Transformer-Based User Clustering for Efficient Downlink NOMA System

Original Scientific Paper

Kanchana Katta*
Indian Institute of Information Technology Senapati, Manipur
Department of Electronics and Communication Engineering
Imphal, India
kanchana@iiitmanipur.ac.in

Ramesh Chandra Mishra
Indian Institute of Information Technology Senapati, Manipur
Department of Electronics and Communication Engineering
Imphal, India
m.ramesh@iiitmanipur.ac.in

Navanath Saharia
Indian Institute of Information Technology Senapati, Manipur
Department of Computer Science and Engineering
Imphal, India
nsaharia@iiitmanipur.ac.in

*Corresponding author

Abstract *– As a result of the growing requirement for intelligent and adaptive resource allocation in future wireless networks, the growing interest in next-generation (NG) wireless networks has promoted the use of sophisticated user clustering methods within non-orthogonal multiple access (NOMA) systems. This paper proposes a novel deep learning framework based on a Transformer encoder for efficient user clustering and pairing in downlink NOMA. Instead of relying on text-based tokenization, the numerical channel state information (CSI) is mapped into dense feature embeddings, which are processed through multi-head self-attention to learn fine-grained inter-user relationships. This enables the model to capture interference patterns and contextual channel dependencies that conventional clustering approaches cannot represent. Using user distance, channel gain, SINR, and power allocation, we generated a synthetic dataset that meets the requirements of 3GPP TR 38.901 for use in evaluating performance in real-world fading conditions. We compared the performance based on a Transformer encoder approach with standard clustering methods (K-means, Balanced K-means, DBSCAN). The simulation results indicate that the proposed Transformer-based user clustering framework consistently outperformed all other clustering methods with respect to the key performance indicators of bit error rate (BER), throughput, user fairness, energy efficiency, and outage probability. For each of the SNR regimes, we achieved lower BERs, greater potential rate, better fairness indices, and less outage than the other clustering approaches. These results highlight the strong potential of Transformer-based architectures as scalable and intelligent solutions for NOMA user clustering and resource optimization in emerging 6G wireless networks.*

Keywords: *non-orthogonal multiple access, deep learning, K-means, balanced K-means, transformer encoder, successive interference cancellation*

Received: January 14, 2026; Received in revised form: March 16, 2026; Accepted: March 16, 2026

1. INTRODUCTION

The rapidly evolving nature of next-generation wireless network technologies requires the development of advanced technologies that support large numbers of devices simultaneously connecting, enhance the available spectrum utilization, ensure user fairness, and allocate resources based on intelligent systems utilized by the network users [1]. Power-domain non-orthogonal multiple access methods are being explored as a po-

tential solution to provide these capabilities. In power-domain NOMA, multiple users can simultaneously share the same time or frequency resources through superposition coding and power-domain multiplexing techniques [2-4]. In downlink NOMA, users with weaker channel conditions are allocated higher transmit power, while users with stronger channels are allocated lower transmit power. Stronger users employ successive interference cancellation (SIC) to decode and subtract the high-power interfering signals before retrieving their own data [5, 6]. This hierarchical decoding mechanism effectively mitigates intra-cluster interference and significantly enhances spectral efficiency and connectivity across heterogeneous environments [7, 8].

The application of NOMA has expanded beyond physical-layer performance enhancement to support emerging network architectures and services, including mobile edge computing (MEC) and integrated communication–computation frameworks [9]. These studies highlight that efficient user grouping, effective interference management, and adaptive resource allocation are fundamental to realizing the full potential of NOMA in complex and heterogeneous network environments. To ensure user fairness in NOMA-enabled systems, recent studies have formulated max–min optimization problems [10] that aim to maximize the minimum task computation or service rate among users. These approaches typically involve the joint optimization of offloading decisions, NOMA decoding order, transmission power allocation, and time resource scheduling, leading to challenging mixed-integer and nonlinear optimization problems. Despite these advantages, the performance of NOMA is strongly influenced by how users are clustered and paired for resource sharing. Poorly selected user groups can result in degraded throughput, unstable SIC operation, fairness imbalance and increased outage probability [11].

1.1. RELATED WORKS

Traditional pairing schemes such as fixed gain-difference pairing, distance-based grouping and random pairing lack adaptability to dynamic wireless environments and fail to capture multi-dimensional channel state information (CSI), limiting their effectiveness under mobility, fading and real-world user distribution conditions [12-14]. To improve adaptability, several clustering approaches have been investigated. Classical unsupervised clustering methods such as K-means, hierarchical clustering and DBSCAN group users based on statistical similarity in distance or channel gain. K-means provides simple partitioning but suffers from sensitivity to initialization and assumes spherical cluster shapes. Hierarchical clustering constructs a tree like structure of clusters by iteratively splitting user groups; it often incurs high computational complexity and may not scale efficiently with a large number of users. DBSCAN identifies clusters of arbitrary shapes but depends heavily on appropriate density thresholds and may be unstable in irregular topologies [15]. Non-parametric mean shift clustering automatically discovers cluster modes, yet its computational complexity increases rapidly with growing user numbers and high dimensional CSI, reducing its practicality for large-scale deployments [16, 17].

To overcome these limitations, recent works have incorporated deep learning (DL) into NOMA clustering and pairing. Convolutional neural network (CNN) based approaches capture spatial CSI features to guide clustering, while long short-term memory (LSTM) based models exploit temporal channel variations to support dynamic user grouping [18-21]. Although these approaches outperform classical clustering, their architectural constraints limit receptive fields in CNNs and sequential processing in LSTMs, which limits their ability to capture global multi user interactions, which are essential for optimal NOMA performance. This has led to growing interest in deep neural networks (DNNs) and graph neural networks (GNNs) for user association and resource allocation. In [22], a DNN-driven user clustering strategy for downlink NOMA was proposed, showing performance gains over heuristic methods, but their approach requires task-specific supervision and retraining under changing network conditions. However, the proposed method relies on task-specific supervision and requires retraining when network conditions change. Similarly, GNNs can significantly outperform traditional clustering methods for user grouping, beamforming and power allocation in hybrid NOMA networks. Recent research indicates that intelligent clustering techniques such as graph-based methods can overcome the limitations of traditional NOMA pairing techniques [23, 24]. Although graph-based models perform well for structured relational learning, the applicability of these models in rapidly evolving wireless environments is limited by reliance on predefined graph structures and high computational requirements.

Recently, transformer architectures have been successfully applied to capture global relationships among high-dimensional features in various applications. Transformers employ multi-head self-attention to capture relationships across multiple diverse CSI features simultaneously, enabling the modeling of complex feature interactions. In addition, transformers process users in parallel, unlike recurrent architectures, which rely on sequential processing, thereby avoiding delays caused by waiting for prior computations before processing subsequent users [25-27].

These characteristics of transformers are therefore suitable for use in wireless applications such as power allocation with attention-driven methods, resource scheduling, link adaptation, channel prediction and 6G holistic resource optimization [28]. Moreover, decision transformer frameworks introduce offline training with online generalization capabilities for base station scheduling and radio resource management. Surveys confirm that transformers are rapidly reshaping machine learning driven wireless optimization owing

to their robustness and generalization strengths [29, 30]. The Transformer models have also been applied to multi-antenna signal processing [31], beam selection [32], and channel estimation [33]. In a number of wireless learning tasks, these studies show that self-attention outperforms conventional convolutional or recurrent networks in modeling relationships in high-dimensional signal spaces.

However, despite these advancements, a critical gap remains, as no existing work provides a unified transformer based CSI embedding and user clustering framework specifically designed for downlink NOMA, integrating contextual CSI representations with adaptive clustering and intelligent user pairing. Most existing studies primarily focus on power allocation, channel prediction, or general resource optimization but do not exploit transformer embeddings for NOMA user grouping. Additionally, prior works seldom provide comprehensive performance evaluations that include BER, fairness, throughput, outage probability and energy efficiency within a single framework. Furthermore, they do not address interpretability via attention visualization, computational complexity and practical deployment feasibility.

To address these challenges, this work proposes a transformer encoder based CSI embedding and clustering framework. The proposed approach transforms raw CSI into contextual embeddings, captures global dependencies among users through a multi-head self-attention mechanism, and enables adaptive, similarity aware user pairing for NOMA systems. The proposed framework advances NOMA resource allocation while addressing long standing challenges in scalability, fairness, energy efficiency and interference management.

1.2. MOTIVATION AND CONTRIBUTIONS

From the existing literature, it is evident that traditional clustering algorithms and classical deep learning models are insufficient for capturing the dynamic, non linear and high dimensional characteristics of user channels in downlink NOMA systems. These limitations result in suboptimal user pairing, reduced SIC performance and degraded system throughput. To address these challenges, this paper proposes a deep learning framework for clustering downlink NOMA users that captures contextual relationships among channel features using window-based temporal representations. The key contributions of this study are detailed below.

- A transformer encoder based framework is proposed for downlink NOMA user clustering to encode multi-dimensional CSI into contextual embeddings. The proposed transformer-based embedding model is evaluated against conventional clustering algorithms, including K-means, DBSCAN and Balanced K-means. In contrast, the transformer encoder generates robust contextual embeddings that improve clustering stability, enhance interference awareness and capture complex inter-user relationships.
- A controlled dataset was created by simulating user distances using the 3GPP TR 38.901 propagation models. Each user is represented using realistic CSI parameters, which include distance to the site, path loss, propagation channel gain, transmitted power and signal to interference plus noise ratio (SINR). This approach provides reliable, reproducible conditions and allows for practical propagation modeling to evaluate clustering performance.
- The system performance was evaluated using key metrics such as bit error rate, throughput, energy efficiency, outage probability and fairness index. The results consistently show that the transformer encoder based system achieves superior performance and more stable SIC operation compared to traditional clustering techniques.

The rest of this article is organized as follows: Section 2 presents the system model. Section 3 discusses the proposed architecture and clustering methods. Section 4 discusses the simulation results. Lastly, section 5 concludes the article.

2. SYSTEM MODEL

The downlink NOMA system is considered, where a single base station (BS) serves a set of N users denoted by $U=\{u_1, u_2,..., u_N\}$ within a single cell. All users experience distance dependent path loss and Rayleigh fading. The BS utilizes power domain multiplexing to transmit data simultaneously to multiple users over the same time-frequency resources with different power levels assigned based on users channel conditions. At the receiver side, a strong user employs successive interference cancellation to decode its signal, where the receiver first decodes and removes the high power signal intended for the weak user before decoding its own signal.

Let the total system bandwidth be B, while the BS operates under a maximum transmit power constraint P_{max}. The available bandwidth is uniformly divided into k frequency resource blocks, each with a bandwidth of B/k. Likewise, the total transmission power is equally allocated, assigning P_{max}/k to each resource block. In the power domain NOMA, each user experiences a composite channel gain and is ordered based on channel gains, where u_1 experiences poor channel which is located farthest from the BS and u_2 is relatively closer to the BS with a better channel, and this pattern continues up to u_N, who is nearest to the BS and has the strongest channel. Accordingly, the users channel gains are ordered such that $|h_1|^2 < |h_2|^2 <\cdots|h_N|^2$ and power allocation follows $P_1 \geq P_2 \geq\cdots\geq P_N$ to ensure that users with weaker channels receive more power. The BS transmits a superimposed signal comprising information symbols intended for each user:

$$x= \sum_{i=1}^{N} \sqrt{P_i}\, s_i \quad (1)$$

Where P_i is the power allocated to user u_i, with the constraint $\sum_{i=1}^{N} P_i \leq P_{max}$ and s_i is the information symbols intended for user u_i. The received signal at user u_i is:

$$y_i = h_i x + n_i = \sum_{i=1}^{N} h_i \sqrt{P_i} s_i + n_i \quad (2)$$

$h_i = g_i d_i^{-\alpha/2}$ $\forall i \in \{1,2,\ldots,N\}$, $g_i \sim CN(0,1)$ is small-scale fading, d_i is the users distance from the BS, α is the path-loss exponent, and $n_i \sim CN(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at user u_i. With SIC decoding at the receiver, the signal-to-interference plus noise ratio (SINR) at user u_i is given by:

$$\gamma_i = \frac{P_i |h_i|^2}{\sum_{j=i+1}^{N} P_j |h_i|^2 + \sigma^2} \quad (3)$$

The throughput for user u_i is given by:

$$R_i = B \log_2 (1 + \gamma_i) \quad (4)$$

Where B is the system bandwidth allocated to the communication channel.

3. PROPOSED ARCHITECTURE

As the number of users increases, the network creates clusters before implementing NOMA. This study uses a synthetic CSI dataset generated following the 3rd Generation Partnership Project Technical Report (3GPP TR) 38.901 propagation model, enabling controlled evaluation under realistic fading, path loss and user mobility conditions. Since real CSI datasets are rarely accessible due to operator restrictions, a simulation-based dataset ensures reproducibility and scalability for large NOMA deployments, consistent with existing literature. Fig. 1 shows the block diagram of the proposed architecture, which has different stages and gives a detailed overview of each stage.

In deep learning models, including transformer-based architectures, we often need to feed the system with considerable data, called the training set. During the training phase, the model learns meaningful feature representations from the input data by minimizing a loss function using labeled samples. In supervised learning, the training dataset contains both input features and corresponding target outputs, allowing the model to learn the mapping between them. The proposed method captures complex relationships among input features through its self-attention mechanism. After training, the learned model is evaluated using a separate test dataset to assess the performance of the transformer based NOMA system.

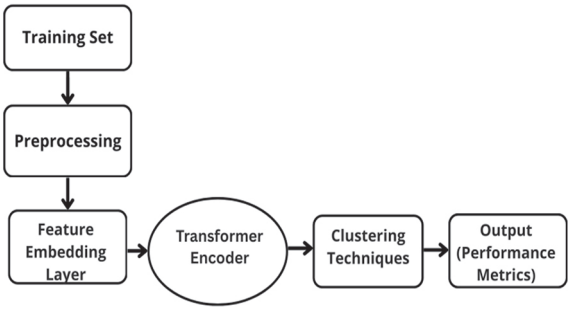

Fig. 1. Proposed architecture block diagram

3.1. TRAINING SET PREPARATION

This experiment starts with creating a dataset that consists of channel gains of users and transmitted power based on user behavior within a NOMA environment. This dataset acts as the backbone of the entire system and provides all necessary information for downstream tasks, such as clustering and embedding generation. To optimize the operation of the systems, we consider several metrics involved in curating the dataset, such as distance from the BS, channel gain, power allocation and SINR from the users perspective to develop meaningful data.

Let the input for each user u_i be represented as a feature vector: $x_i = [d_i, h_i, P_i, \gamma_i]$ where d_i is distance, h_i is channel gain, P_i is power allocation, and γ_i is SINR. All features are normalized to zero mean and unit variance prior to training.

3.2. PREPROCESSING

Preparing raw data for future analyses is the primary function of a dataset. The problem with raw datasets is that they typically have a variable scale and magnitude across their data features. This can create bias in the results of machine learning models, specifically for machine learning algorithms, where the distance between data points affects performance significantly. To solve this problem, the dataset is put through a standard scaler, which is a data normalization method that adjusts the dataset such that the mean of the dataset is equal to 0 and the standard deviation of the dataset is equal to 1.

Mathematically, it can be expressed as: $Z=(X-\bar{X})/\sigma$ where $\bar{X}$ and σ are the mean and standard deviation of X and Z is the standardized data. By scaling all of the features, the dataset ensures that channel gain and power allocation, which may originally have different numerical ranges, do not dominate the clustering process. Thus, by training the model on the normalized data, the model is able to assign equal weight to all of the different features of the data, resulting in a more accurate and meaningful clustering result. Without this normalization step, features with larger magnitudes could significantly bias the clustering outcome, leading to degraded model performance.

3.3. FEATURE EMBEDDING LAYER

The proposed system directly processes numerical CSI and does not require tokenization, as opposed to the BERT based text models. A learnable feature embedding (LFE) layer is introduced to transform the normalized CSI input into multi-dimensional tensor features for use by attention-based algorithms [25].

A trainable projection embedding is used to represent the relationship between users and their CSI properties.

$$e_i = W_e \cdot x_i + b_e \quad (5)$$

where x_i is the normalized feature vector, W_e is the embedding weight matrix, b_e is the embedding bias, and e_i is the embedded CSI representation.

3.4. TRANSFORMER ENCODER

The Transformer encoder processes the embedded CSI vectors to learn users contextual relationships with their respective neighborhoods by modeling global dependencies through multi-head self-attention [27]. Let N represent the total number of users in the system. The input to the encoder is represented as an embedding matrix.

$$\mathrm{Z} = [\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_\mathrm{N}]^\mathrm{T} \tag{6}$$

where e_k represents the embedded CSI vector of user k.

There are two main components present in the Transformer encoder layer: multi-head self-attention (MSA) and feed-forward network (FFN). Both components contain residual connections and layer normalization, which help stabilize training and improve the learning capability of the model. The self-attention computes the similarity between each user and all other users within the system, enabling the model to capture the level of interaction and potential interference among users sharing the same channel.

The MSA approach allows the model to establish important associations among users who share similar channel condition characteristics or have a poor channel condition. Therefore, MSA and FFN together make up the fundamental parts of the Transformer encoder, which allows the model to identify more complicated relationships in the set of CSI than traditional cluster algorithms could identify. After MSA has processed the user embeddings created by the model, an element-wise feed-forward network is applied to create high-level representations that enhance the model's ability to identify nonlinear interactions among the various channel characteristics. Additionally, by including residual connections, original CSI data will maintain its information content while incorporating the features learned from the user data. The encoder layers are stacked together to produce a final embedding for all of the users, which will consist of individual characteristics and the relationships between users that were learned from the model.

$$\mathrm{h_i} = \mathrm{Encoder}(\mathrm{e_i}) \tag{7}$$

Where h_i reflects both individual CSI behavior and learned inter-user dependencies.

As a result, the Transformer encoder produces a structured feature space where users with similar behavior are naturally grouped and strong-weak distinctions become more prominent. These properties significantly improve the performance of downstream clustering algorithms and enable more efficient NOMA user pairing. Further, the Transformer encoder is trained using a contrastive reconstruction loss, defined as

$$L = (1/K)\Sigma_{i=1}^{\mathrm{K}} ||x_i - \hat{x}_i||^2 \tag{8}$$

Where x_i denotes the input CSI feature vector and $\hat{x}_i$ represents the reconstructed output, and K is the number of training samples. This loss encourages the Transformer encoder to learn compact and information-preserving representations suitable for clustering.

The Transformer model is trained using the Adam optimizer with a learning rate of 10^{-4} and batch size of 128. The network consists of 4 Transformer layers, each with 8 attention heads and an embedding dimension of 128. Training is conducted for 100 epochs, and the validation set is used for hyperparameter tuning and convergence monitoring. All random seeds and initialization parameters are fixed to ensure reproducibility. The Transformer encoder converts raw CSI into contextual embeddings that capture global inter-user relationships using a self-attention mechanism.

These embeddings provide a structured feature space in which conventional clustering algorithms such as K-Means, Balanced K-Means, and DBSCAN can operate more effectively. By working on the transformed embedding space H instead of the raw CSI, clusters become more coherent, balanced, and aligned with strong weak pairing requirements in NOMA systems. This significantly improves throughput, fairness, SIC stability, and overall system performance.

3.5. CLUSTERING METHODS

Clustering or partitioning of users is one of the widely used techniques and plays a crucial role in order to enhance the performance of the NOMA system. Among various clustering schemes, some commonly used traditional clustering schemes are K-means, hierarchical clustering, and DBSCAN. Traditional clustering schemes often fail to achieve balanced cluster sizes, leading to suboptimal user pairing and reduced system performance. In this paper, we propose a Balanced K-means and transformer encoder model to address this challenge. These schemes prioritize balanced cluster sizes and optimize user pairing to maximize the utilization of NOMA, spectral efficiency and throughput of the NOMA network.

K- means remains attractive for NOMA because of its low computational cost and its ability to partition users based on similarities in CSI or distance. Algorithm 1 iteratively assigns each user to the nearest centroid and updates the centroid based on the cluster members. Although K-means is popular, it has a number of limitations that make it less than ideal for use with NOMA. First, K-means clustering is highly sensitive to the selection of the initial centroids; it does not provide a mechanism for constraining the size of the clusters, which often leads to unbalanced clusters with less diversity in the power domain and poorer performance for SIC. Second, K-means is not well suited for irregular user distributions, dynamic user mobility or the presence of extreme outlier values. Even a small number of these extreme values can significantly change the position of centroids, leading to unstable clustering results and reduced reliability of the overall system.

Algorithm 1: K-Means Clustering

Require: User set $U=\{u_1, u_2,\ldots, u_N\}$ with feature vectors x_i ; number of clusters m.

Ensure:user clusters $C=C_1, C_2,\ldots, C_m$

1. Initialize cluster centroids $\mu_1, \mu_2,\ldots, \mu_m$ randomly
2. Repeat
3. Reset all clusters: $C_k \leftarrow \emptyset$ for k=1, 2,…,m
4. for each user u_i do
5. Compute distance $d(x_i, \mu_k)$ to each centroid μ_k
6. Assign u_i to the closest cluster C_k with minimum distance
7. end for
8. for each cluster C_k do
9. Update centroid $\mu_k = 1/|C_k|) \sum(x_{i\in Ck}) x_i$
10. end for
11. Until cluster assignments do not change
12. Return C

The clustering technique used by DBSCAN is based instead on local density. This allows DBSCAN to discover clusters with arbitrary shapes and to easily identify the presence of outliers (noise) or user errors that affect clustering ability. These characteristics make it suitable for heterogeneous network environments such as metropolitan areas where there are large numbers of users clustered in high-density places and small numbers of users in low-density areas such as cell-edge locations. However, DBSCAN introduces its own challenges when applied to NOMA systems. Its performance strongly depends on two parameters, the neighborhood radius ϵ and the minimum number of points, making it extremely sensitive to variations in user density, fading conditions and mobility. Even small changes in these parameters can produce significantly different clustering outcomes, the formation of too many clusters, too few clusters or a large number of users classified as noise, resulting in unstable user grouping. Furthermore, the computational complexity also increases rapidly with the number of users and dimensions of CSI, making it less suitable for large scale, real time base station scheduling. The detailed procedure of the DBSCAN based user clustering is summarized in Algorithm 2.

Algorithm 2: DBSCAN Clustering

Require: User feature vectors $X=\{x_1, x_2,\ldots, x_N\}$; neighborhood radius ϵ; minimum points MinPts

Ensure: Clusters C and noise set N

1. Mark all users as unvisited
2. $C=\emptyset$, $N=\emptyset$
3. for each user x_i do
4. If x_i is unvisited then
5. Mark x_i as visited
6. Retrieve neighbors: $N\in(x_i)$
7. if $|N\in(x_i)| <$ MinPts then
8. Mark x_i as noise; add to N
9. else
10. Create new cluster C_k
11. Expand cluster C_k using $N\in(x_i)$
12. Add C_k to C
13. end if
14. end if
15. end for
16. return C and N

The expand cluster step is used in DBSCAN to grow a cluster from an initial core point by adding all neighboring points that meet the density requirement. It ensures that all points that are density-reachable, including neighbors of neighbors, are included in the same cluster, allowing DBSCAN to detect full, arbitrarily shaped clusters rather than only immediate neighbors.

Expand Cluster Procedure

Procedure Expand Cluster (x_i, Neighborset, C_k)

1. Add x_i and C_k
2. for each user x_j Neighborset do
3. If x_j is unvisited then
4. Mark x_j as visited
5. Retrieve neighbors: $N_\epsilon(x_j)$
6. if $|N_\epsilon(x_j)| \geq MinPts$ then
7. Neighborset =Neighborset U $N_\epsilon(x_j)$
8. end if
9. end if
10. if x_j not in any cluster then
11. Add x_j and C_k
12. end if
13. end for

Balanced K-Means extends the classical K-Means algorithm by enforcing cluster size constraints so that each cluster contains approximately an equal number of data points, denoted as "n" . During the assignment phase, each data point is allocated to its nearest centroid subject to the constraint that no cluster exceeds its capacity. This ensures that users are evenly distributed among "m" clusters, avoiding load imbalance that may degrade system performance. The implementation of the clustering method is presented in Algorithm 3.

Algorithm 3: Balanced K-Means Clustering for NOMA User Grouping

Require: User set $U=\{u_1, u_2,\ldots, u_N\}$ with feature vectors; number of clusters m.

Ensure: Balanced user clusters $C_1, C_2,\ldots, C_m$

1. Initialize cluster centroids $\mu_1, \mu_2,\ldots, \mu_m$
2. while not converged do
3. Reset all clusters: $C_j \leftarrow \emptyset$ for j=1, 2,…,m
4. for each user $u_i \in U$ do
5. Compute distance $d(u_i, \mu_j)$ to each centroid μ_j
6. Assign u_i to the closest cluster C_j such that $C_j| <\lceil N/m \rceil$

7. end for
8. for each cluster C_j do
9. Update centroid $\mu_j = (1/|C_j|) \sum_{u \in Cj} u$
10. end for
11. if no user reassignment occurs then
12. break
13. end if
14. end while
15. return: User clusters $C_1, C_2, \ldots, C_m$.

4. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed transformer-based user clustering and pairing framework in a downlink NOMA system. MATLAB simula-tions were conducted for 5,000 synthetic users under Rayleigh fading, with the signal-to-noise ratio varied across a wide operational range to assess performance under differ-ent channel conditions. The dataset is divided into 70% training, 15% validation, and 15% testing to ensure general-ization to unseen channel realizations. The trained encoder can be directly applied at the base station using real-time CSI measurements, without requiring labeled data or re-training. In comparing the transformer encoder driven clus-tering method proposed herein with the three baseline clustering algorithms (K-Means, DBSCAN, and Balanced K-Means), several key performance metrics will be examined (throughput, energy efficiency, Jain's fairness index, outage probability and bit error rate) that together describe how well the proposed approach maintains reliable communica-tion, utilizes spectrum efficiently and distributes resources fairly among users.

The simulation parameters and system configuration used for this work are presented in Table 1, with subsequent sections providing in-depth comparisons for each of the five performance indicators.

Table 1. Parameters utilized for simulation.

S.No	Parameters	Range
1	Number of users	5000
2	SNR range	0 to 20 dB
3	Noise power σ^2	1 (normalized)
4	Transmit power P_{total}	1Watt (normalized)
5	Path loss exponent	4

The throughput performances of all clustering techniques discussed are summarized in Fig. 2.

The transformer-based model produces the highest throughput rates for each SNR tested, as well as significantly higher average throughput than any of the traditional clustering methods. The transformer-based models ability to generate optimized user groupings, along with consideration for interferences, resulted in significantly improved resource utilization efficiencies and increased spectral efficiencies compared with the other clustering approaches. Both Balanced K-means and DBSCAN clustering demonstrate increased throughput compared to traditional clustering methods by providing improved stability, while the throughput generated by classical K-means is significantly lower than that of the other methods because of its sensitivity to centroid initialization and its failure to form effective clusters, which results in less than optimal intra-cluster power distribution and decreased SIC decoding reliability.

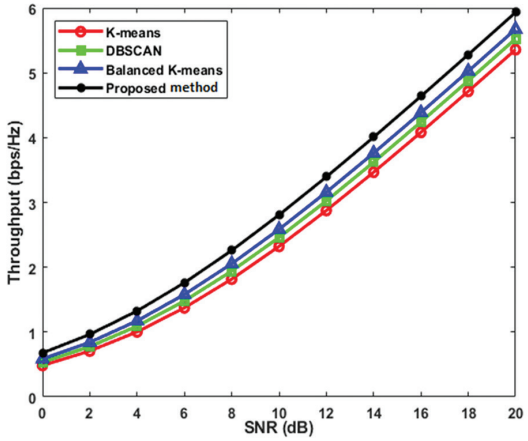

Fig. 2. Throughput comparison with proposed method

The energy efficiency (EE) of each of the evaluated clustering methods is represented in Fig. 3. In essence, EE is defined as the amount of throughput attained divided by the total power used during transmission. The transformer-based framework achieves the highest EE for all conditions of SNR. In particular, it provides approximately 20% higher EE than the K-means method at low and moderate SNR levels. This improvement is mainly due to the model's capability to form interference-aware clusters and perform more effective power allocation, which allows the transmission of a greater number of useful bits per unit of consumed energy. Balanced K-means also demonstrates competitive performance because it maintains relatively balanced cluster sizes and reduces unnecessary power consumption. In contrast, DBSCAN and conventional K-means show lower energy efficiency, mainly because their clustering structures are often irregular and less optimized for efficient power utilization.

The transformer encoder is trained in an offline stage, and its inference complexity increases approximately linearly with the number of users in each batch. During deployment, inference is performed at the base station using standard GPU or accelerator hardware. Therefore, the computational processing associated with the model does not influence the transmission-side energy model considered in this work. As a result, the reported improvements in energy efficiency originate from more effective cluster formation and better power allocation, rather than from reductions in hardware-level power consumption.

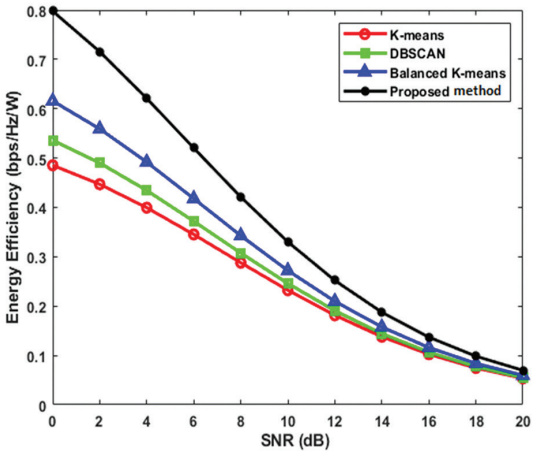

Fig. 3. Comparison of proposed method for EE vs SNR

Fig. 4 compares Jain's fairness index obtained for the different clustering approaches, indicating how evenly the available network resources are distributed among users. The proposed model consistently achieves the highest fairness, with values ranging from 0.81 to 0.93, indicating a more balanced distribution of throughput and significantly reduced user starvation.

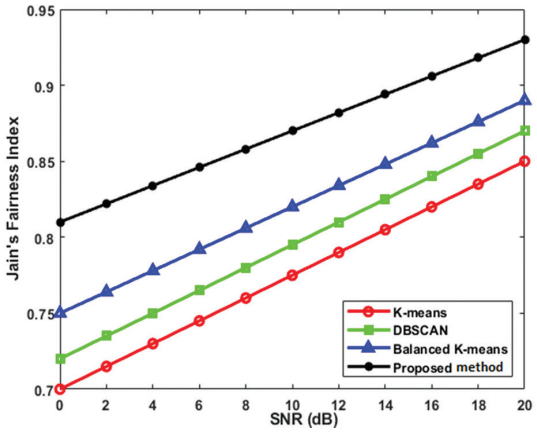

Fig. 4. Jain's fairness index comparison of all the schemes

The proposed model's fairness improvements were found to be statistically significant when compared with all baseline methods. Balanced K-means also delivers competitive fairness by enforcing cluster size constraints, while traditional K-means and DBSCAN show more uneven fairness due to their tendency to form irregular or imbalanced clusters.

Fig. 5 illustrates the BER performance of the different clustering methods across the entire SNR range. As expected, BER decreases for all schemes as SNR increases. The transformer-based clustering method proposed achieves the best BER at all SNR values. This improvement is due to the encoder's ability to learn more about the context associated with CSI over a longer period than the other methods, allowing it to create better user groupings that characterize potential interference and therefore allow for more accurate signal separation when utilizing SIC. Additionally, Balanced K-means produces reasonable BERs due to the ability to form clusters in an organized and controlled manner. Traditional K-means and DBSCAN perform less well than Balanced K-means due to their reliance on initialization and density thresholding methods, both of which introduce uncertainty into the cluster assignments generated by those algorithms.

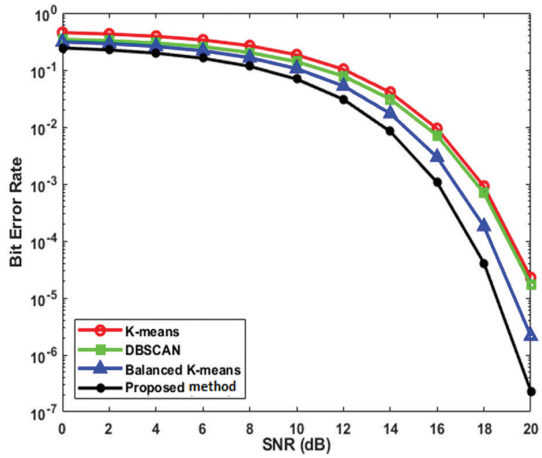

Fig. 5. BER comparison for NOMA with different methods

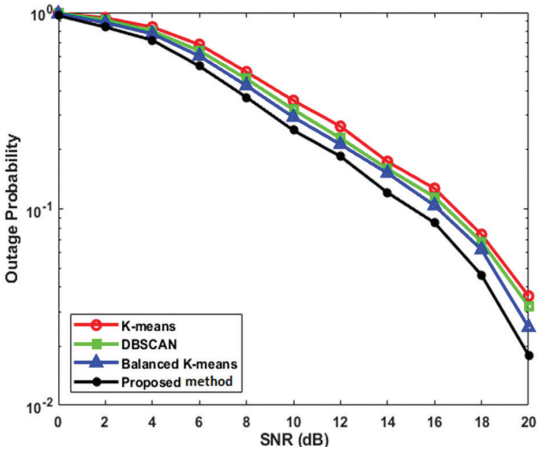

Fig. 6. Outage probability comparison for NOMA with different methods

Fig. 6 represents the outage probability of the evaluated clustering schemes under Rayleigh fading, where an outage probability event is defined as a users instantaneous achievable rate falling below a predefined threshold. The proposed method has the lowest outage probability of all scheme implementations, which indicates it has a high capability of being robust against fluctuations due to channel fading and interference in the SNR levels.

The achieved lower outage probability is attributed to the use of a transformer model, which enables the creation of more coherent user groupings, thereby allowing for greater separation of received signals and an increase in the reliability of power allocation. Balanced K-means clustering also provides a substantial gain over conventional clustering methods, as cluster sizes

are enforced to be uniform throughout all clusters, thereby reducing the number of overloaded clusters. Classical K-means and DBSCAN implementations have higher outage probabilities than the proposed method throughout all SNR levels, especially during low SNR levels, due to the sensitivity of the methods to initialization and density parameter selection, resulting in poor user grouping and degraded achievable rate performance. The results presented in the figure represent the average of multiple simulation runs and therefore show statistical robustness.

To ensure statistical rigor, all performance curves were averaged over 20 independent Monte-Carlo realizations of user distributions and Rayleigh fading. For each performance metric, the variance across runs was extremely small due to the large number of users (5000), which results in inherently stable estimates. Therefore, confidence intervals largely overlapped with the mean curves and are omitted for clarity.

5. CONCLUSION

This work presented a transformer driven deep learning framework for adaptive user clustering and pairing in downlink NOMA systems. By employing a transformer encoder to generate contextual CSI embeddings, the proposed model effectively captures complex inter user relationships that traditional clustering approaches fail to represent. Using a 3GPP based synthetic dataset, extensive simulations demonstrate that the transformer based framework significantly improves clustering stability and NOMA performance. Compared with K-means, balanced K-means and DBSCAN, the proposed method achieved significantly lower BER, higher throughput, improved fairness, enhanced energy efficiency and reduced outage probability across different SNR regimes. The results demonstrate significant improvements in key performance indicators, indicating that transformer-based structure can be leveraged for learning specific CSI relations in order to optimally allocate NOMA resources. These results indicate that transformer-based methods offer a promising direction for supporting future 6G and AI-enabled wireless communication systems. Future research should focus on examining the effectiveness of hybrid transformer/GNN architectures while also incorporating real-world CSI data sets into their design and optimizing computational aspects for utilization by real-time base station equipment.

DECLARATION OF THE USE OF AI TOOLS

The authors used Grammarly and Copilot to improve language clarity and readability, including assistance with sentence restructuring and text refinement. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the integrity and accuracy of this work. No AI tool is listed as an author, and AI tools are not cited as primary sources.

REFERENCES

[1] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends", IEEE Communications Magazine, Vol. 53, No. 9, 2015, pp. 74-81. doi: 10.1109/MCOM.2015.7263349.

[2] Y. Liu, Z. Ding, M. Elkashlan, H. V. Poor, "Cooperative Non-orthogonal Multiple Access With Simultaneous Wireless Information and Power Transfer", IEEE Journal on Selected Areas in Communications, Vol. 34, No. 4, 2016, pp. 938-953. doi: 10.1109/JSAC.2016.2549378.

[3] M. Vaezi, Z. Ding, H. V. Poor, "Multiple access techniques for 5G wireless networks and beyond", IEEE Communications Magazine, Vol. 57, No. 4, 2019, pp. 80-86. doi: 10.1007/978-3-319-92090-0.

[4] A. Akbar, S. Jangsher, F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques", Computer Networks, Vol. 190, 2021, p. 107950, doi: 10.1016/j.comnet.2021.107950.

[5] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, T. Nakamura, "NOMA: From concept to standardization", Proceedings of the IEEE Conference on Standards for Communications and Networking, Tokyo, Japan, 28-30 October 2015, pp. 18-23. doi: 10.1109/CSCN.2015.7390414.

[6] S. M. R. Islam, M. Zeng, O. A. Dobre, K.-S. Kwak, "Resource Allocation for Downlink NOMA Systems: Key Techniques and Open Issues", IEEE Wireless Communications, Vol. 25, No. 2, 2018, pp. 40-47. doi: 10.1109/MWC.2018.1700099.

[7] N. S. Mouni, A. Kumar, P. K. Upadhyay, "Adaptive User Pairing for NOMA Systems With Imperfect SIC", IEEE Wireless Communications Letters, Vol. 10, No. 7, 2021, pp. 1547-1551. doi: 10.1109/LWC.2021.3074036.

[8] Z. Ding et al. "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks", IEEE Communications Magazine, Vol. 55, No. 2, 2017, pp. 185-191. doi: 10.1109/MCOM.2017.1500657CM.

[9] F. Fang, K. Wang, Z. Ding, V. C. M. Leung, "Energy-Efficient Resource Allocation for NOMA-MEC Net-

works With Imperfect CSI", IEEE Transactions on Communications, Vol. 69, No. 5, 2021, pp. 3436-3449. doi: 10.1109/TCOMM.2021.3058964.

[10] B. Zhu, S. Zhu, K. Chi, S. Mumtaz, W. Bazzi, "Max-Min Computation Optimization in Multi-BS WPT-MEC Networks via Multi-Agent Reinforcement Learning", IEEE Transactions on Mobile Computing. doi: 10.1109/TMC.2025.3637266. (in press)

[11] N. S. Mouni, A. Kumar, P. K. Upadhyay, "Adaptive User Pairing for NOMA Systems With Imperfect SIC", IEEE Wireless Communications Letters, Vol. 10, No. 7, 2021, pp. 1547-1551. doi: 10.1109/LWC.2021.3074036.

[12] M. Shirvanimoghaddam, M. Dohler, S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations", IEEE Communications Magazine, Vol. 55, No. 9, 2017, pp. 55-61. doi: 10.1109/MCOM.2017.1600618.

[13] K. Katta, R. C. Mishra, K. Deka, "Optimal user pairing using the shortest path algorithm", Proceedings of the IEEE International Conference on Advanced Networks and Telecommunications Systems, Hyderabad, India, 13-16 December 2021, pp. 1-6. doi: 10.1109/ANTS52808.2021.9937012.

[14] M. S. Ali, H. Tabassum, E. Hossain, "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems", IEEE Access, Vol. 4, 2016, pp. 6325-6343. doi: 10.1109/ACCESS.2016.2604821.

[15] S. M. Hamedoon, J. N. Chattha, M. Bilal, "A novel user clustering and efficient resource allocation in non-orthogonal mutliple access for IoT networks", PLoS One, Vol. 19, No. 9, 2024, p. e0309695. doi: 10.1371/journal.pone.0309695.

[16] Q. Wu, R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network", IEEE Communications Magazine, Vol. 58, No. 1, 2020, pp. 106-112. doi: 10.1109/MCOM.001.1900107.

[17] J. Cui, Z. Ding, P. Fan, N. Al-Dhahir, "Unsupervised Machine Learning-Based User Clustering in Millimeter-Wave-NOMA Systems", IEEE Transactions on Wireless Communications, Vol. 17, No. 11, 2018, pp. 7425-7440. doi: 10.1109/TWC.2018.2867180.

[18] S. A. H. Mohsan, Y. Li, A. V. Shvetsov, J. Varela-Aldás, S. M. Mostafa, A. Elfikky, "A Survey of Deep Learning Based NOMA: State of the Art, Key Aspects, Open Challenges and Future Trends", Sensors, Vol. 23, No. 6, 2023.

[19] Z. Elsaraf, F. A. Khan, Q. Z. Ahmed, "Deep Learning Based Power Allocation Schemes in NOMA Systems: A Review", Proceedings of the 26th International Conference on Automation and Computing, Portsmouth, United Kingdom, 2-4 September 2021, pp. 1-6. doi: 10.23919/ICAC50006.2021.9594173.

[20] T. S. Anu and T. Raveendran, "CNN-based Channel Estimation using NOMA for mmWave Massive MIMO System", Proceedings of the IEEE Statistical Signal Processing Workshop, Hanoi, Vietnam, 2-5 July 2023, pp. 349-353. doi: 10.1109/SSP53291.2023.10207968.

[21] P. Kumaresan, T. C. Keong, N. Hoe, "An Efficient Stacked-LSTM Based User Clustering for 5G NOMA Systems", Computers, Materials & Continua, Vol. 72, No. 1, 2022, pp. 6119-6140. doi: 10.32604/cmc.2022.027223.

[22] A. Dejonghe, C. Antón-Haro, X. Mestre, L. Cardoso, C. Goursaud, "Deep Learning-Based User Clustering For Mimo-Noma Networks", Proceedings of the IEEE Wireless Communications and Networking Conference, Nanjing, China, 29 March - 1 April 2021, pp. 1-6. doi: 10.1109/WCNC49053.2021.9417426.

[23] Shen, Yifei et al. "Graph Neural Networks for Scalable Radio Resource Management: Architecture Design and Theoretical Analysis", IEEE Journal on Selected Areas in Communications, Vol. 39, 2020, pp. 101-115, doi: 10.1109/JSAC.2020.3036965.

[24] Y. Hou et al. "NOMANet: A Graph Neural Network Enabled Power Allocation Scheme for NOMA", IEEE Transactions on Vehicular Technology, 2026. doi: 10.1109/TVT.2026.3654971. (in press)

[25] M. K. Naeem, R. Abozariba, M. Asaduzzaman, M. Patwary, "Mobility Support for MIMO-NOMA User Clustering in Next-Generation Wireless Networks", IEEE Transactions on Mobile Computing, Vol. 22, No. 10, 2023, pp. 6011-6026. doi: 10.1109/TMC.2022.3186430.

[26] L. Huang, B. Zhu, R. Nan, K. Chi, Y. Wu, "Attention-Based SIC Ordering and Power Allocation for Non-Orthogonal Multiple Access Networks", IEEE Transactions on Mobile Computing, Vol. 24, No. 2, 2025, pp. 939-955. doi: 10.1109/TMC.2024.3470828.

[27] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gündüz, H. V. Poor, "Transformer-Empowered 6G Intelligent Networks: From Massive MIMO Processing to Semantic Communication", IEEE Wireless Communications, Vol. 30, No. 6, 2023, pp. 127-135. doi: 10.1109/MWC.008.2200157.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017.

[29] T. Zhou et al. "Transformer Network Based Channel Prediction for CSI Feedback Enhancement in AI-Native Air Interface", IEEE Transactions on Wireless Communications, Vol. 23, No. 9, 2024, pp. 11154-11167. doi: 10.1109/TWC.2024.3379123.

[30] Y. Elouargui, M. Zyate, A. Sassioui, M. Chergui, M. El Kamili, M. Ouzzif, "A Comprehensive Survey On Efficient Transformers", Proceedings of the IEEE 10th International Conference on Wireless Networks and Mobile Communications, Istanbul, Turkiye, 26-28 October 2023, pp. 1-6. doi: 10.1109/WINCOM59760.2023.10322921.

[31] Y. Wen, "Adaptive beamforming in millimeter-wave MIMO systems via reinforcement learning", Journal of Computational Methods in Sciences and Engineering, 2025. doi: 10.1177/14727978251352133.

[32] H. Ju, S. Jeong, B. Lee, B. Shim, "Transformer-based Predictive Channel Estimation for mmWave Massive MIMO Systems", Proceedings of the IEEE 100th Vehicular Technology Conference, Washington, DC, USA, 7-10 October 2024, pp. 1-5. doi: 10.1109/VTC2024-Fall63153.2024.10758002.

[33] J. Gao, M. Hu, C. Zhong, G. Y. Li, Z. Zhang, "An Attention-Aided Deep Learning Framework for Massive MIMO Channel Estimation", IEEE Transactions on Wireless Communications, Vol. 21, No. 3, 2022, pp. 1823-1835. doi: 10.1109/TWC.2021.3107452.