Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Efficient User Clustering Framework for Non-Orthogonal Multiple Access Systems

Kanchana Katta[1*], Aditi Mishra[2], Navanath Saharia[3], Ramesh Ch Mishra[4]
[1]Department of Electronics and Communication Engineering, IIIT Manipur, India*
[2]Department of Electronics and Communication Engineering, IIIT Manipur, India
[3]Department of Computer Science and Engineering, IIIT Manipur, India.
[4]Department of Electronics and Communication Engineering, IIIT Manipur, India
Corresponding author: kanchana@iiitmanipur.ac.in

ARTICLE INFO

Received: 25 Sept 2025

Revised: 30 Sept 2025

Accepted: 15 Oct 2025

ABSTRACT

Introduction: This work introduces an efficient user clustering framework for Non-orthogonal multiple access (NOMA) systems along with a structured, high-resolution Bluetooth-based dataset to facilitate advanced research in next generation wireless communication networks. As modern NOMA deployments increasingly rely on adaptive user grouping and interference-aware pairing, the lack of publicly available real world data has posed a significant limitation to empirical validation. In this paper, we focus on the capturing real-world signal characteristics in realistic indoor and urban environments, the dataset includes device-level metrics such as received signal strength indicator (RSSI), estimated distance, device type, and detection frequency parameters essential for modeling dynamic user heterogeneity and clustering behavior. Also, this paper presents a detailed classification of user clustering techniques specifically designed for NOMA-based wireless communication systems. Using these clustering schemes, we conducted extensive simulations to evaluate their performance in NOMA-based systems. The proposed clustering algorithm consistently outperforms existing techniques in terms of sum rate and energy efficiency (EE). The dataset serves as a critical asset for bridging theoretical models with field deployable NOMA architectures. It is particularly relevant for 5G and emerging 6G research, where intelligent access schemes must operate reliably under dense device populations and rapidly varying channels.

Keywords: *Non-orthogonal multiple access, Dataset, User clustering, Received signal strength indicator, Balanced K-means*

1. INTRODUCTION

The exponential growth of mobile users, IoT devices, and latency-sensitive applications has pushed wireless communication systems to adopt more spectrum efficient technologies. NOMA has emerged as a key enabler in this evolution, allowing multiple users to simultaneously access the same time-frequency resources through superposition coding and successive interference cancellation (SIC) [1, 2]. User clustering and pairing are fundamental to maximizing NOMA performance. Pairing users with distinct channel gains can significantly improve both spectral efficiency and user fairness [3-5].

Traditionally, static user grouping and heuristic based pairing have been employed, but they struggle to adapt under dynamic environments [6-8]. In recent years, deep learning and data-driven optimization have attracted significant attention in wireless systems [9-11]. These methods aim to model complex, non-linear relationships in real time and are highly effective in adapting to diverse user and channel profiles [12-14]. However, most existing models rely heavily on simulation generated data, which fails to reflect the stochastic nature of real-world deployments [15, 16].

To address this critical gap, our study presents a Bluetooth-based dataset collected across realistic settings involving thousands of unique devices. The dataset provides extensive variation in signal quality, proximity, and temporal characteristics, suitable for evaluating NOMA user clustering and pairing schemes. These features are crucial for developing robust interference management and power allocation strategies required in dense NG environments. Furthermore, proximity aware and context aware datasets enable real-time modeling of heterogeneous user groups, helping network designers validate algorithms under practical constraints. This allows for more accurate assessment of fairness, latency, and energy trade-offs, which are central to future intelligent access schemes.

1.1. LITERATURE REVIEW

Study during the experiment reflects extensive progress in user clustering for NOMA systems using synthetic or simulated data, however real-world validation remains scarce. Table 1 consolidates key references, highlighting their dataset types, clustering methods, system focus, major findings, and limitations. This comparison underscores the novelty of our proposed framework using balanced K-Means algorithm applied on a structured Bluetooth dataset, bridging the gap between theoretical modeling and practical system deployment.

Table 1. Comparison of Existing NOMA Clustering Studies, Dataset Types, and Limitations

Reference	Dataset	Clustering Method	System Focus	Key Findings	Limitations
Dileepa et al. [17]	Simulated mmWave channel / users	Hierarchical clustering	Downlink mmWave-NOMA	Maximizes sum-rate while satisfying QoS without pre-specified cluster count	Synthetic simulation only; no real-world dataset
Santos YP et al. [18]	Simulated / dynamic user variations	DenStream-based adaptive clustering	Dynamic NOMA (user arrivals/departures)	Adapts cluster membership; ~10% gains over OMA	Simulation only; complexity & real-data validation missing
Celik A et al. [19]	Synthetic / modeled user signals	Partition-based clustering	Grant-Free NOMA / IoT	Clustering & power control in milliseconds; close to optimal sum-rate	Synthetic model; no Bluetooth dataset
R. Kim et al. [20]	Channel gains (simulated)	Joint clustering + beamforming (greedy)	Downlink NOMA + beamforming	Gains by jointly optimizing clustering & beamforming	Simulation-based; clustering not Bluetooth-featured; scaling issues
Kumaresan et al. [12]	Simulated channel/ resource allocation	Deep neural network clustering	Fixed-power downlink NOMA	DNN-based clustering improves throughput vs heuristic	Simulation only; lacks real-world validation

1.2. MOTIVATION AND CONTRIBUTIONS

Based on the discussions in the introduction and literature review, it is evident that significant progress has been made in the development of user clustering strategies for NOMA systems. Existing works employing K-Means, DBSCAN, hierarchical clustering, and other heuristics have demonstrated performance improvements in terms of spectral efficiency and fairness under controlled conditions. However, the majority of these studies are built upon synthetic or randomly generated datasets, which lack the diversity and realism of actual wireless environments.

Although a few attempts have been made to use Bluetooth or IoT traces, these datasets are generally limited in scale, unstructured, and not specifically tailored for NOMA research. Moreover, conventional clustering approaches often suffer from imbalanced group formation, where near users dominate cluster assignments while far users are underrepresented, leading to reduced fairness and degraded energy efficiency. These limitations highlight a clear gap between theoretical research and practical deployment. To address this gap, there is a strong need for a real-world, Bluetooth-based dataset combined with a balanced clustering algorithm that ensures fairness and improves system-level performance metrics. Such an approach is particularly vital for 5G and emerging 6G networks, where intelligent, data-driven user access schemes must operate effectively under dense connectivity and rapidly varying channel conditions.

The major contributions of this work are listed below.

1. We introduce a structured, high-resolution Bluetooth dataset specifically designed to capture real-world device heterogeneity and user interaction patterns in indoor and urban wireless environments. Unlike synthetic or random datasets, this dataset reflects practical propagation conditions suitable for empirical validation of NOMA systems.
2. The dataset includes critical features such as RSSI, estimated distance, device type, and detection frequency, enabling realistic modeling of user heterogeneity, mobility, and interference, aspects often overlooked in existing studies.
3. We propose a novel user clustering framework in NOMA using Balanced K-Means approach that leverages device-level Bluetooth features to achieve fair and optimized user grouping, improving interference management compared to conventional clustering techniques.
4. Using the developed dataset, we evaluate multiple clustering strategies under NOMA scenarios and demonstrate significant improvements in sum rate and energy efficiency bridging the gap between theoretical modeling and practical system deployment.

The remaining paper is organized as follows: Section 2 presents the system model, detailing the NOMA framework and proposed clustering algorithm.. Section 3 discusses the dataset construction and dataset features. Section 4 demonstrates the simulation setup, including performance evaluation metrics and comparative results. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2. SYSTEM MODEL

We considered a downlink scenario in a NOMA system, where a single base station (BS) transmits to M users simultaneously within the same time-frequency resource. NOMA achieves spectrum efficiency by exploiting the power domain for multiple access, using superposition coding at the transmitter and SIC at the receivers. These users are grouped into distinct clusters to enable user-specific resource allocation and interference mitigation. These users are distributed into K clusters, where each cluster contains multiple users who are geographically close or exhibit similar channel characteristics which is shown in Figure 1. This clustering approach enables more efficient power-domain multiplexing by leveraging user diversity.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

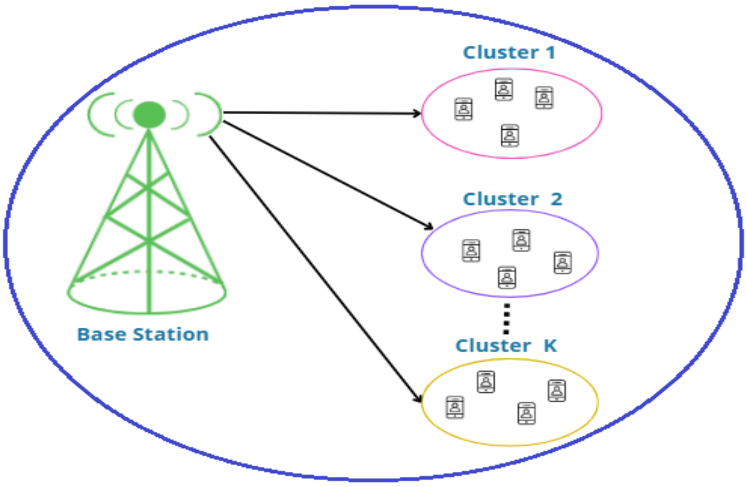

Figure 1. System model for NOMA user clustering

The transmitted signal from the BS is a superposition of individual user signals, expressed as:

$$x = \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sqrt{P_{k,m}}\, s_{k,m} \qquad (1)$$

where $s_{k,m}$ represents the unit-power symbol intended for the m-th user in the k-th cluster, $P_{k,m}$ is the corresponding allocated power, and M_k denotes the number of users in cluster k. The total transmit power from the BS is constrained by $\sum_{k,m} P_{k,m} \le P_{total}$. This superimposed transmission is received by each user in the presence of channel gain and noise.

The received signal at user m in cluster k is:

$$y_{k,m} = h_{k,m}x + n_{k,m} \qquad (2)$$

where $h_{k,m}$ denotes the complex channel gain between the BS and the user (assumed Rayleigh distributed), while $n_{k,m}$ is the additive white Gaussian noise (AWGN) with variance σ^2. Within each cluster, users are sorted by their channel conditions such that $|h_{k,1}|^2 \le |h_{k,2}|^2 \le \cdots \le |h_{k,M_k}|^2$, allowing for the use of SIC. The user with the weakest channel gain receives the highest power and decodes its message directly, while stronger users decode weaker users messages first, subtract them, and then decode their own.

The signal-to-interference-plus-noise ratio (SINR) for the m-th user in the k-th cluster is calculated as:

$$\gamma_{k,m} = \frac{P_{k,m}|h_{k,m}|^2}{\sum_{j=m+1}^{M_k} P_{k,j}|h_{k,m}|^2 + \sigma^2} \qquad (3)$$

This SINR formulation accounts for the fact that each user can cancel interference from users with weaker channels but must treat stronger users' signals as noise.

The achievable data rate for user m in cluster k is given by:

$$R_{k,m} = B\log_2(1+\gamma_{k,m}) \qquad (4)$$

where B is the available system bandwidth.

2.1. PROPOSED APPROACH

User clustering plays a crucial role in NOMA systems, where multiple users are grouped into clusters and served simultaneously on the same frequency-time resource. The performance of NOMA strongly depends on how users are paired or clustered, since SIC relies on sufficient differences in users' channel conditions. A good clustering strategy ensures that strong and weak users are grouped together, thereby improving, sum rate, spectrum efficiency, fairness, and stability of SIC decoding. Several clustering techniques have been widely studied in NOMA research, these methods while effective in certain contexts, do not explicitly enforce balanced grouping between near and far users, which is essential for practical NOMA performance. In this paper we propose a Balanced K-Means algorithm that explicitly enforces equal cluster sizes.

Algorithm 1: Balanced K-Means Clustering

```
Input: $X = \{x_1, x_{21}, \ldots, x_M\}$ - user feature set, K - number of clusters, M/K - balanced cluster size
Output: Cluster assignments $C = \{c_1, c_2, \ldots, c_K\}$
Initialize K centroids randomly from dataset X
repeat
    Compute distance d($x_i$, $c_j$) between each user $x_i$ and each centroid $c_j$
    Construct cost matrix M with distances and capacity M/K per cluster
    Solve balanced assignment using Hungarian algorithm (or min-cost flow)
    Update cluster assignments $c_j$ based on optimal assignment
    Recompute centroids $c_j$ = mean of users in cluster j
until convergence (no centroid change or max iterations reached)
return C
```

To manage the complexity of decoding and resource allocation in a large scale system, users are grouped into K clusters based on their spatial proximity and signal characteristics such as RSSI, estimated distance, and channel gain.

User clustering is optimized using two complementary objective functions. First, the clustering assignment C is determined to maximize total sum-rate:

$$\max_{C} \sum_{k=1}^{K} \sum_{m=1}^{M_k} R_{k,m} \qquad (5)$$

Second, to ensure user similarity within each cluster, a distance-based clustering loss is minimized:

$$\min_{C} \sum_{k=1}^{K} \sum_{m \in C_k} | u_m - \mu_k |^2 \qquad (6)$$

where u_m is the feature vector of the m-th user, and μ_k is the centroid of cluster k.

3. DATASET CONSTRUCTION

In the context of NOMA-based user clustering and pairing in wireless communication networks, dataset collection plays a crucial role in enabling the training and evaluation of intelligent algorithms such as machine learning and deep learning models. However, due to the absence of standardized real-world datasets in this domain, researchers typically rely on synthetically generated small datasets. In this paper, we focus on the capturing real-world signal characteristics in realistic indoor and urban environments.

The developed dataset comprises of 16 parameters associated with identity, signal profile and environmental profiles of the scanned devices. The developed tool detects bluetooth-enabled devices in a defined intervals. Each scan logs timestamps, device identifiers, signal characteristics, and estimated spatial relationships. The implementation differentiates between classic Bluetooth and BLE devices while maintaining a unified data structure. Signal quality metrics are derived from RSSI values, with distance estimations calculated using environment-specific path loss models. For mobile devices, the system additionally estimates cellular network parameters including RSRP and RSRQ values. Device persistence is tracked through detection counts and first/last seen timestamps, enabling movement and presence analysis. The used parameters along with definitions are discussed below.

Dataset Features:

a) **Scan Time**: The timestamp when a device was detected, formatted as *YYYY-MM-DD HH:MM:SS*.

b) **Device Name**: The advertised name of the device, if available.

c) **MAC Address**: A unique identifier (hardware address) of the device, formatted as six hexadecimal pairs separated by colons.

d) **Device Type**: Specifies the kind of device, such as Mobile Phone, Laptop, or BLE Device.

e) **RSSI**: Received Signal Strength Indicator in decibels per milliwatt (dBm), ranging from > -60 dBm (strong) to < -80 dBm (very weak).

f) **Signal Quality**: Categorized signal strength: Strong, Moderate, or Weak, based on RSSI thresholds.

g) **Estimated Distance**: Estimated distance between base station and device, derived from RSSI values and environmental factors (ranging from 0.1m to 47.9m).

h) **SNR**: Signal-to-noise ratio (in dB), indicating the clarity of the signal in a noisy environment.

i) **Network Technology**: Network technology used by the device, such as LTE.

j) **RSRP**: Reference Signal Received Power, derived from RSSI, indicating cellular signal strength.

k) **RSRQ**: Reference Signal Received Quality, estimated from RSSI tiers to measure interference.

l) **Major Device Class**: General classification of the device, such as Phone.

m) **Service UUIDs**: Unique identifiers for services offered by the device.

n) **Detection Count**: Number of times the same device was detected.

o) **First Seen**: Timestamp of first detection.

p) **Last Seen**: Timestamp of most recent detection.

A summary of the key characteristics of the dataset is provided in Table 2. The dataset was collected through continuous scanning in a urban environment. A total of 93216 device entries were recorded across multiple scan sessions, representing approximately 9740 unique devices based on MAC addresses. The majority of detected devices (96.4%) were BLE devices, with only 3384 devices of Classic Bluetooth type identified. Signal strengths varied considerably, with 3.7 % showing strong signals (-60 dBm), 17.5 % moderate signals (-60 to -80 dBm), and 78.9% weak signals (-80 dBm). The estimated proximity ranged from very close (0.1m) to distant (50m), with a median distance of approximately 22.9m. Detection persistence varied significantly, with one device detected 1439 times while 5.1% of devices were detected only once.

Table 2. Statistics of the collected NOMA user dataset

Metric	Value
Total device entries	93,216
Unique devices	9,740
BLE devices	89,832
Classic Bluetooth devices	3,384
Strong signal devices	3,438
Moderate signal devices	16,272
Weak signal devices	73,506
Closest proximity	0.1m
Furthest detection	50.0m
Highest detection count	1,439

4. RESULTS AND DISCUSSIONS

This section presents the performance evaluation of different clustering methods for downlink NOMA using the proposed Bluetooth-based dataset. Unlike purely synthetic simulations, the dataset captures realistic RSSI and distance information, enabling a more accurate modeling of user heterogeneity in indoor and urban environments. The dataset was processed in MATLAB, where user channels were derived from the distance and RSSI features combined with Rayleigh fading and a path-loss exponent.

Figure 2 illustrates the sum-rate performance of different clustering methods Balanced K-Means, Hierarchical, DBSCAN, and standard K-Means—under varying SNR conditions from -10 dB to 20 dB. The Balanced K-Means algorithm consistently achieves the highest sum rate across all SNR levels due to its ability to maintain balanced clusters containing both near and far users. This balanced clustering facilitates efficient power allocation and improves the accuracy of successive interference cancellation (SIC), resulting in superior throughput. In contrast, Hierarchical and DBSCAN clustering methods show intermediate performance, as they capture user similarity but do not explicitly enforce balanced cluster sizes. Standard K-Means exhibits the lowest sum rate at lower SNRs, primarily due to uneven clustering that leads to suboptimal near-far user pairing. Overall, these results highlight the critical importance of user clustering strategies in enhancing the spectral efficiency of NOMA systems.

Figure 3 depicts the energy efficiency (EE) of the four clustering methods as a function of SNR. EE is computed as the ratio of total system throughput to total power consumption, including both transmit and circuit power. The Balanced K-Means algorithm achieves the highest energy efficiency across almost the entire SNR range, with a noticeable peak at mid-range SNR levels. This peak occurs because the algorithm achieves high throughput without excessive power consumption. In comparison, Hierarchical, DBSCAN, and standard K-Means show gradually increasing EE curves, but they remain consistently lower than Balanced K-Means. At very high SNR, the EE growth begins to saturate or slightly decline, reflecting the dominance of transmit power over throughput gains. These observations confirm that enforcing balanced clusters significantly improves energy utilization in NOMA systems, making Balanced K-Means a promising approach for energy-efficient wireless communications.

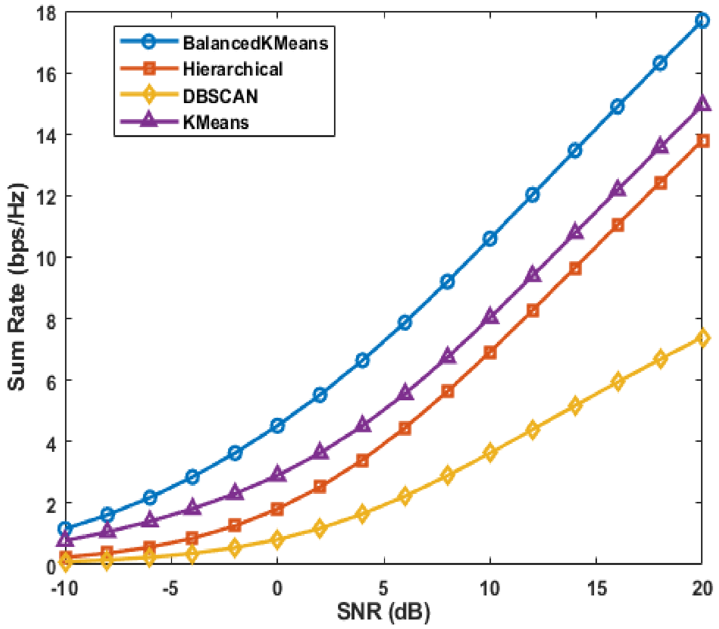

Figure 2. Sum rate performance using user clustering schemes

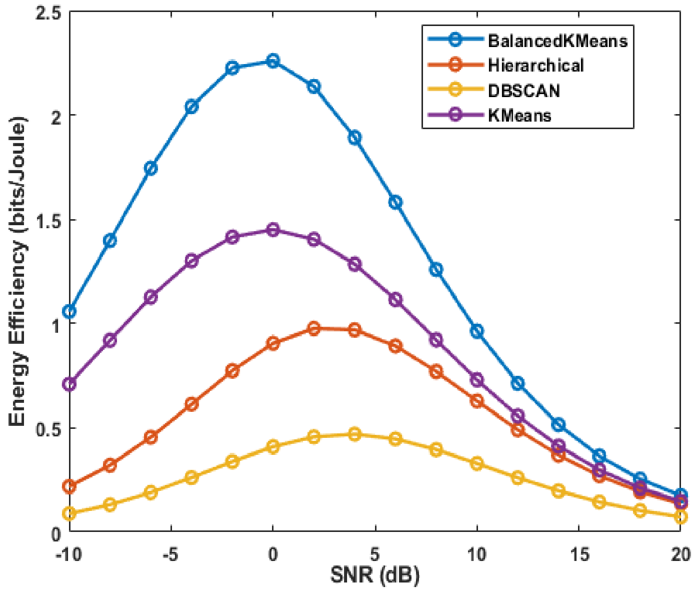

Figure 3. Energy efficiency Vs SNR using different clustering schemes

5. CONCLUSION

This work introduced a structured Bluetooth-based dataset designed for advancing NOMA research in realistic wireless environments. By capturing device-level metrics such as RSSI, distance, and detection frequency, the dataset enables effective modeling of user heterogeneity and clustering behavior. A classification of clustering techniques was presented, and simulation results confirmed that the proposed clustering algorithm achieves superior performance in terms of sum rate and energy efficiency compared to existing methods. For future work, this dataset and framework can be extended by integrating deep learning-driven clustering strategies, incorporating IRS-assisted NOMA, and evaluating large scale multi user scenarios under practical 6G network conditions.

REFERENCES

[1] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," IEEE Communications Magazine, vol. 53, no. 9, pp. 74-81, September 2015.

[3] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), Dresden, Germany, 2013, pp. 1-5.

[4] Z. Zhang, Y. Cheng, L. Yang, B. Jiao, and X. Shen, "User Pairing and Power Allocation for Downlink NOMA Systems," IEEE Transactions on Wireless Communications, vol. 15, no. 12, pp. 8287-8300, Dec. 2016.

[5] K. Katta, R. C. Mishra and N. Saharia, "Energy Efficient Intelligent Reflecting Surface Assisted Downlink Non-Orthogonal Multiple Access System," 2025 IEEE Guwahati Subsection Conference (GCON), Itanagar, India, 2025, pp. 1-5.

[6] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 721-742, third quarter 2017.

[7] Y. Liu, Y. Fu, R. Schober, and D. W. K. Ng, "Adaptive User Clustering and Power Allocation for NOMA Systems," IEEE Access, vol. 6, pp. 70130-70144, 2018.

[8] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal User Pairing and Power Allocation for Downlink NOMA Systems," IEEE Transactions on Communications, vol. 66, no. 12, pp. 5645-5657, Dec. 2018.

[9] Q. Zhang, Y. Xu, S. Gong, L. Dai, and Z. Wang, "NOMA Meets Machine Learning: Opportunities and Challenges," IEEE Communications Magazine, vol. 58, no. 7, pp. 56-61, July 2020.

[10] A. Dejonghe, C. Antón-Haro, X. Mestre, L. Cardoso and C. Goursaud, "Deep Learning-Based User Clustering For Mimo-Noma Networks," 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, 2021, pp. 1-6.

[11] Fa-Long Luo, "Machine Learning for Optimal Resource Allocation," in Machine Learning for Future Wireless Communications , IEEE, 2020, pp.85-103, doi: 10.1002/9781119562306.ch5.

[12] Kumaresan, S.P.; Tan, C.K.; Ng, Y.H. Deep Neural Network (DNN) for Efficient User Clustering and Power Allocation in Downlink Non-Orthogonal Multiple Access (NOMA) 5G Networks. Symmetry 2021, 13, 1507.

[13] S. Kiani, M. Dong, S. ShahbazPanahi, G. Boudreau and M. Bavand, "Learning-Based User Clustering in NOMA-Aided MIMO Networks With Spatially Correlated Channels," in IEEE Transactions on Communications, vol. 70, no. 7, pp. 4807-4821, July 2022.

[14] K. Sayrafian, B. Cloteaux, V. Marbukh and C. Emiyah, "Evaluation of the Bluetooth-based Proximity Estimation for Automatic Exposure Determination," 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2022, pp. 683-686.

[15] Y. Yuan et al., "NOMA for Next-Generation Massive IoT: Performance Potential and Technology Directions," in IEEE Communications Magazine, vol. 59, no. 7, pp. 115-121, July 2021

[16] Umar Ghafoor, Mudassar Ali, Humayun Zubair Khan, Adil Masood Siddiqui, Muhammad Naeem, NOMA and future 5G & B5G wireless networks: A paradigm, Journal of Network and Computer Applications, Volume 204,2022

[17] Marasinghe, Dileepa et al. “Hierarchical User Clustering for mmWave-NOMA Systems.” 2020 2nd 6G Wireless Summit (6G SUMMIT) (2020): 1-5.

[18] Santos YP, Silveira LFQ. “Adaptive Clustering of Users in Power Domain NOMA.” Sensors (Basel). 2023 Jun 3;23(11).

[19] Celik A. Grant-Free NOMA: A Low-Complexity Power Control through User Clustering. Sensors (Basel). 2023 Oct 4;23(19).

[20] H. R. Kim, J. Chen and J. Yoon, "Joint User Clustering and Beamforming in Non-Orthogonal Multiple Access Networks," in IEEE Access, vol. 8, pp. 111355-111367, 2020.