocate them into multiple clusters and each cluster is supported by each antenna sub-array. FPC is used to define how meaningful the data points can be clustered and it is a value between 0 and 1. Fig. 4(b) and (c) represent FPC values for clustered users in both the layers which are found maximum for the case of two cluster heads.

We observed the FPC performance by varying the number of cluster heads from 2 to 10 while considering 100 users. The result has been taken for both the BS layers L_1 and L_2 and comparing them with the conventional non-layering case as shown in Fig 5. The comparison shows a better representation of users can be obtained when the users are classified through antenna layering. A higher FPC is achieved for the case with fewer cluster heads because of the nature of user distribution. A similar analysis on the performance of FCM clustering is shown in Fig. 6. The maximum FPC for different clustering cases in which the number of users is varied from 10 to 100 is recorded. The cases involving BS layering yield higher FPC values, and the clustering is more meaningful in the presence of fewer users.

For the preliminary evaluation of the multi-agent DRL algorithm for the proposed system, we considered two users per cluster and set the hybrid beamformer parameters as $\mathcal{P} = 4$ and $\mathcal{D} = 2$. For the THz channel, the carrier frequency is chosen as 0.34 THz in directional propagation in order to provide large channel capacity and avoid path loss peak [23]. The AWGN power spectral density σ^2 is considered as -174 dBm/Hz. We compute the path loss term between BS and user as given in (11), where the parameters used for frequency dependent molecular absorption coefficient $\kappa(f)$ can be found in [24]. The maximum transmit power for the BS layers is set to 37 dBm. Typical values are used for the other parameters: The power consumption per RF chain P_{rf}, at closed switch P_{sw}, for the 4-bit phase shifter P_{ph}, of the combiner P_{com}, at the power amplifier P_{amp}, and of the baseband $P_{bb}$is 160 mW [1], 24 mW [25], 42 mW [26], 6.6 mW [27], 60 mW [28], and 200 mW [1], respectively.

The hyper-parameter used in the multi-agent DQN algorithm is shown in Table I. In our algorithm, an agent trains a DQN which has an input layer, two fully connected hidden layers, and an output layer. The hidden layers have 64 and 32 neurons, respectively. The total number of input ports is

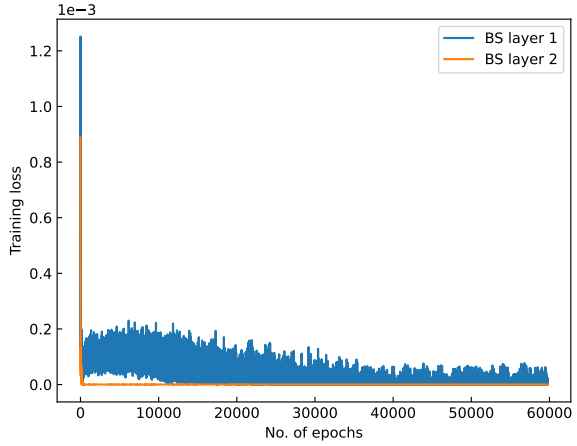

Fig. 7. Training loss function for the two agents in both the BS layers.

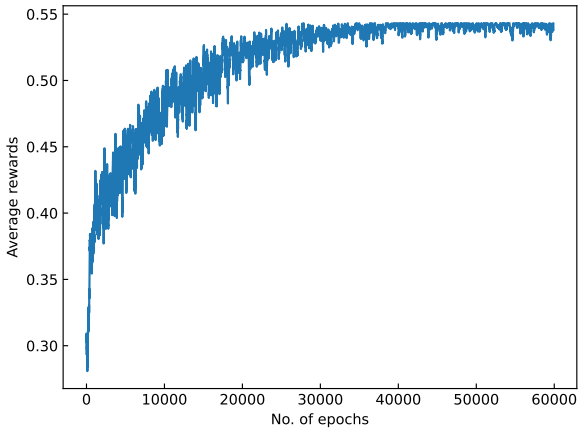

Fig. 8. Average rewards with a moving average value over the last 1000 steps.

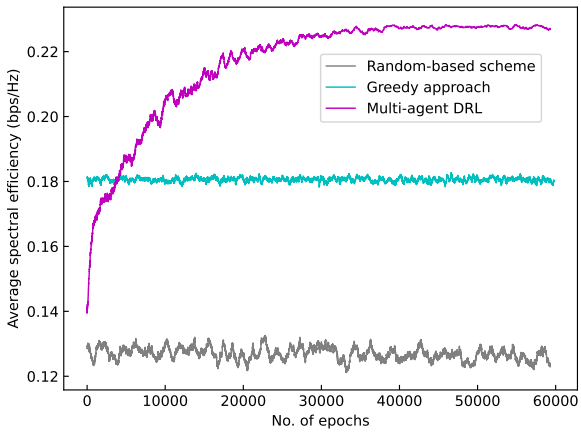

Fig. 9. Average spectral efficiency with a moving average value over the last 1000 steps.

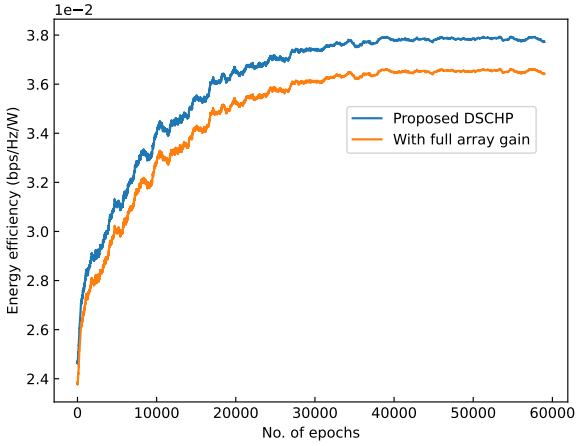

Fig. 10. Energy efficiency compared with full array gain scheme when $N_{rf} = 2$. Representations are a moving average value over the last 1000 steps.

determined from the states described in the Section IV(B). The value of U_l is considered as 4. Therefore, we have a total of $7U_l + N_{sw,l} + 7(L-1) = 39$ input ports in the input layer. In the output layer, the number of output ports is equivalent to the action space length described in the Section IV(B). We considered the length of the beam directional vector set $\mathcal{D} = 2$ and set of the available power levels for cth cluster $\mathcal{P} = 2$. As we considered two sub-arrays per layer, there are four power levels for the lth layer. Therefore, the number of member elements in the action space (i.e., output ports) is $\mathcal{D}\sum_{c=1}^{C}\mathcal{P}_{l,c} + 2N_{sw,l} = 16$. The memory size of the experience pool is set to 500 and the size of the mini-batch is fixed at 32. The target DQN updates its parameters after 100 time slots, i.e., $N_{step} = 100$. We used rectifier linear unit (ReLU) function for each hidden layer to be activated. Additionally, RMSprop optimizer is used to update the parameter in which the initial learning rate is set to 0.0005 and $\lambda = 0.5$.

We evaluate the training loss performance of the proposed multi-agent DRL framework under the fixed learning rate of 0.0005. Because the agents in BS layers 1 and 2 train their networks simultaneously, both the loss performances are recorded to observe the required no. of epochs for convergence. As shown in Fig. 7, the loss function value decreases as training progresses and stabilizes after 150 epochs. The Q values predicted in the later slots are more accurate and stable than those in the initial slots. The obtained average rewards over each iteration are presented graphically in Fig. 8. The reward function is evaluated from (33) which subtracts the received penalty from the system's utility for a particular action. Therefore, based on the training model, the average reward an agent is receiving at every epoch is gradually increasing in pace with the average spectral efficiency. Therefore, the agent learns well the environment gradually.

The average spectral efficiency for the proposed multi-agent DRL-based DSCHP scheme is evaluated. The performance of the proposed scheme is compared with that of the random policy and the greedy approaches. In the random policy,

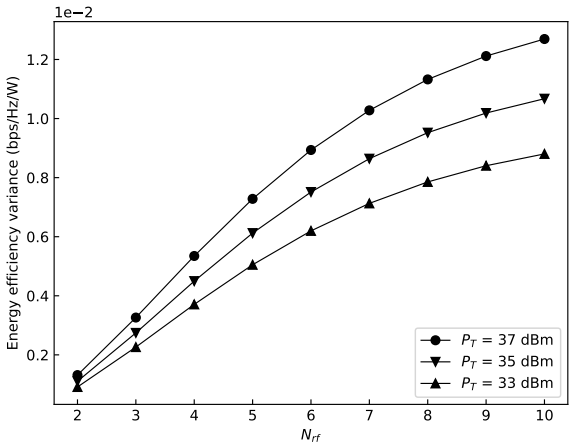

Fig. 11. The deviation of the energy efficiency value of the proposed scheme with respect to the full array gain scheme.

each agent randomly chooses actions. Whereas, in the greedy approach, each layer (i.e., agent) uses the best beamformer to achieve higher channel gain without considering the interference power to the neighboring layers. Which eventually degrades the average spectral efficiency. The proposed system shows better performance as the agent learns about the suitable beamformer by receiving a reward based on the action played. This is because the decision-making policy is enhanced by updating the weights of the trained DQN regularly. As shown in Fig. 9, the proposed scheme starts performing better than the greedy policy after approximately 4,000 time slots and eventually reaches a relatively stable situation in approximately 45,000 epochs. The results show that the multi-agent DRL-based DSCHP scheme can learn the trade-off between maximizing EE and minimizing the interference power to the neighboring layers.

Further, we evaluate the EE of the proposed scheme and compare it with the traditional method. In the traditional approach, to maximize the performance, the full array gain

is exploited by connecting all the switches linked to the RF chain. In this framework, the additional power consumed by the switches and phase shifters decreases the overall EE. Fig. 10 shows the EE for the proposed DSCHP system and scheme with full array gain. As the systems converge, the proposed scheme, in its preliminary configuration, outperforms the traditional scheme by achieving a 3.6% higher EE. The EE can be further enhanced by increasing the number of RF chains in the case of ultra-massive MIMO antenna arrays. The EE variance of the proposed approach and FCHP scheme under different N_{rf} values is shown in Fig. 11. The number of connecting switches increases with increasing N_{rf}. Consequently, in the case of the DSCHP scheme, the number of switches maintaining the OFF state increases with increasing N_{rf}. In addition, the phase shifters connected with these switches become idle. Therefore, the power consumption defined in (19) considerably decreases, thereby enhancing the EE variance with the increment in N_{rf}. The results for three values of P_T indicate that a higher P_T corresponds to an enhanced EE variance. The objective function we have formulated in (20) is to maximize the EE while maintaining the key constraints. The EE is defined as the ratio of the spectral efficiency to the total power consumption. However, for the system, the optimal solution can be defined as achieving maximum spectral efficiency with minimum power consumption which is practically not feasible. Therefore, we find a near-optimal solution that can improve the EE while maintaining a satisfactory achievable sum rate with the least possible power consumption.

VI. Conclusion

This paper proposes a multi-layer DSCHP architecture for THz mMIMO-NOMA systems to enhance their spectral efficiency and EE. The DSCHP scheme allows the number of RF chains connected to the subarray antenna elements to be increased or decreased by sensing the channel conditions and requirements. A multi-agent DRL-based approach is applied to solve the problem of maximizing the utility function. In the proposed framework, an agent is responsible for a BS layer. The agents train a DQN centrally that periodically shares the updated parameters, and the agents execute actions until a stable solution is obtained. The FCM clustering algorithm is used to group users under each subarray and efficiently distribute the power coefficients. The simulation results show that the proposed approach can achieve an excellent spectral efficiency. Moreover, the EE can be enhanced when mMIMO antenna arrays with a larger number of RF chains are used.

References

[1] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. M. Leung, "Energy Efficient User Clustering, Hybrid Precoding and Power Optimization in Terahertz MIMO-NOMA Systems," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2074-2085, Sept. 2020.

[2] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can Terahertz Provide High-Rate Reliable Low Latency Communications for Wireless VR?," *IEEE Internet Things J.*, doi: 10.1109/JIOT.2022.3142674.

[3] L. Yan, C. Han, and J. Yuan, "A Dynamic Array-of-Subarrays Architecture and Hybrid Precoding Algorithms for Terahertz Wireless Communications," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2041-2056, Sept. 2020.

[4] M. Vaezi et al., "Interplay Between NOMA and Other Emerging Technologies: A Survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 900-919, Dec. 2019.

[5] B. Ling, C. Dong, J. Dai and J. Lin, "Multiple Decision Aided Successive Interference Cancellation Receiver for NOMA Systems," *IEEE Wirel. Commun. Lett.*, vol. 6, no. 4, pp. 498-501, Aug. 2017.

[6] K. Higuchi and A. Benjebbour. "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access." *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403-414, 2015.

[7] 5. O. Elkharbotly, E. Maher, A. El-Mahdy, and F. Dressler, "Optimal Power Allocation in Cooperative MIMO-NOMA with FD/HD Relaying in THz Communications," *9th IFIP International Conference on Performance Evaluation and Modeling in Wireless Networks (PEMWN)*, 2020, pp. 1-6.

[8] H. W. Oleiwi, N. Saeed, and H. Al-Raweshidy, Cooperative SWIPT MIMO-NOMA for Reliable THz 6G Communications. *Network*, vol. 2, pp. 257-269, 2022.

[9] H. Sarieddeen, A. Abdallah, M. M. Mansour, M. -S. Alouini, and T. Y. Al-Naffouri, "Terahertz-Band MIMO-NOMA: Adaptive Superposition Coding and Subspace Detection," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2628-2644, 2021.

[10] S. A. Busari et al., "Generalized Hybrid Beamforming for Vehicular Connectivity Using THz Massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8372-8383, Sept. 2019.

[11] Q. Sun, S. Han, C. L. I, and Z. Pan, "On the Ergodic Capacity of MIMO NOMA Systems," *IEEE Wirel. Commun. Lett.*, vol. 4, no. 4, pp. 405–408, 2015.

[12] Z. Ding, R. Schober, and H. V. Poor, "A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 6, pp. 4438–4454, 2016.

[13] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the Sum Rate of MIMO-NOMA and MIMO-OMA Systems," *IEEE Wirel. Commun. Lett.*, vol. PP, no. 99, p. 1, Jun. 2017.

[14] S. Chinnadurai et al., "User clustering and robust beamforming design in multicell MIMO-NOMA system for 5G communications," *AEU-Int. J. Electron. Commun.*, vol. 78, pp. 181-191, 2017.

[15] S. Solaiman, L. Nassef, and E. Fadel, "User Clustering and Optimized Power Allocation for D2D Communications at mmWave Underlaying MIMO-NOMA Cellular Networks," *IEEE Access*, vol. 9, pp. 57726-57742, 2021.

[16] X. Xu, Q. Chen, X. Mu, Y. Liu, and H. Jiang, "Graph-Embedded Multi-Agent Learning for Smart Reconfigurable THz MIMO-NOMA Networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 259-275, Jan. 2022.

[17] J. Ge, Y. -C. Liang, J. Joung and S. Sun, "Deep Reinforcement Learning for Distributed Dynamic MISO Downlink-Beamforming Coordination," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6070-6085, Oct. 2020.

[18] C. Huang et al., "Multi-Hop RIS-Empowered Terahertz Communications: A DRL-Based Hybrid Beamforming Design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663-1677, June 2021.

[19] Y. S. Nasir and D. Guo, "Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239-2250, Oct. 2019.

[20] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The Fuzzy C-Means Clustering Algorithm," *Comput. Geosci.*, vol. 10, no. 2-3, pp. 191-203, 1984.

[21] Y. Huang, C. Xu, C. Zhang, M. Hua, and Z. Zhang, "An Overview of Intelligent Wireless Communications using Deep Reinforcement Learning," *Journal of Communications and Information Networks*, vol. 4, no. 2, pp. 15-29, 2019.

[22] L. Baird, "Residual Algorithms: Reinforcement Learning with Function Approximation," in *proc. of the Twelfth International Conference on Machine Learning*, Tahoe City, California, 1995, p.30-37.

[23] C. Han and I. F. Akyildiz, "Distance-Aware Bandwidth-Adaptive Resource Allocation for Wireless Systems in the Terahertz Band," *IEEE Trans. THz Sci. Technol.*, vol. 6, no. 4, pp. 541-553, July 2016.

[24] J. M. Jornet and I. F. Akyildiz, "Channel Modeling and Capacity Analysis for Electromagnetic Wireless Nanonetworks in the Terahertz Band," *IEEE Trans. Wirel. Commun.*, vol. 10, no. 10, pp. 3211-3221, October 2011.

[25] Y. Kim, H. Lee, and S. Jeon, "A 220–320 GHz single-pole single-throw switch," in *Proc. IEEE RFIT*, Aug. 2016, pp. 1–3.

[26] Y. Kim et al., "A 220–320-GHz vector-sum phase shifter using single Gilbert-cell structure with lossy output matching," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 1, pp. 256–265, Jan. 2015.

[27] Y. Shang, H. Yu, H. Fu, and W. M. Lim, "A 239–281 GHz CMOS receiver with on-chip circular-polarized substrate integrated waveguide antenna for sub-terahertz imaging," *IEEE Tran. THz Sci Techn.*, vol. 4, no. 6, pp. 686–695, Nov. 2014.

[28] L. A. Samoska, "An overview of solid-state integrated circuit amplifiers in the submillimeter-wave and THz regime," *IEEE Trans. THz Sci. Technol.*, vol. 1, no. 1, pp. 9–24, Aug. 2011.

MD. SHAHJALAL (Member, IEEE) received his B.Sc. degree in Electrical and Electronic Engineering (EEE) from Khulna University of Engineering & Technology (KUET), Bangladesh, in May 2017. In 2019 and 2022 he obtained his M.Sc. and Ph.D, respectively in Electronics Engineering from Kookmin University, South Korea and was awarded for his academic excellence. In 2022, he joined the Department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh as an Assistant Professor. He has published around 55 technical papers and patents. He has served as a Reviewer for many international journals and IEEE conferences. His research interests include optical and THz wireless communications, wireless security, non-orthogonal multiple access (NOMA), internet of things, low-power wide-area network, and 6G.

Md. Habibur Rahman (Student Member, IEEE) received B.Sc. degree in Electrical and Electronic Engineering from the Khulna University of Engineering & Technology, Khulna, Bangladesh, in March 2019, and M.Sc. degree in Electronics Engineering from Kookmin University, Seoul, South Korea, in February 2022. He is currently pursuing Ph.D. degree in Electrical and Electronics Engineering at Virginia Tech, VA, USA. Currently, his research interests include machine learning/deep learning (ML/DL) for open radio access network (O-RAN), automating ML/DL pipeline, application of reinforcement learning for wireless resource allocation, cloud computing, and federated learning.

Md Morshed Alam (Graduate Student Member, IEEE) received the B.Sc. degree in Electrical and Electronic Engineering from the Khulna University of Engineering and Technology, Khulna, Bangladesh, in May 2018, and the M.Sc. degree in Electronics Engineering from Kookmin University, Seoul, South Korea. He is currently pursuing his Ph.D. degree from the University of Technology Sydney, NSW, Australia. He completed an exchange program on power systems from the University of Porto, Porto, Portugal, in September 2017. He has authored or co-authored more than 35 technical articles and patents. His research interests include intelligent grid systems, optimization, management and control systems, renewable energy, machine learning and deep learning algorithms, low-power wide-area network, and optical wireless communications. Mr. Morshed was the recipient of the Academic Excellence Award during his M.Sc. from Kookmin University.

Mostafa Zaman Chowdhury (Senior Member, IEEE) received the B.Sc. degree in Electrical and Electronic Engineering from the Khulna University of Engineering & Technology (KUET), Bangladesh, in 2002, and the M.Sc. and Ph.D. degrees in Electronics Engineering from Kookmin University, South Korea, in 2008 and 2012, respectively. In 2003, he joined the Electrical and Electronic Engineering Department, KUET as a Lecturer, where he is currently working as a professor. He worked as a Head, Department of Biomedical Engineering, KUET during April 2022 to December 2023. He was a Postdoctoral Researcher with Kookmin University from 2017 to 2019 supported by National Research Foundation, Korea. He has published research articles in top quality journals such as IEEE Communications Surveys and Tutorials, IEEE Transactions on Services Computing, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Instrumentation and Measurement, IEEE Systems Journal, IEEE Communications Magazine, IEEE Communications Letters, IEEE Consumer Electronics Magazine, IEEE Access, and Nature Scientific Reports. In 2008, he received the Excellent Student Award from Kookmin University. His three papers received the Best Paper Award at several international conferences around the world. He was involved in many Korean government projects. His research interests include convergence networks, QoS provisioning, small-cell networks, Internet of Things, eHealth, 5G and beyond communications, and optical wireless communication. He received the Best Reviewer Award 2018 by ICT Express journal. Moreover, he received the Education and Research Award 2018 given by Bangladesh Community in South Korea. He was a TPC Chair of the International Workshop on 5G/6G Mobile Communications in 2017 and 2018. He has been serving as Publicity Chair of the International Conference on Artificial Intelligence in Information and Communication (from 2019 to 2023) and International Conference on Ubiquitous and Future Networks (2022 and 2023). He served as a Reviewer for many international journals (including IEEE, Elsevier, Springer, ScienceDirect, MDPI, and Hindawi published journals) and IEEE conferences. He has been working as an Executive Editor for ICT Express, an Associate Editor of IEEE ACCESS, an Associate Editor of Frontiers in Communications and Networks, a Lead Guest Editor for Wireless Communications and Mobile Computing, and a Guest Editor for Applied Sciences. He has served as a TPC member for many IEEE conferences.

YEONG MIN JANG (Member, IEEE) received the B.E. and M.E. degrees in Electronics Engineering from Kyungpook National University, Daegu, South Korea, in 1985 and 1987, respectively, and the doctoral degree in Computer Science from the University of Massachusetts, MA, USA, in 1999. From 1987 to 2000, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. Since 2002, he has been with the School of Electrical Engineering, Kookmin University, Seoul, South Korea, where was the Director of the Ubiquitous IT Convergence Center from 2005 to 2010. He has been the Director of the LED Convergence Research Center, Kookmin University, since 2010, the Director of the Internet of Energy Research Center, Kookmin University, since 2018, and the Director of the AI Mobility Research Institute, Kookmin University, since 2021. His research interests include AI mobility, the internet of energy, the IoT platform, AI platform, cloud platform, optical wireless communications, optical camera communication, 5G/6G mobile communications, and the internet-of-things. He has co-authored over 500 technical papers and holds over 120 patents. Dr. Jang is also a fellow of the Korean Institute of Communications and Information Sciences (KICS). He had served as an Executive Director for KICS from 2006 to 2014. He was the President of KICS in 2019. He received the Young Scientist Award from the Korean Government from 2003 to 2006. He was a recipient of the Dr. Irwin Jacobs Award in 2018. He had also served as the Founding Chair for the KICS Technical Committee on Communication Networks in 2007 and 2008. He was/has been the Steering Chair of the Multi-Screen Service Forum from 2011 to 2019, and the Society Safety System Forum from 2015 to 2021, and the ESG Convergence Forum, since 2022. He has also served as the Chairman for the IEEE 802.15 Optical Camera Communications Study Group in 2014, and the IEEE 802.15.7m Optical Wireless Communications Task Group from 2015 to 2019 and successfully published IEEE 802.15.7-2018 and ISO 22738:2020 standard. He has been the Chairman of IEEE 802.15.7a Higher Rate and Longer Range OCC TG, since 2020. He has organized several conferences and workshops, such as the International Conference on Ubiquitous and Future Networks from 2009 to 2017, the International Conference on ICT Convergence from 2010 to 2016, the International Workshop on Optical Wireless LED Communication Networks from 2013 to 2016, the International Conference on Information Networking in 2015, and the International Conference on Artificial Intelligence in Information and Communication since 2019. He is also the Editor-in-Chief of ICT Express.