

Optimizing Spectral Efficiency through User Clustering in User-centric Cell-free Massive MIMO Environment for 6G Wireless Network

Pooya Hejazi and Masoud Sabaei

Abstract—Delivering low-latency, high-throughput wireless communication in the terahertz (THz) frequency range presents challenges for 6G implementation, including signal distortion and destructive interference. To overcome these limitations, we propose a novel approach using multi-antenna beamforming with massive MIMO technology at base-stations (BS) and user equipment (UE). Our centralized optimization mechanism improves spectral efficiency (SE) by clustering UEs within a user-centric cell-free massive MIMO framework. We propose an optimization problem to jointly determine user clustering and BS assignment to maximize total throughput. In this architecture, multi-antenna beamforming at the BSs enables spatially directed transmission and reception, while non-orthogonal multiple access (NOMA) is employed within each cluster to allow multiple UEs to share subcarriers. Uplink signals are resolved using successive interference cancellation (SIC), and downlink precoding leverages estimated uplink channels. This combination ensures efficient resource reuse and robust connectivity even under high user density. Experimental results demonstrate its superiority over existing methods in diverse 6G scenarios, effectively addressing the challenges and requirements of future wireless networks while maintaining computational efficiency.

Index Terms—6G wireless networks, deep-learning, multi-user massive MIMO, spectral efficiency, UE clustering.

I. INTRODUCTION

THE proliferation of innovative applications such as autonomous vehicles, virtual/augmented reality, holographic images, and machine-type communications has necessitated the development of a communication infrastructure characterized by low latency and high throughput to ensure an acceptable quality of experience (QoE) [1]. Traditional microwave spectrum alone proves inadequate in meeting the connectivity demands imposed by these applications over wireless communication channels. Consequently, millimeter-wave (mmWave) spectrum (over 6 GHz) and terahertz (THz) communication (ranging from 100 GHz to 1 THz) have emerged as indispensable elements in addressing these requirements [2]–[4].

Manuscript received January 16, 2025; revised September 27, 2025; approved for publication by Li, Xingwang Li, Division 2 Editor, November 6, 2025.

The authors are with Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran, email: {pooya.hejazi, sabaei}@aut.ac.ir.

M. Sabaei is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2025.000085

However, the utilization of high-frequency electromagnetic waves in the mmWave and THz spectrums presents challenges such as heightened signal distortion and the occurrence of destructive interference. To surmount these issues, the deployment of massive multiple antenna communication, known as multiple-input multiple-output (MIMO), becomes essential. MIMO facilitates directed empowered beamforming towards user equipment (UE), offering beamforming and multiplexing gains. These gains significantly enhance spectral efficiency (SE) but also introduce complexities in channel estimation [5].

Traditional channel estimation methods, such as least square error (LSE) and minimum mean square error (MMSE) [6], exhibit limitations including communication delays and potential waste of wireless resources, especially in low signal-to-noise ratio (SNR) scenarios. Deep learning-based channel estimation methods have emerged as a solution offering low complexity and highly efficient channel estimation in complex massive MIMO (mMIMO) environments [7]–[14].

To optimize user throughput and alleviate backhaul traffic, user-centric cell-free mMIMO (UC-CF mMIMO) [15]–[17] has been introduced. The joint design of beamforming and user-centric base station (BS) clustering has been proposed to optimize backhaul traffic in beyond 5G networks [18]. However, existing literature often overlooks the benefits of multi-user mMIMO (MU-mMIMO) and fails to fully exploit its multiplexing gains.

Several recent works have investigated machine learning techniques for user clustering and resource allocation in mMIMO and UC-CF mMIMO systems. For example, reinforcement learning (RL) frameworks have been applied to dynamic scheduling, clustering, and power control, as in [19], where a hierarchical deep RL model is used for joint AP clustering and power allocation in CF-mMIMO downlink systems. Similarly, [20] proposes a soft actor-critic-based scheduling agent for mMIMO that balances quality of service (QoS) with computational efficiency. Deep learning approaches have also been adopted in [21], which integrates user association and resource allocation via supervised neural models, and [22], which employs semi-supervised learning for joint non-orthogonal multiple access (NOMA) user pairing and power assignment. In contrast to these standalone ML or RL methods, our approach introduces a hybrid framework that combines optimization and learning in a novel way. First, we formulate and solve a constrained optimization problem to generate ground-truth cluster assignments and BS

Creative Commons Attribution-NonCommercial (CC BY-NC).

This is an Open Access article distributed under the terms of Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided that the original work is properly cited.

mappings under strict QoS and interference criteria. These optimal assignments are then used to train an ensemble deep learning model with a custom loss function that embeds the constraints directly into the training process. This allows the model to generalize feasible and near-optimal decisions without compromising performance or violating resource and QoS limits. In this way, our method achieves the robustness and interpretability of optimization with the fast inference and scalability of deep learning. This hybrid design, unlike prior RL-only [19], [20] or unsupervised clustering-based approaches [23], [24], results in significantly reduced computational complexity during inference while maintaining high SE, as shown in our evaluations. In parallel, unsupervised clustering—especially K-means—has been widely used to form user groups or reduce pilot contamination in CF-mMIMO networks. For example, [23] applies K-means to improve SE in mmWave NOMA scenarios, while [24] uses it for cell-free MIMO pilot reuse and AP grouping. Although effective for reducing dimensionality, such heuristic approaches typically overlook QoS constraints, inter-cluster interference, or non-orthogonal access limitations. Additionally, dynamic clustering has gained attention as a key enabler for channel-aware optimization in wireless networks. Lim *et al.* [25] proposed a cross-entropy-based UE clustering strategy for mmWave-NOMA systems, jointly optimizing beamforming and power control while accounting for imperfect successive interference cancellation (SIC). Their channel-aware dynamic clustering strategy enhances SE in mMIMO contexts. Similarly, Humadi *et al.* [26] presented a user-centric dynamic BS clustering mechanism tailored for hybrid sub-6 GHz–mmWave–THz networks. Their approach continuously adjusts serving BSs based on instantaneous channel quality to improve reliability and coverage. In parallel, Wang *et al.* [27] developed a training channel-based technique to estimate mmWave MIMO channel parameters in complex urban mobility scenarios. Their method improves dynamic channel modeling accuracy by exploiting spatio-temporal correlations between beam-domain training and actual channels—providing useful insights for real-time user clustering and mobility-aware optimization. In the context of secure multiuser communications, Guo *et al.* [28] proposed a threshold-based scheduling algorithm to improve physical layer security in multiuser satellite systems. Their focus is on maximizing secrecy rate by selecting users under eavesdropping constraints, whereas our work emphasizes SE in a terrestrial UC-CF mMIMO architecture. Although both works involve constrained user scheduling, our approach addresses interference-aware clustering and BS assignment using a hybrid optimization and deep learning framework, without targeting security aspects or satellite environments.

Furthermore, since THz communication suffers from high path loss and fading, we model our system under LOS-dominant short-range conditions and compensate for fading using multi-antenna diversity and BS selection, as described in Section II.C.

In this paper, we address these gaps by leveraging the MU-mMIMO model to achieve high-rank channel coefficient matrices, maximizing multiplexing gain. We exploit channel

reciprocity in time division duplexing (TDD) and perform channel estimation solely in the uplink (UL) direction to optimize pilot signal duration within the coherence interval. To mitigate destructive interference stemming from NOMA [29], SIC is utilized and UEs are clustered based on maximized achievable SE. This approach ensures interference-free channels when the number of UEs within each cluster is sufficiently high.

For downlink (DL) communications, we utilize the estimated channel matrix as a precoder to cancel interference and noise at each UE, leveraging the concept of favorable propagation [30]. To enable multiple connectivity between different BSs and a single cluster, we assign orthogonal frequency division multiplexing (OFDM) sub-carriers to each BS/Cluster pair, enabling simultaneous communication between each UE cluster and multiple BSs while eliminating intra-cluster interference.

We propose an optimization problem to maximize total throughput, determining UE membership to clusters and BS assignments. The output of the network is split into two parts: one for predicting UE-cluster membership (x) and another for BS assignments (y). By solving this optimization problem, the model presented in Fig. 1 is proposed to map the optimization problem to its alternative deep learning model to solve the radio resource management (RRM) problem with less computational complexity and highest achievable accuracy. Based on the model, we present ensemble deep learning-based approaches with custom loss function to solve this optimization problem with precision comparable to mathematical models but with lower computational complexity. The ensemble deep learning model brings us higher precision while the customized loss function adapt the learning model with the constraints used in the optimization problem.

The contributions of our research can be summarized as follows:

- 1) Proposal of a UE clustering method based on estimated channel coefficients to minimize destructive interference in MU-mMIMO systems.
- 2) Formulation of an optimization problem to maximize total throughput and proposal of deep learning-based solutions to address it.
- 3) Linearize optimization problem in order to be solved in polynomial time.
- 4) Generate training/testing datasets by solving the optimization problem to be used for the deep learning model.
- 5) Develop the multi-layer perceptron (MLP) models with customized loss function to solve the UE clustering and BS assignment problem with less computational complexity and approximately same accuracy.
- 6) Considering batch normalization and custom loss function, developed based on the constraint of the optimization problem and ensemble learning to improve the decision making of BS assignment and UE membership.

The rest of the paper is organized as follows: Section II explains the system model. Section III contains analytic results, and Section IV concludes the paper.

Notation: Bold lower-case letters and upper-case letters denote vectors and matrices, respectively. $\mathbf{x}^T$ and $\mathbf{x}^H$ refer

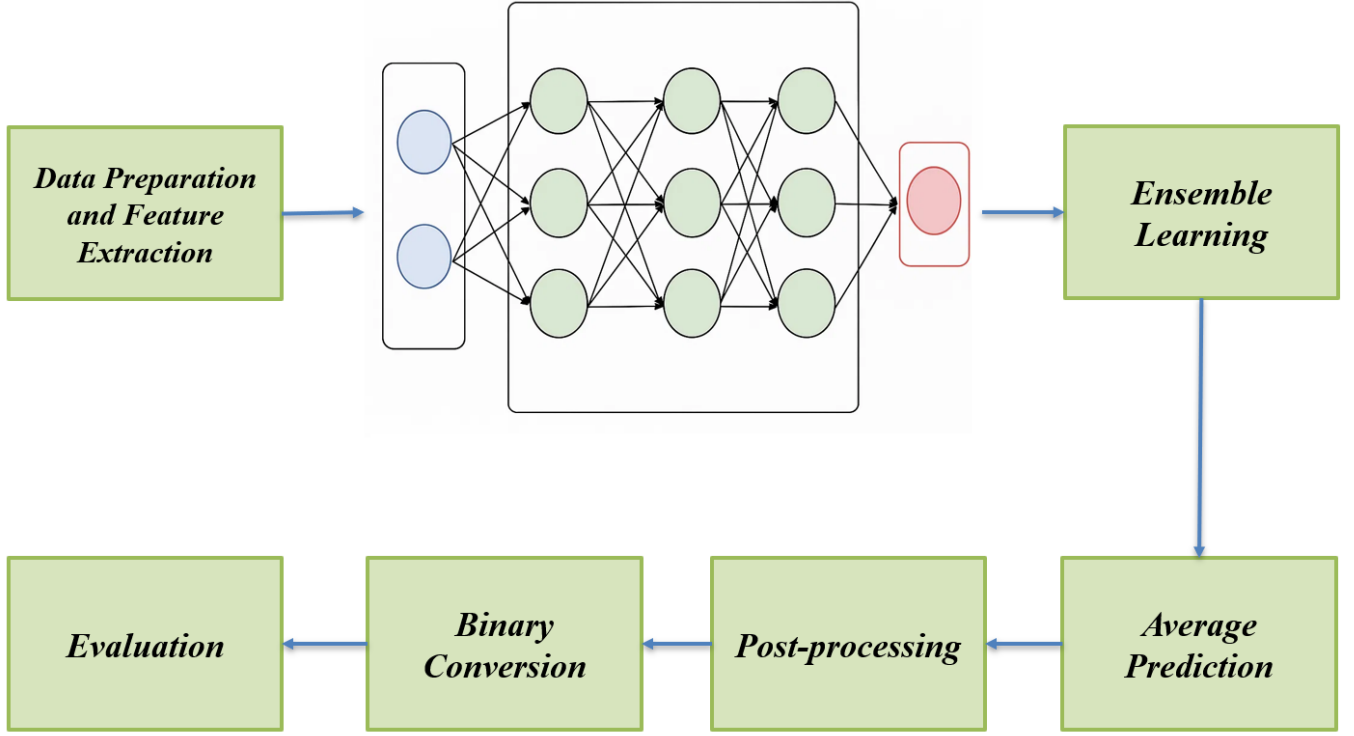


Fig. 1. Optimization problem mapping process to its alternative deep learning model.

to the transpose and Hermite transpose of the vector.

II. SYSTEM MODEL

To accurately model the RRM in our study, we begin by formulating optimization-based framework to solve UE clustering in a UC-CF mMIMO scenario. We assume a quasi-static topology where UEs are stationary or slowly moving, allowing us model the wireless channel as slow-fading, with consistent channel state information (CSI) during the coherence interval.

Each BS is equipped with multiple antennas, and each UE is served by all available BSs. To optimize the channel estimation, we leverage TDD and exploit channel reciprocity models, which enables the channel to be estimated using only UL pilot transmission, thereby minimizing pilot overhead.

We adopt the DeepMIMO dataset framework proposed in [31], which provides realistic channel matrices derived from ray-tracing in mmWave and THz scenarios. This dataset supports our model with geometry-consistent CSI and allows comparison with other distributed RRM approaches in high-frequency wireless systems.

To ensure efficient utilization coherence interval (denoted as τ_c), we divide it into three non-overlapping segments:

- 1) **Pilot transmission (UL):** UEs transmit predefined pilot sequences for UL channel estimation.
- 2) **Data transmission (UL):** UEs send UL data using the estimated channel.
- 3) **Data transmission (DL):** The BSs transmit DL data using the UL-estimated CSI via TDD reciprocity.

This structured division ensures synchronized and efficient access while supporting both UL and DL operation using minimal overhead.

A. UE Clustering

In the first phase of our system design, we focus on efficient clustering of UEs to maximize total SE. As illustrated in Fig. 2, the proposed clustering mechanism enables the deployment of UCCF-mMIMO, where distributed BSs cooperatively serve UE clusters over shared OFDM subcarriers.

To achieve this, we formulate an optimization problem that jointly determines the membership of each UE in a specific cluster, and the assignment of BSs and subcarriers to each cluster.

This cluster-based resource allocation ensures that communication within each cluster remains efficient, while minimizing interference between clusters.

Within each cluster, we employ NOMA [32] techniques to enable UEs to transmit their signals to the BS in the UL direction. Taking advantage of favorable propagation conditions and employing SIC, we can ensure that UEs can transmit their signals with minimal interference.

The combination of advanced clustering, resource-aware subcarrier allocation, and SIC-enabled uplink reception ensures that each user experiences high data rates and robust connectivity. This architecture is particularly well suited for high-density and high-frequency 6G networks.

We define the system components and variables as follows.

- $U = u_1, u_2, \dots, u_i$: the set of UEs.
- $B = b_1, b_2, \dots, b_j$: the set of distributed BSs.

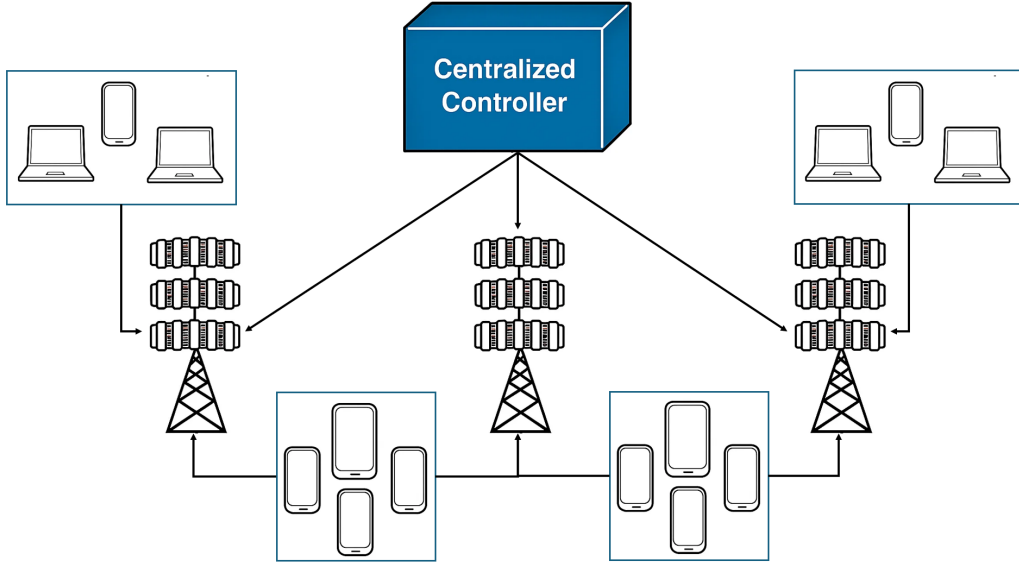


Fig. 2. UE clustering in UC-CF mMIMO environment.

- $C = c_1, c_2, \dots, c_k$: the set of UE clusters.
- $S = s_1, s_2, \dots, s_l$: the set of OFDM subcarriers available.
- $A_b = a_1, a_2, \dots, a_m$: the set of antennas deployed in BS b .
- $\mathbf{H}_{i \times j \times l \times m}$: the full-dimensional channel matrix, where each element represents the channel coefficient from UE u_i to BS b_j in subcarrier s_l and antenna a_m .
- x_c^u : binary decision variable that indicates whether UE u belongs to cluster c .
- $y_c^{b,s}$: binary decision variable that indicates whether BS b is assigned to the cluster c on subcarrier s .

Based on these definitions, the UE clustering and resource allocation problem is formulated. The objective function F is given by (1):

$$F = (\tau_c - \zeta) \times \sum_{b \in B} \sum_{s \in S} \sum_{c \in C} y_c^{b,s} \times R_c^{b,s}. \quad (1)$$

Here,

- τ_c is the coherence interval. In a typical 5G scenario, we can consider the Bandwidth Interval and Time Interval as 100 KHz and 1 msec respectively. Therefore the typical value of τ_c can be presented in (2):

$$\tau_c = 100000 \times 0.001 = 100 \text{ Symbols}. \quad (2)$$

- ζ is maximum pilot length and can be calculated by (3):

$$\text{argmax}_{c \in C} \sum_{u \in U} x_c^u. \quad (3)$$

- $y_c^{b,s}$ is a decision variable representing the assignment of BS b to cluster c on sub-carrier s .
- $R_c^{b,s}$ is the rate function for cluster c on sub-carrier s and BS b . It's defined in (4):

$$R_c^{b,s} = \log(1 + \rho_{ul}^c \|\mathbf{G}_c^{b,s}\|^2), \quad (4)$$

where $\|\mathbf{G}_c^{b,s}\|^2 = \sum_{u \in U} \sum_{a \in A_b} x_c^u \times \|\mathbf{H}_u^{b,s,a}\|^2$. In addition,

$$\rho_{ul}^c = \rho_{ul} \sum_{u \in UE} x_c^u.$$

Here, ρ_{ul} is a parameter representing the signal power.

1) *Channel gain calculation* $\alpha_c^{u,b,s}$: In order to compute the channel gain between UE u , BS b , and sub-carrier s , we calculate $\alpha_c^{u,b,s}$ as the sum of squared channel coefficients across antennas, adjusted by path loss. Specifically:

$$\alpha_c^{u,b,s} = \frac{1}{\text{path_loss_ratio}} \times \sum_{a \in A_b} (\|\mathbf{H}_u^{b,s,a}\|)^2. \quad (5)$$

Here,

- $\mathbf{H}_u^{b,s,a}$ is the channel coefficient between the UE u and the BS b on the subcarrier s and the antenna a .
- The path loss ratio is calculated as the ratio between the distance of UE u to BS b at the current and original positions, raised to the power of the path loss exponent.

2) *Constraints:*

- 1) **UE membership constraint:** Each UE must be assigned to exactly one cluster.

$$\sum_{c \in C} x_c^u = 1, \quad \forall u \in U. \quad (6)$$

- 2) **BS assignment constraint:** Each cluster can be assigned to at most one BS per sub-carrier.

$$\sum_{s \in S} y_c^{b,s} \leq 1, \quad \forall b \in B, \forall c \in C. \quad (7)$$

- 3) **Sub-carrier assignment constraint:** Each BS can be assigned to at most one cluster per sub-carrier.

$$\sum_{c \in C} y_c^{b,s} \leq 1, \quad \forall b \in B, \forall s \in S. \quad (8)$$

- 4) **Maximum pilot constraint:** The number of UEs assigned to any cluster cannot exceed the maximum pilot limit.

$$\sum_{u \in U} x_c^u \leq \zeta, \quad \forall c \in C. \quad (9)$$

- 5) **Cluster connectivity constraint:** Each UE u assigned to a cluster c must ensure that the cluster c is connected to at least one BS b on a sub-carrier s . This constraint ensures that a UE can only be part of a cluster if the cluster has connectivity with a BS, which guarantees communication for the UE within the assigned cluster.

$$x_c^u \leq \sum_{b \in B} \sum_{s \in S} y_c^{b,s}, \quad \forall u \in U, \forall c \in C. \quad (10)$$

Here, x_c^u represents the membership of UE u in cluster c , and $y_c^{b,s}$ denotes the assignment of BS b to cluster c on sub-carrier s .

- 6) **Receiver sensitivity rule:** The channel SNR for any UE u communicating with a BS b on sub-carrier s within a cluster c must meet the minimum receiver sensitivity requirement. This rule ensures that the communication between the UE and BS is of sufficient quality.

$$\alpha_c^{u,b,s} \times y_c^{b,s} \geq \gamma \times x_c^u, \quad \forall u \in U, \forall b \in B, \forall s \in S, \forall c \in C. \quad (11)$$

Here, $\alpha_c^{u,b,s}$ denotes the channel gain between the UE u and BS b for sub-carrier s in cluster c , and γ represents the minimum required receiver sensitivity.

These constraints ensure that the optimization problem is properly formulated and reflects the requirements of the problem scenario.

B. Non-linear to Linear Translation

The original optimization problem involves non-linear components, particularly in the objective function and some constraints, which makes it challenging for linear solvers to handle. The primary sources of non-linearity include:

- The product of decision variables x_c^u (indicating UE membership in a cluster) and $y_c^{b,s}$ (indicating the assignment of a BS to a cluster on a sub-carrier).
- The logarithmic rate function $R_c^{b,s}$, which models the achievable SE.

To address these non-linearities, we apply the following transformations:

1) *Linearization of the product of decision variables:* To linearize the product of x_c^u and $y_c^{b,s}$, we introduce an auxiliary binary variable $z_c^{u,b,s}$, which represents the interaction between the UE-cluster membership and the BS-sub-carrier assignment. This auxiliary variable allows us to replace the non-linear product $x_c^u \cdot y_c^{b,s}$ with linear constraints, as follows:

$$z_c^{u,b,s} \leq x_c^u, \quad z_c^{u,b,s} \leq y_c^{b,s}, \quad z_c^{u,b,s} \geq x_c^u + y_c^{b,s} - 1. \quad (12)$$

These constraints ensure that $z_c^{u,b,s}$ equals 1 only when both x_c^u and $y_c^{b,s}$ are 1, and 0 otherwise.

2) *Linear approximation of the rate function:* The original rate function $R_c^{b,s}$ involves a logarithmic expression, which makes it non-linear. The rate function is defined as:

$$R_c^{b,s} = \log(1 + \rho_{ul}^c \cdot \|\mathbf{G}_c^{b,s}\|^2), \quad (13)$$

where $\mathbf{G}_c^{b,s}$ is a channel gain between UE u , BS b , and sub-carrier s . To linearize this, we approximate the logarithmic rate function by using a first-order Taylor expansion, which results in the following linear approximation:

$$R_c^{b,s} \approx v \cdot (\rho_{ul} \cdot \alpha_c^{u,b,s}) + w. \quad (14)$$

Here,

- $\alpha_c^{u,b,s}$ is the pre-computed channel gain coefficient between UE u , BS b , and sub-carrier s , which is linear in nature.
- v and w are constants used to approximate the logarithmic function.

3) *Linearized objective function:* After applying the linearization of both the product of decision variables and the rate function, the linearized version of the objective function becomes:

$$F = \sum_{c \in C} \sum_{u \in U} \sum_{b \in B} \sum_{s \in S} ((\tau_c - \zeta) z_c^{u,b,s} (v(\rho_{ul} \alpha_c^{u,b,s}) + w)). \quad (15)$$

This linearized objective function can now be efficiently solved by linear optimization solvers.

4) *Auxiliary constraints:* In addition to linearizing the objective function, we ensure that the auxiliary variable $z_c^{u,b,s}$ is properly constrained. These constraints, listed below, guarantee that the auxiliary variable correctly reflects the product of x_c^u and $y_c^{b,s}$:

$$z_c^{u,b,s} \leq x_c^u, \quad z_c^{u,b,s} \leq y_c^{b,s}, \quad z_c^{u,b,s} \geq x_c^u + y_c^{b,s} - 1. \quad (16)$$

These constraints ensure the correct behavior of the linearized variables, allowing us to maintain the problem's original structure and relationships between decision variables in a linear format.

To handle the non-linearity in our CF-mMIMO model, we introduce linearization techniques. The product of decision variables, representing the interaction between UE membership and BS assignment, is linearized using an auxiliary variable. The logarithmic rate function is approximated using a linear function with parameters determined by the signal power and channel coefficients. These adjustments make the problem solvable by standard linear optimization solvers, enabling efficient data generation for the training of deep learning models that aim to solve the same optimization problem. The final linear model is as below:

Objective Function:

The objective function F is given by (17):

$$F = \sum_{c \in C} \sum_{u \in U} \sum_{b \in B} \sum_{s \in S} ((\tau_c - \zeta) z_c^{u,b,s} (v(\rho_{ul} \alpha_c^{u,b,s}) + w)). \quad (17)$$

subject to:

$$1) \sum_{c \in C} x_c^u = 1, \quad \forall u \in U.$$

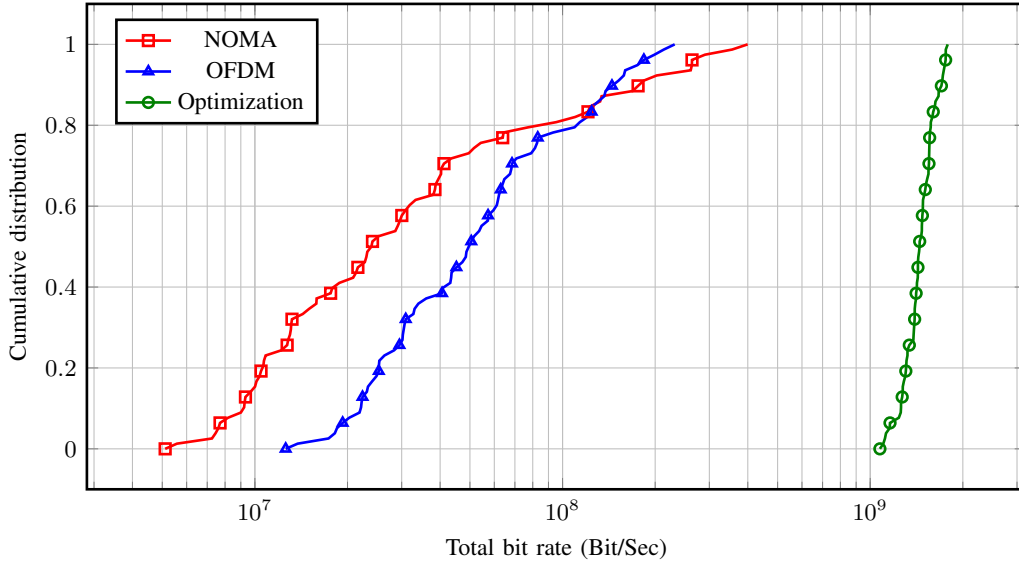


Fig. 3. CDF comparison of total SE among NOMA, OFDM, and optimization-based methods.

- 2) $x_c^u \leq \sum_{b \in B} \sum_{s \in S} y_c^{b,s}, \quad \forall u \in U, \forall c \in C.$
- 3) $\sum_{s \in S} y_c^{b,s} \leq 1, \quad \forall b \in B, \forall c \in C.$
- 4) $\sum_{c \in C} y_c^{b,s} \leq 1, \quad \forall b \in B, \forall s \in S.$
- 5) $\sum_{u \in U} x_c^u \leq \zeta, \quad \forall c \in C.$
- 6) $\alpha_c^{u,b,s} \times y_c^{b,s} \geq \gamma \times x_c^u, \quad \forall u \in U, \forall b \in B, \forall s \in S, \forall c \in C.$
- 7) To linearize the product of decision variables x_c^u and $y_c^{b,s}$, an auxiliary variable $z_c^{u,b,s}$ is introduced, and the following constraints are used:
 - $z_c^{u,b,s} \leq x_c^u, \quad \forall c \in C, \forall u \in U, \forall b \in B, \forall s \in S.$
 - $z_c^{u,b,s} \leq y_c^{b,s}, \quad \forall c \in C, \forall u \in U, \forall b \in B, \forall s \in S.$
 - $z_c^{u,b,s} \geq x_c^u + y_c^{b,s} - 1, \quad \forall c \in C, \forall u \in U, \forall b \in B, \forall s \in S.$

TABLE I
SIMULATION PARAMETERS.

Parameter	Value
Number of BSs	5
Number of antennas per BS	32
Number of sub-carriers	4
Number of UEs	1100
Area size	500 × 500
Path loss exponent	3.5
Receiver sensitivity	−210 dBm
Number of iterations	110
OFDM bandwidth	3 GHz
Number of paths per UE	10
Max pilot length (ζ)	10
Uplink power (p_{ul})	1000 mW
Noise power density (ν)	−174 dBm

C. Channel Characteristics and THz Fading Compensation

To realistically represent THz and high-frequency mmWave communication behavior, our channel model integrates the effects of rapid signal attenuation, molecular absorption, and sparse scattering—typical in these bands. We apply the following assumptions and mitigation strategies:

- **Short-range, static UE setting:** We simulate a quasi-static scenario with low mobility and UE-to-BS distances tailored for hotspot and indoor THz deployments, reducing Doppler and alignment issues.
- **Directional beamforming:** We utilize the DeepMIMO dataset [31], which models directional beamforming and frequency-aware path loss, calibrated for high-frequency bands. The environment is dominated by LOS and first-order reflections.
- **Channel gain aggregation:** Our optimization and learning models compute aggregated channel energy across multiple antennas and OFDM subcarriers. This provides spatial and frequency diversity that statistically smooths out fading and improves link reliability.
- **Flexible BS selection:** UEs are clustered such that each cluster can be associated with multiple BSs, enabling

selection of the best-quality link per UE from among available BSs—an effective strategy for overcoming deep fades or link blockage.

These considerations ensure that while the fading challenge in THz communication is significant, it is systematically addressed in our framework via modeling choices, architectural redundancy, and optimization-aware learning.

III. ANALYTIC RESULTS

In our study, we conducted extensive model execution to validate the proposed UE clustering approach and its impact on system performance. The results were obtained using a combination of optimization techniques and neural network models, as described in the methodology. We considered a network consisting of 5 BSs, each equipped with 32 antennas, 4 sub-carriers (clusters), and 1100 UEs (110 scenarios of 10 UEs). The UEs and BSs were randomly positioned within a 500 × 500 area. The path loss exponent was set to 3.5, and the receiver sensitivity was set to −210 dBm, converted to linear scale for the calculations. The parameters used in the model execution are summarized in Table I.

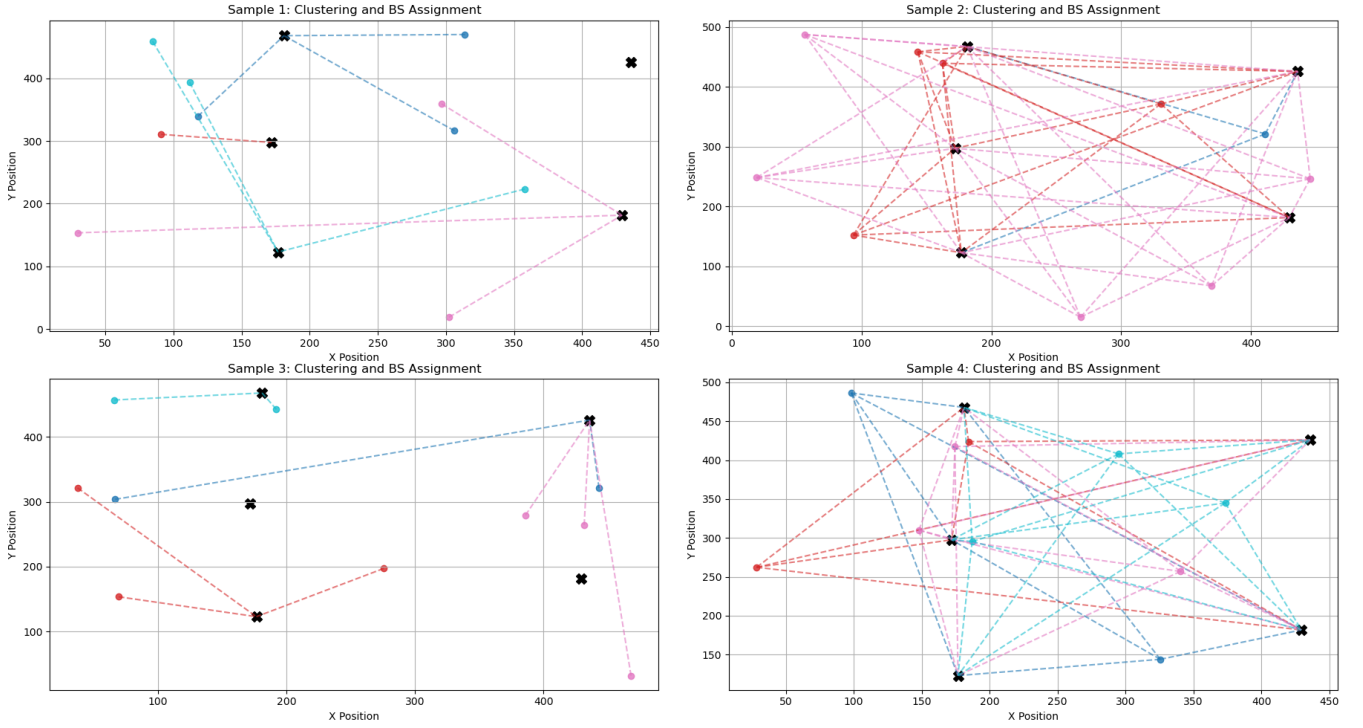


Fig. 4. Samples of cluster membership and BS assignment.

We formulated an optimization problem to maximize the total SE by assigning UEs to clusters and clusters to BSs while satisfying various constraints, such as UE-cluster membership, cluster-BS assignment, cluster-sub-carrier allocation, maximum pilot length, and receiver sensitivity. The problem was solved using the CBC linear solver, and the results showed efficient clustering of UEs with optimal resource allocation. To provide a comparative evaluation of UE throughput across multiple schemes, we plotted the cumulative distribution function (CDF) of total bit rates achieved by the proposed optimization-based method, NOMA, and OFDM (Fig. 3). The bit rate is calculated by multiplying the SE (bit / s / Hz) by the system bandwidth, which is set to 60 GHz (bandwidth = 60×10^9) in our model. As the plot reveals, our optimization approach consistently delivers higher bit rates to a larger proportion of UEs, reflecting its ability to intelligently assign UEs and BS under interference and QoS constraints. In contrast, NOMA and OFDM exhibit more gradual increases in CDF, indicating that fewer UEs achieve high throughput. This highlights the performance gain achieved by jointly optimizing UE clustering and sub-carrier assignment in our model.

The cluster membership and BS assignment samples are illustrated in Fig. 4. As shown in the figure, the channel coefficients play the key role in creating cluster membership and BS assignment. The figure contains four subplots, each showing the clustering and BS assignments for a different sample of UEs within a spatial domain. UEs are represented by dots colored uniquely per cluster, whereas BSs are marked with black “X” symbols. Dashed lines connect UEs to their assigned BSs, with line colors matching the cluster colors to emphasize the UE-cluster relationships.

To further evaluate the robustness and stability of the

proposed optimization framework, we analyze the behavior of the model in training executions. Fig. 5(a) illustrates the histogram of UE-level symbol rate, showing that the majority of UEs’ SE concentrated between 130 and 160 bit/sec/Hz. The mean symbol rate for all UEs is 149.82 bit/sec/Hz, which confirms consistent throughput and fairness. Fig. 5(b) presents the usage trends of sub-carriers over all optimization runs. Each sub-carrier is reused frequently and evenly, with less than 5% standard deviation among runs. This confirms that the resource allocation mechanism efficiently and consistently utilizes the available spectrum resources. Fig. 5(c) illustrates the evolution of the size of the clusters in different executions. Although some variation exists due to dynamic UE distributions, each cluster consistently maintains UE membership within narrow 95% confidence bounds. This confirms that the clustering process adapts well without leading to over- or under-clustering. These analyses highlight the stability of the model under various conditions and provide additional evidence of its practical viability in massive, user-centric, large-scale MIMO environments without cell networks.

To further enhance the clustering process, we trained a neural network model using the features and labels extracted from the optimization problem results. The neural network model consisted of five dense layers with dropout regularization to prevent overfitting. The model achieved high accuracy in predicting the UE assignments to cluster and Clusters assignment to BS.

A. Training and Inference Time Analysis

The proposed system separates training and inference to maximize performance and efficiency. The training phase solves a constrained integer programming problem for each

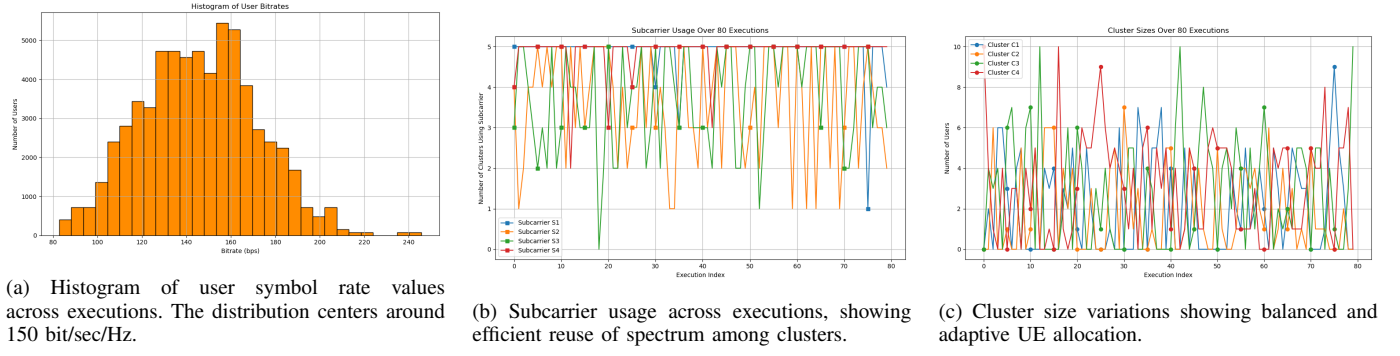


Fig. 5. Analysis of optimization stability over all executions: (a) UE bitrates, (b) subcarrier usage, and (c) cluster size variation. These results confirm robustness and fairness in clustering and resource allocation.

training sample using the CBC solver and trains an ensemble of five neural networks with a constraint-aware custom loss function. Each MLP consists of five layers with 256, 1024, 512, 256, and m neurons. The online inference time for each model is $\mathcal{O}(950,272 + 128m)$, resulting in a total inference complexity of $\mathcal{O}(4.75 \times 10^6 + 640m)$ for the ensemble. Since m (the number of decision variables) remains constant for a given scenario, the inference remains lightweight and supports real-time deployment.

IV. MLP MODEL ARCHITECTURE

This section provides a detailed analysis of the deep learning model developed to predict UE clustering and BS assignment in UCCF-mMIMO networks. The model integrates complex operational constraints into its architecture through a custom loss function and utilizes ensemble learning to enhance prediction accuracy and stability. The model uses input features such as UE positions, BS positions, and channel coefficients and extract the features based on sample data provided by optimization results. The feature extraction process comprises flattening UE positions, BS positions, and channel coefficients into feature vectors, which are then used as input for the deep learning model. The input features are then normalized. This normalization ensures that the features have consistent scales, which accelerates training convergence and helps the model learn effectively.

A. Custom Loss Function

The core component for enforcing the optimization constraints is handled in the custom loss routine, which integrates the following constraints into the learning process:

- **Binary crossentropy loss:** The baseline loss function is binary crossentropy, ensuring proper binary classification of x and y values.
- **Penalty terms for constraints:** To enforce the optimization constraints during training, weighted penalty terms are added for each constraint in optimization problem, ensuring the training model meets the optimization problem requirements.

This loss function effectively combines the classification loss with constraint enforcement, guiding the model toward solutions that satisfy the underlying optimization problem.

B. Model Architecture

The deep learning model is a MLP (Fig. 1) with several layers:

- **Dense layers:** Multiple dense layers are used to capture complex interactions between input features.
- **Batch normalization and dropout:** Batch normalization is applied to stabilize and accelerate training, while dropout prevents overfitting by randomly setting a fraction of inputs to zero.
- **Sigmoid activation:** The final layer uses a sigmoid activation function to predict probabilities for binary x and y values.

The output of the network is split into two parts: one for predicting UE-cluster assignments (x) and another for cluster-BS assignments (y).

C. Ensemble Learning

An ensemble of models is trained to improve prediction accuracy and robustness. The ensemble training routine builds multiple instances of the same model and trains them independently:

- **Model diversity:** By training 10 different models, the ensemble technique helps reduce overfitting and increases the model's ability to generalize to unseen data.
- **Averaging predictions:** During inference, the predictions from all models in the ensemble are averaged to produce the final binary outputs for x and y .

D. Prediction and Post-processing

After training, the model predicts x and y values on test samples:

- **Binary conversion:** The predicted outputs from the sigmoid function are converted to binary values using rounding, ensuring that x and y remain discrete as required by the problem.
- **Saving predicted results:** The predictions, along with relevant data such as predicted bit rates and objective values, are saved in JSON format for analysis.

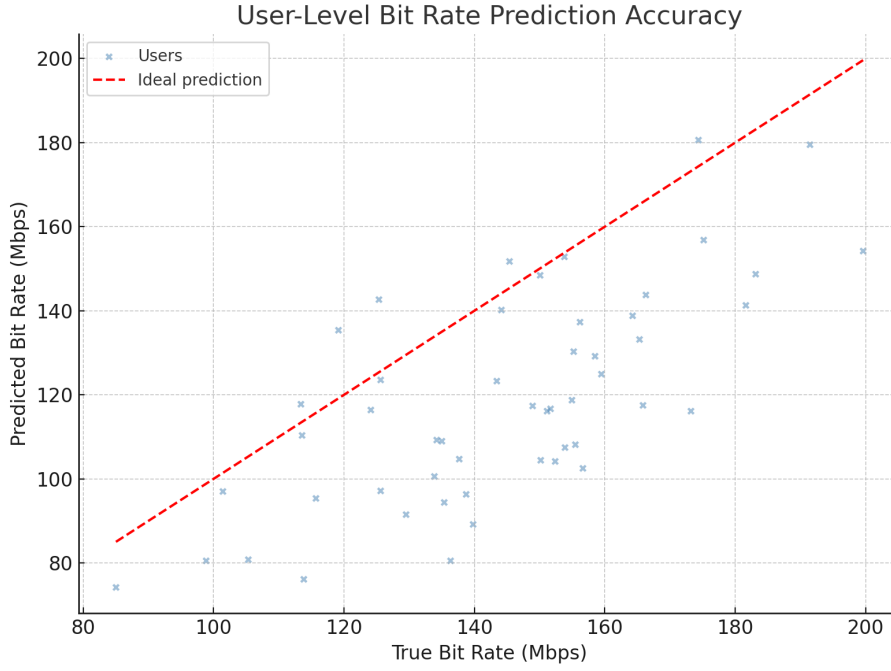


Fig. 6. User-Level bit rate prediction accuracy. Each point compares predicted vs. actual UE bit rate. The dashed red line indicates perfect prediction.

TABLE II
DEEP LEARNING MODEL PERFORMANCE EVALUATION.

Metric	Value
Cluster membership accuracy (X)	71.00%
BS assignment accuracy (Y)	79.88%
Mean absolute SE error (bit/sec/Hz)	0.85
Mean squared SE error (bit/sec/Hz) ²	2.91

and an average absolute SE error below 0.9 bit/sec/Hz, indicating that the model's predictions remain practically close to the optimization-generated ground truth even without post-processing or constraint projection.

E. Model Performance Evaluation

To evaluate the prediction performance of our deep learning model, we conduct a comprehensive analysis comparing the predicted and actual user-level bit rates on the test set. This evaluation comprises:

- **User-level bit rate prediction accuracy:** Scatter plot analysis comparing predicted and real values.
- **Statistical error metrics:** Including mean absolute error (MAE) and mean squared error (MSE).
- **Classification accuracy:** Evaluating how accurately the model predicts cluster membership (X) and BS assignment (Y).

Fig. 6 visualizes the relationship between predicted and actual UE bit rates across the entire test dataset. Each point represents one UE's result. The dashed red line corresponds to the ideal case where the predicted value exactly matches the real one. As shown, most data points cluster closely around the identity line, confirming that the model maintains a high prediction fidelity. The slight variance from the diagonal highlights some inherent complexity in the nonlinear structure of the optimization problem, but the general trend confirms the model's reliability.

Furthermore, Table II summarizes the key metrics. The model achieves 71% accuracy for cluster membership prediction (X), 79.88% accuracy for BS-subcarrier assignment (Y),

V. CONCLUSION

In this paper, we have presented a novel optimization clustering-based problem and deep-learning-based approach for high-precision RRM in the context of mmWave and THz communication, specifically targeting MU-mMIMO scenarios. Our proposed approach leverages MU-mMIMO, UC-CF mMIMO, efficient UE clustering, and optimized resource allocation to improve the accuracy and computational efficiency.

We provide a comprehensive evaluation of our proposed method and compare its performance with traditional algorithms such as NOMA and OFDM. The results of our evaluation demonstrate that our optimization approach outperforms traditional methods in terms of total SE and fairness while maintaining a manageable computational complexity using the alternative deep learning model.

Overall, our research contributes to the advancement of wireless communication by providing a deep-learning-based solution for precise resource allocation in MU-mMIMO scenarios. The proposed approach offers a balance between computational complexity and accuracy, making it suitable for practical implementation in next-generation wireless systems. We believe that our findings will pave the way for further advancements in resource allocation techniques and contribute to the development of efficient and reliable mmWave and THz communication systems.

REFERENCES

- [1] E. C. Strinati *et al.*, “6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication,” *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, 2019.
- [2] P. Zhou *et al.*, “Deep learning-based beam management and interference coordination in dense mmWave networks,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592–603, 2018.
- [3] M. Xiao *et al.*, “Millimeter wave communications for future mobile networks,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, 2017.
- [4] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, “Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2017.
- [5] M. Belgiovine, K. Sankhe, C. Bocanegra, D. Roy, and K. R. Chowdhury, “Deep learning at the edge for channel estimation in beyond-5G massive MIMO,” *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 19–25, 2021.
- [6] A. Farzamia, N. W. Hlaing, M. K. Haldar, and J. Rahebi, “Channel estimation for sparse channel OFDM systems using least square and minimum mean square error techniques,” in *Proc. ICET*, 2017.
- [7] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, “Deep learning-based channel estimation,” *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 652–655, 2019.
- [8] Q. Peng, J. Li, and H. Shi, “Deep learning based channel estimation for OFDM systems with doubly selective channel,” *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2067–2071, 2022.
- [9] Q. Bai, J. Wang, Y. Zhang, and J. Song, “Deep learning-based channel estimation algorithm over time selective fading channels,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 125–134, 2019.
- [10] B. Xiang, D. Hu, and J. Wu, “Deep learning-based downlink channel estimation for FDD massive MIMO systems,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 699–702, 2023.
- [11] C.-J. Chun, J.-M. Kang, and I.-M. Kim, “Deep learning-based joint pilot design and channel estimation for multiuser MIMO channels,” *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1999–2003, 2019.
- [12] S. Lyu, X. Li, T. Fan, J. Liu, and M. Shi, “Deep learning for fast channel estimation in millimeter-wave MIMO systems,” *J. Syst. Eng. Electron.*, vol. 33, no. 6, pp. 1088–1095, 2022.
- [13] K. Kim, Y. K. Tun, M. S. Munir, W. Saad, and C. S. Hong, “Deep reinforcement learning for channel estimation in RIS-aided wireless networks,” *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 2053–2057, 2023.
- [14] H. Hirose, T. Ohtsuki, and G. Gui, “Deep learning-based channel estimation for massive MIMO systems with pilot contamination,” *IEEE Open J. Veh. Technol.*, vol. 2, pp. 67–77, 2020.
- [15] S. Buzzi, C. D’Andrea, A. Zappone, and C. D’Elia, “User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, 2019.
- [16] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, “User-centric cell-free massive MIMO networks: A survey of opportunities, challenges and solutions,” *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 611–652, 2021.
- [17] F. Göttsch, N. Osawa, I. Kanno, T. Ohseki, and G. Caire, “Fairness scheduling in user-centric cell-free massive MIMO wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11942–11957, 2024.
- [18] M. Ito *et al.*, “Clustering and beamforming for user-centric cell-free massive MIMO with backhaul capacity limitation,” *IEEE Access*, vol. 12, pp. 382–395, 2023.
- [19] Y. Tan, L. Zhao, and R. Zhang, “Hierarchical deep reinforcement learning for access point clustering and power control in cell-free massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 684–699, 2023.
- [20] Y. An, C. Wang, and L. Zhang, “Soft actor-critic based resource scheduling in massive MIMO systems,” *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1231–1245, 2023.
- [21] A. Sher, A. Khan, and S. Ahmed, “Joint user association and resource allocation for heterogeneous networks using deep learning,” *IEEE Access*, vol. 11, pp. 50112–50125, 2023.
- [22] T. Luo, Y. Xu, and J. Huang, “Semi-supervised learning for resource allocation in NOMA-enabled networks,” *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 5862–5875, 2022.
- [23] H. Cui, X. Fang, and Z. Huang, “User clustering and power allocation for mmwave-NOMA using K-means and convex optimization,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4985–4996, 2023.
- [24] W. Zhang, X. Gao, and S. Li, “Joint pilot assignment and AP clustering for cell-free massive MIMO: A K-means-based method,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 8, pp. 1372–1376, 2023.
- [25] S. Lim, J. Kim, and I. Lee, “Joint user clustering, beamforming, and power allocation for mmwave-NOMA with imperfect SIC,” *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1601–1615, 2023.
- [26] H. Humadi, A. Uwaechia, and B. Adebisi, “User-centric cluster design and analysis for hybrid sub-6 GHz–mmWave–THz dense networks,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9807–9819, 2022.
- [27] H. Wang, P. Xiao, and X. Li, “Channel parameter estimation of mmWave MIMO system in urban traffic scene: A training channel-based method,” *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 754–762, 2022.
- [28] K. Guo *et al.*, “Physical layer security for multiuser satellite communication systems with threshold-based scheduling scheme,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13471–13485, 2020.
- [29] L. Dai *et al.*, “Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, 2015.
- [30] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, “Massive MIMO networks: Spectral, energy, and hardware efficiency,” *Foundations Trends® Signal Processing*, vol. 11, no. 3–4, pp. 154–655, 2017.
- [31] A. Alkhateeb, “DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications,” 2019 *arXiv:1902.06435*.
- [32] M. Shahjalal, M. H. Rahman, M. O. Ali, B. Chung, and Y. M. Jang, “User clustering techniques for massive MIMO-NOMA enabled mmWave/THz communications in 6G,” in *Proc. ICUFN*, 2021.



Pooya Hejazi is a Ph.D. candidate in Computer Networks at Amirkabir University of Technology (AUT), Tehran, Iran. He earned his B.Sc. degree in Software Engineering from Shahid Beheshti University of Technology, Tehran, in 2006, and his M.Sc. degree in Computer Networks from AUT in 2009. His research focuses on cutting-edge topics such as 6th generation (6G) cellular networks, machine and deep learning, wireless sensor networks (WSNs), and optimization of radio resource management (RRM) in wireless and IoT networks. Additionally, Pooya Hejazi has extensive experience in the security domain, with research on red teaming, next generation firewalls, and advanced web application firewalls. His multidisciplinary approach combines network optimization with security concerns, aiming to enhance the performance and robustness of modern communication systems.



Masoud Sabaei received his B.Sc. degree in Computer Engineering from Esfahan University of Technology, Esfahan, Iran, in 1992, and his M.Sc. and Ph.D. degrees in Computer Engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 1995 and 2000, respectively. He has been a Professor in the Department of Computer Engineering at Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, since 2002. His research interests include software-defined networking, the Internet of things, wireless networks, and telecommunication network management.