

Joint Throughput And Energy Optimization In GSM Networks Using Watts–Strogatz Graph Modeling And Reinforcement Learning

Imam Qasim Mousa¹, Kheirolah Rahsepar Fard¹

¹Department of Computer Engineering and Information Technology, University of Qom, Ghadir boulevard, Qom, Iran

Kheirolah Rahsepar Fard : rahseparfard@gmail.com

Imam Qasim Mousa : imamqasim888@gmail.com

Abstract

In this paper, the problem of simultaneous optimization of throughput and energy consumption in GSM networks is studied as a multi-objective problem. For realistic modeling of the network structure, the cell topology is represented as a “small world” graph using the Watts–Strogatz stochastic graph model, and by exploring the graph, important structural and traffic features of each cell, including degree, centrality, traffic load, and QoS indices, are extracted. These features, together with the instantaneous channel information and the power status of the base stations, form the state space of a deep reinforcement learning agent. Using a neural network-based Q-value function approximation, the DRL agent learns control actions such as adjusting the transmit power, turning cells on/off, and routing traffic between neighboring cells in such a way that a reward function including normalized throughput, normalized energy consumption, and QoS violation penalty is maximized. Simulation results in the scenario of one macro cell and 12 micro cells with 200 to 2000 users show that the proposed method has a significant advantage over the three reference methods (basic, heuristic and classical Q-learning); so that the average cumulative reward of the DRL algorithm reached 3.41, while for Q-learning and the heuristic method it was 2.28 and 0.7, respectively, and for the basic method it was reported to be negative 3.55. Also, the network energy efficiency, in terms of the ratio of total throughput to base station power consumption, is maintained at a stable level of close to 2.5 for the proposed method, while this value is about 2.05 for Q-learning and 1.56 and 1.21 for the two non-intelligent methods, respectively. These results indicate that the combination of small-world graph modeling and deep reinforcement learning can provide an efficient framework for intelligent energy and capacity management in GSM networks.

Keywords: GSM cellular networks, energy efficiency, throughput optimization, Watts–Strogatz small-world graph, deep reinforcement learning

1-Introduction

In recent decades, cellular networks based on the GSM standard have been the main infrastructure for providing mobile telecommunications services in many countries and continue to support a significant portion of voice and data traffic [1]. The increasing number of users, the diversity of services, and the requirements for quality of service (QoS) have made the optimal use of radio and hardware resources of these networks a major challenge. On the other hand, the increase in energy consumption in base stations not only increases the operating costs of operators, but also creates significant environmental impacts and conflicts with the goals of green and sustainable networks [2]. Estimates show that a major share of energy consumption in cellular networks is allocated to base stations, and any improvement in energy efficiency in this sector can have a direct impact on reducing costs and greenhouse gas emissions [3].

In such circumstances, simultaneous optimization of throughput and energy consumption in GSM networks has gained special importance as a multi-objective problem. In most classical studies, the optimization problem in GSM networks has been raised as a single objective or with an emphasis on a main criterion such as maximizing

capacity, reducing call drop, or minimizing interference [4]. In these approaches, other criteria are usually assumed to be fixed or implicitly and indirectly considered. However, in real scenarios, the operator is forced to establish a dynamic balance between conflicting objectives, such as increasing throughput and reducing energy consumption. For example, increasing transmission power can improve signal quality and thus throughput, but at the cost of increasing energy consumption and inter-cell interference. On the other hand, reducing power and turning off some equipment may help save energy, but if not managed properly, it will degrade the QoS of users. Therefore, multi-objective optimization approaches that can manage this trade-off in a systematic and intelligent manner are needed [5].

One of the important prerequisites for designing intelligent algorithms in cellular networks is to have a proper model of the network structure and dynamics. In this regard, modeling the GSM network as a graph with “small-world” properties can provide a more realistic view of the relationships between cells [6]. Several studies have shown that many real networks – from social networks to communication networks – have similar structural features to small-world networks, namely a combination of high local clustering and short average path lengths between nodes [7]. Such features can also be observed in cellular networks: each cell interacts directly with a number of nearby neighbors, but more distant cells can also be reached through a small number of intermediate links. These properties make the use of stochastic graph models such as the Watts–Strogatz model attractive for approximating the topological structure of cellular networks [8].

The Watts–Strogatz model is a well-known framework for generating small-world random graphs, in which each node is first connected to a few of its nearest neighbors in a regular manner, and then some of the edges are redrawn with a certain probability and connected to more distant nodes. The result of this process is a graph that simultaneously benefits from relatively high local clustering and small average path length. If we consider cells or base stations of a GSM network as nodes of this graph, the edges can represent neighborhood relationships, handover possibilities, or dominant paths of users moving between cells. In this way, the Watts–Strogatz model provides a suitable platform for realistic simulation of the communication structure between cells and analysis of network behavior [9].

In such a framework, the use of graph exploration techniques on the generated random graph allows the extraction of important structural and traffic features [10, 11]. Indices such as node degree, different types of centrality, clustering and community structure can describe the role of each cell in the network and its importance in traffic routing or the formation of inter-cell interference. The analysis of these features on the Watts–Strogatz graph corresponding to the GSM network allows the identification of critical cells, high-traffic areas and energy-sensitive points [12]. This information can play a key role in designing resource management policies and adjusting parameters as input to a learning or optimization algorithm. In recent years, reinforcement learning and especially its deep variants have been considered as one of the new approaches to solving complex sequential decision-making problems in telecommunication systems. In this framework, the learning agent learns control policies by interacting directly with the environment in such a way that it maximizes a long-term reward function [13].

An important feature of reinforcement learning is that it does not require a detailed analytical model of the environment and can learn optimal or quasi-optimal policies based on experience and observed data in situations where explicit system modeling is difficult or expensive. This feature makes reinforcement learning very

attractive for application in cellular networks with high complexity, uncertainty, complex graph structure, and time-varying traffic [14-16].

Despite the significant growth of research in applying reinforcement learning to resource management in next-generation networks, several important gaps can be identified in the field of GSM networks [17]. First, a significant part of the existing work mainly focuses on a specific objective such as capacity maximization or coverage improvement, and the problem of simultaneous optimization of throughput and energy consumption as a multi-objective framework has not been addressed. Second, in many studies, the network structure is either considered very simple and regular or general graph models are used without considering the small-world properties. Third, the coherent combination of graph modeling based on the Watts–Strogatz model, graph exploration and reinforcement learning to extract structural features and use them in designing optimal control policies has been less studied [18-20].

The present paper, titled “Simultaneous Optimization of Throughput and Energy Consumption in GSM Networks Using Graph Modeling, Graph Exploration and Reinforcement Learning”, seeks to address these gaps. In the proposed approach, first the GSM network is modeled as a small-world stochastic graph based on the Watts–Strogatz model, in which the number of local neighbors of each cell and the probability of edge redrawing determine the topological structure of the network. Then, using graph exploration methods, a set of structural and traffic features are extracted from this graph that well describe the state of each cell and its role in the network. These features, together with traffic information and current control parameters, form the state space of the reinforcement learning agent.

In the next step, a multi-objective reinforcement learning framework is designed in which the agent, by dynamically and intelligently adjusting parameters such as base station transmission power and radio resource allocation pattern, tries to maximize a reward function that simultaneously maximizes network throughput and minimizes energy consumption. The proposed reward function is defined in such a way that, in addition to combining throughput and energy criteria, user quality of service constraints are also included as a penalty; such that violating QoS constraints leads to a reduction in reward and guides the agent towards acceptable policies in terms of service provision. In this way, the learning agent learns to create an appropriate balance between conflicting goals and adopts policies that improve the overall network performance in the long run.

The main innovation of this work can be summarized in three axes: first, the systematic use of the Watts–Strogatz stochastic graph model to approximate the topological structure of the GSM network and the application of graph exploration to enrich the state space; second, the design of a multi-objective reinforcement learning framework that can simultaneously optimize throughput and energy consumption; and third, the comprehensive evaluation of the proposed method in simulated scenarios and its comparison with conventional methods based on fixed tuning or simple control rules. The results are expected to show that combining the structural insight obtained from small-world graphs with the adaptability of reinforcement learning can lead to significant improvements in network performance in terms of throughput and energy efficiency.

The remaining structure of the paper is organized as follows. In the next section, a comprehensive review of related works in the areas of energy and throughput optimization in cellular networks, graph modeling based on small-world networks, and applications of reinforcement learning in resource management is provided. Then, in the third section, the system model and the GSM network modeling method based on the

Watts-Strogatz model are described along with the graph exploration methods used. The fourth section is dedicated to the formulation of the multi-objective optimization problem and the definition of the main components of the reinforcement learning framework. In the fifth section, the simulation results and comparative analysis of the performance of the proposed method are presented and finally, in the final section, a conclusion and suggestions for further research in this field will be presented.

2-Related works

Wen et al. (2015) [21] formulate the problem of turning off base stations in CoMP scenario as a sequential decision problem and use multi-stage Q-learning to select the sleep strategy of cells. The goal is to reduce BS energy consumption while maintaining downstream performance and quality of service. Simulations show that the learned policy provides simultaneous improvements in energy efficiency and throughput efficiency compared to simple threshold methods.

Ali et al. (2016) [22] formulate the problem of maximizing the ratio of network throughput to power consumption as a nonlinear fractional programming in a HetNet equipped with relays and D2D communication. Using the Charnes–Cooper transform and the Outer Approximation algorithm, they obtain the ε -optimal solution for power, spectrum allocation, and cell selection, and show that the EE is significantly improved over the reference schemes.

Shahid et al. (2017) [23] present a self-organizing framework for different layers in a heterogeneous D2D-based network that jointly regulates the physical and MAC layers to reduce energy consumption and improve spectral efficiency. By modeling user and cell interactions in a game format and designing adaptive policies, they show that the energy consumption per bit is reduced while maintaining capacity and QoS.

Li et al. (2018) propose an optimal sleep strategy for dense heterogeneous networks, in which small cells are clustered based on an interference graph. The nodes of the graph represent BSs and edges indicate strong interference. By shutting down under-loaded clusters and reconfiguring the coverage, the energy consumption is reduced, while the total throughput and user coverage are kept under QoS constraints [24].

Nasir&Guo (2019) formulate dynamic power control in a multi-cell network as a multi-agent reinforcement learning problem where each link is a DRL agent. Using a joint DNN architecture and simultaneous updating, the proposed algorithm optimizes the total network rate in the presence of interference. The results show that this approach achieves higher throughput while consuming less power per target rate than classical optimization algorithms such as WMMSE [25].

Zhang (2020) introduces a centralized DRL framework for resource allocation in next-generation networks, which aims to maximize a profit function including a combination of energy efficiency and throughput. The non-convex problem is modeled as an MDP, and the DQN agent learns spectrum and power allocation decisions by observing the current load, channel, and power status. The results show that the proposed scheme provides higher EE while maintaining or improving the aggregate rate compared to convex optimization-based methods [26].

Hsieh et al. (2021) propose a DRL framework for joint optimization of power allocation and user association in heterogeneous networks. The state includes the load of each BS, link quality, and channel state, and the DQN agent selects the BS and power level for each user to maximize the total energy efficiency. Simulations show that the proposed method simultaneously provides higher EE and acceptable user rate compared to heuristic algorithms [27].

Azimi et al. (2022) study the resource allocation problem in 5G RAN slicing and propose the EE-DRL-RA algorithm. In this framework, a coarse-scale DL network determines the

resource amount of each slice, and a fine-scale DRL agent performs the precise allocation of power and radio resources to users of each slice. With this hierarchical structure, both the functional separation of slices is maintained and the energy efficiency and network capacity are improved [28].

Mughees et al. (2023) design a multi-agent parametric DRL (MA-PDRL) approach to solve the joint user association and power allocation problem in 5G HetNet. The continuous power and discrete user association action space are modeled in the form of parametric actions. The results show that the MA-PDRL method achieves higher energy efficiency, higher aggregation rate, and a higher percentage of users with acceptable QoS than WMMSE, non-cooperative games, and traditional DRL [29].

Choi et al. (2024) present a dynamic and energy-efficient eCIC scheme for H-CRAN networks, in which the DRL agent adjusts the Almost Blank Subframes scheduling and power parameters according to the interference situation and network load. The reward function is a combination of energy efficiency and quality of service. Simulations show that compared to static eCIC, energy consumption is reduced and spectral efficiency and user throughput are improved [30].

Azimi et al. (2024) propose an energy-aware federated deep reinforcement learning framework for networks beyond 5G, in which base stations train resource allocation policies in a distributed manner and only share gradients with the central server. The reward function design considers both energy efficiency and quality of service for mobile users; results show that this approach reduces signaling overhead and energy consumption while maintaining model accuracy [31].

Shilpa et al. (2024) investigate the radio resource scheduling problem in 5G networks using DRL. The state includes traffic queue, channel conditions, and service priority, and the DRL agent selects users and allocates RBs to maximize the total rate under delay constraints. The study shows that the proposed approach provides higher throughput and lower latency than classical schedulers such as PF and MLWDF [32].

Kim&So (2025) propose a distributed multi-agent deep reinforcement learning-based power control algorithm for cellular networks, where each BS is a local DRL agent. The agents learn a policy with limited information exchange that reduces inter-cell interference, ensures link quality, and lowers power consumption. Results on multi-cell scenarios show that this method increases the network EE without significantly reducing user throughput [33].

QU et al. (2025) analyze the relay selection and resource allocation problem in a relay-based D2D network with the aim of improving energy efficiency while maintaining throughput. By combining analytical modeling and optimization algorithms, they show that intelligent relay selection and joint allocation of power and bandwidth can significantly improve network EE and D2D user capacity [34].

3-Proposed method

Figure (1) shows the system model of the proposed method. On the left, the physical network including macro cells, small cells and mobile users with wireless links is shown, and on the right, the same network is drawn as an abstract graph with nodes corresponding to BTSs and edges corresponding to interference and handover relationships. The middle arrow represents the process of “graph modeling” during which the complex physical network is transformed into a simpler graph to enable structural analysis and feeding the reinforcement learning module for simultaneous optimization of throughput and energy.

In Figure (2), the overall process of the proposed method for simultaneous optimization of throughput and energy consumption in GSM networks is shown. First, the cellular network structure is modeled as a small-world graph based on the Watts-Strogatz

model; In such a way that the nodes corresponding to the cells and the edges represent the neighbor and interference relationships, and by adjusting the parameters k and p , the graph has a high clustering coefficient and a short average path length. Then, based on this graph, the structural and operational characteristics of each cell, degree, centrality, traffic load, transmission power and QoS indices are extracted and given as a state vector to the reinforcement learning module. By observing these states, the reinforcement learning agent selects control actions such as power adjustment, carrier on/off and user acceptance threshold adjustment, and gradually converges to a balanced policy between throughput and energy consumption with a reward function that combines normalized throughput, normalized energy consumption and QoS violation penalty. Finally, the performance of the method is evaluated based on criteria such as total throughput, energy efficiency, percentage of users with acceptable QoS, and call drop rate, and the results are used to better tune the parameters of the graph model and the reward function.

Figure (1) System model of the proposed method

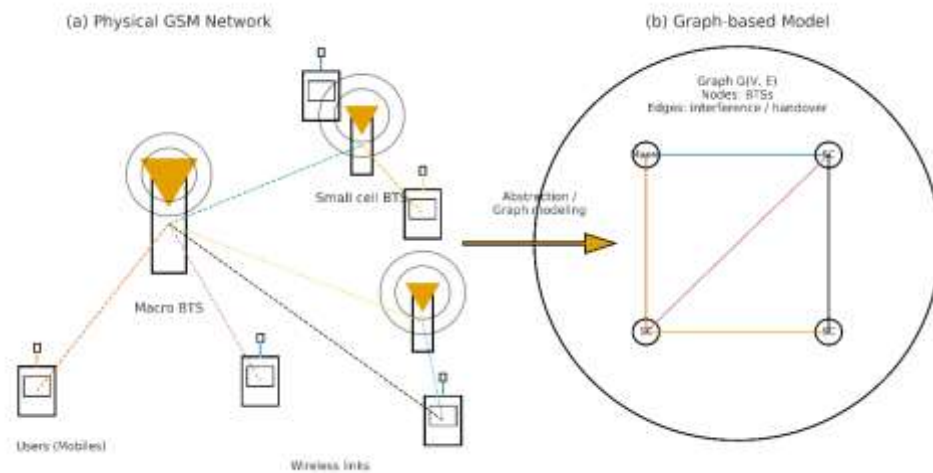
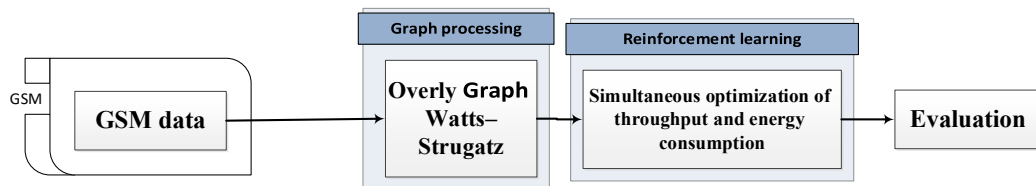


Figure (2) Block diagram of the proposed method



For the proposed method, the communication between cells of the GSM network is modeled as a small-world random graph based on the Watts–Strogatz model. In this model, each cell is first connected to its k nearest neighbors in a regular circular graph, and then some of the edges are randomly redrawn with probability p . Thus, the final graph has both a high clustering coefficient and short paths between more distant cells. By appropriately adjusting the parameters k and p , the network topology adapts to the cell density and traffic pattern, providing a structure that both realistically represents the neighborhood relationships and interference between cells and provides a suitable platform for reinforcement learning. This structure allows the learning agent to learn

low-energy control strategies for power adjustment, cell management, and serving cell selection, and as a result, the network throughput and energy efficiency of base stations are simultaneously improved. Figure 3 shows the process of constructing this graph.

Figure (3) Pseudocode of the proposed method for constructing Watts-Strogatz graphs

Input:
n: Number of cells
k: Number of nearest neighbors each cell connects to
p: Rewiring probability
Output:
G: Watts-Strogatz random graph
Algorithm:
1. Initialize G as an empty graph.
2. Create n nodes in G and arrange them in a circular layout.
3. Connect each node to its k nearest neighbors:
For each node i in G:
For j = 1 to k/2:
Add an edge between cell i and cell (i + j) mod n.
Add an edge between cell i and cell (i - j + n) mod n.
4. Rewire edges with probability p:
For each edge (i, j) in G:
Generate a random number r between 0 and 1.
If r < p:
Remove edge(i, j).
Select a new cell k randomly such that k ≠ i and (i, k) is not already an edge.
Add edge (i, k).
5. Return G.

3-1-Dynamic Learning with Cellular Automata

After modeling the communication between cells based on the Watts–Strogatz graph, the network enters the dynamic learning phase. The goal of this phase is to adaptively adjust the behavior of cells over time in order to balance energy consumption, increase communication stability, and reduce unnecessary signaling traffic. In this phase, each cell (or cell cluster) is modeled as a “learning agent” that has a set of control actions such as direct transmission to the base station, routing through neighboring cells, adjusting transmission power, and changing the functional state (e.g., active/semi-active). At each time interval, cells select one of these actions according to the current state of the network and update their decision-making policy based on the feedback received from the network – including reward or penalty. The reward function is defined based on criteria such as energy consumption, link quality, effective distance to BS and local traffic load, such that actions that lead to reduced energy consumption and improved data transmission quality receive more rewards. This learning process continues gradually and iteratively until the network reaches an equilibrium state in which energy consumption between cells is balanced, communication stability is maximized and the need for centralized control intervention is minimized. In this way, cells, as a self-organized and distributed mechanism, learn their optimal communication and operational strategies within the framework of the Watts–Strogatz graph model and reinforcement learning.

3-2-Steps of a learning automaton based on reinforcement learning

Suppose $(x, y) \in D$ is a selected serving cell. The vector $x \in X \subset R^n$ contains the values of cells where each $x(i)$ is from the set f_i such that $f_i \in F = \{f_1 \dots f_n\}$ and n is the total number of cells. Also, $y \in Y$. Let $c: F \rightarrow R$ be the mapping function of the chosen path f to the desired search cost $c(f)$. Consider the states $\tilde{S} = (x, y, \tilde{F})$ where

$$\tilde{S} = (x, y, \tilde{F}) \in \tilde{S} = X \times Y \times \psi(F) \quad (1)$$

In (1), $\psi(F)$ is the power of the set F . $\tilde{F}$ is an instance (x, y) and the set of cells in question. The agent in question can observe only one of the states (one cell out of the available cells) s , which only includes the cells selected from x , which of course do not have a label. The observed state s belongs to S , that is, $s \in S$, which is a set of $(x^{(i)}, f_i)$ that we will have for each selected cell.

$$f_i := s = \{x^i, f_i\} | \forall f_i \in \tilde{F} \quad (2)$$

The activity $a \in A$ where $a \in A = A_c \cup A_f$ is one of the possible learning activities for the selected cell, i.e. $A_c = y$ or activity $A_f = F$. The control action A_c terminates the part and the agent receives a reward of zero. In the case of correct classification or -1 in the subsequent cases, the action of selecting the best path A_f receives (releases) the corresponding value and the agent receives a negative reward of $-\lambda c(f_i)$. The set of present control states or stability control activities in the present set is limited to only the existing states that have not been examined so far. According to these definitions, the reward function r , which is $r: \tilde{S} \times A \rightarrow R$, is defined as (3)

$$r((x, y, \tilde{F}), a) = \begin{cases} -\lambda c(f_i) & \text{if } a \in A_f, a = f_i \\ 0 & \text{if } a \in A_c \text{ and } a = y \\ -1 & \text{if } a \in A_c \text{ and } a \neq y \end{cases} \quad (3)$$

In equation (4) $\lambda \in R^+$ is a cost factor. The parameter λ is a variable and determining parameter that, by changing this parameter, creates a balance between cost and stability improvement. High values of λ force the agent to prefer lower cost and smaller values of λ prefer high accuracy. (Larger λ indicates the priority of cost reduction over accuracy and smaller λ indicates the priority of accuracy over cost)

In the starting step, the states or state space S_0 , no control signal is provided, in fact $\tilde{S}_0 = (x_1 y_1 o)$ which is selected from the data set. The desired environment is deterministic and the transfer function $t: \tilde{S} \times A \rightarrow \tilde{S} \cup \tau$ will be. τ is the terminal state. In this case, the function t is defined as equation (5):

$$t((x, y, \tilde{F}), a) = \begin{cases} \tau & \text{if } a \in A_c \\ ((x, y, \tilde{F}), \alpha) & \text{if } a \in A_f \end{cases} \quad (5)$$

This property makes the desired environment for reinforcement learning inherently a segment with a maximum segment length of $|F|+1$. Now the main goal is to find a policy π such that it sequentially selects the available suitable paths from F and finally selects the desired cells from y . This operation must be performed in such a way that the desired total reward and expectation are maximized.

An agent's location will only act on the observed states s , so it cannot accurately state which state S has been selected. As a result, each agent chooses its reward and also its transfer function randomly, although the working environment is deterministic. To overcome this weakness, the action of each agent will be based on probabilities. For this reason, optimization will be performed in the reinforcement learning algorithm. For this purpose, if $t(s, a)$ is the transition probability function for the states (possible paths) of the observation, that is, S , then the action a performed is $S: t: S \times A \rightarrow \Delta(SUT)$.

Similarly, the reward function $r(s,a)$ will indicate the expected reward value if the location a is observed in states.

To improve the performance of the reinforcement learning algorithm in this research, the policy improvement method has been used. In the proposed improvement method, the action-state value pair is estimated with a function Q . This estimation is done with the help of a neural network. The function $Q^\pi(s,a)$ indicates the discount start reward in the observed state s , which is done in applying action a and presenting policy π after it. Which has a recursive form. Equation (6)

$$Q^\pi(s,a) = r(s,a) + E \left[\gamma Q^\pi(S, \pi(S)) \right] \quad (6)$$

$$s \sim t(s,a)$$

It is worth noting that $r(s,a)$ is the expected value in all possible rewards. The value of the action-state pair in the terminal state is also zero, that is, $Q(\tau, 0) = 0$

The value of the factor γ , which is also a discrete value, determines the importance of future rewards. It is very useful in non-partial environments or when an approximation function is used. Since the features of interest in this research are discrete features, the environment of interest of the data guarantees the terminating of the algorithm. The numerical value of γ is helpful during the training process and is a determining parameter.

In practice, the optimal function Q^* is desired such that relation (7) holds.

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s,a) \quad (7)$$

Which in this regard

$$Q^*(s_1 a) = r(s_1 a) + E \left(\gamma \max_{a'} Q^*(s', a') \right) \quad (8)$$

$$S \sim t(s,a)$$

In a limited number of cells and low dimensions and a small state space, the function Q is easily determined by the dynamic state. However, if the exact space has large dimensions or the space of interest is continuous, it is almost impossible to determine Q . Inspired by neural networks and dynamic programming, a θ neural network can estimate the function Q^θ , which is estimated by minimizing the MSE between the two sides of equation (9). For the transfer (s,a,r,s') that is empirically determined with the agent of interest with a greedy policy, the following equation

$$\pi^\theta(S) = \operatorname{argmax}_a Q^\theta(s,a) \quad (9)$$

In this transition, s and a represent the action performed and the current state, r represents the reward obtained, which is expected to converge to the value $s(r,a)$, and s is the next state, i.e. $t(s,a)$. To be able to express it in a formula, the parameter θ will iteratively minimize the loss function L_θ for the transition batch, i.e. B .

$$L_\theta(B) = \frac{1}{|B|} \sum_{(s,a,r,s) \in B} (q(r,s) - Q^\theta(s,a))^2 \quad (10)$$

In equation (10), q is the estimate of the objective function Q , which is constant. For values of w , r , and t : it is constant. In fact, the parameters r , w , and t are constant in the optimization step.

$$q(r,s) = r + \gamma \max_a Q^\theta(s,a) \quad (11)$$

As the error increases, the expected function Q^θ will converge to Q^* .

In summary, the model presented at this stage uses reinforcement learning to make optimal cell decisions in the network based on experience and feedback from the environment. Each node (learning agent) observes the current state of the network and evaluates possible paths to choose the action that will bring the highest expected reward

in the future. This iterative process causes the algorithm to gradually converge to the optimal policy π^* ; so that paths with lower cost and higher stability are selected more often. Using the value function $Q(s,a)$ and its continuous update based on the rewards received, makes each node's decisions more optimal and establishes a balance between accuracy, cost, and stability. The pseudocode in Figure 4 shows the implementation steps of the reinforcement learning process for choosing the optimal path:

Figure (4) Pseudocode of the implementation steps of the reinforcement learning process for optimal cell selection

Algorithm: RL-based Cell / Route Selection (Q-function Approximation)

Input:

$D = (x, y)$ // samples: features x , true label/target y
 $f_i \in F = \{f_1 \dots f_n\}$ // candidate cells/routes
 $c: F \rightarrow \mathbb{R}$ // cost of candidate f_i
 $\lambda > 0$ // cost-accuracy trade-off
 $\gamma, \alpha, lr, \varepsilon$ // discount, learning rate, exploration
 R, B // number of episodes, replay buffer size
 $Q\theta(s,a), Q\theta^-(s,a)$ // Q-network and target Q-network

State&Actions:

Latent state: $\tilde{S} = (x, y, \tilde{F}) \in X \times Y \times \psi(F)$
Observable state: $s = \Phi(\tilde{S})$ // features of selected candidates
Actions: $A = A_f \cup A_c$
 A_f : select candidate $f_i \in \tilde{F}$ // continue exploration
 A_c : commit final decision $\hat{y} \in Y$ // terminate
Reward:
if $a \in A_f$ and $a = f_i$: $r = -\lambda \cdot c(f_i)$
if $a \in A_c$ and $\hat{y} = y$: $r = 0$
if $a \in A_c$ and $\hat{y} \neq y$: $r = -1$

Procedure RL_Training

Initialize θ (and $\theta^- \leftarrow \theta$), replay buffer $D_{\text{replay}} \leftarrow \emptyset$
for episode = 1 ... R do
Sample (x, y) from D
Build initial $\tilde{F}_0$ and $\tilde{S}_0 = (x, y, \tilde{F}_0)$
 $s \leftarrow \Phi(\tilde{S}_0)$; done $\leftarrow$ false
while not done do
// 1) ε -greedy action selection
with prob ε : $a \leftarrow$ random feasible action
else: $a \leftarrow \operatorname{argmax}_a Q\theta(s,a)$

// 2) Environment response
if $a \in A_f$ (select $f_i \in \tilde{F}$):
 $r = -\lambda \cdot c(f_i)$
Update $\tilde{F} \rightarrow \tilde{F}'$ and internal metrics
 $\tilde{S}' = (x, y, \tilde{F}')$, $s' = \Phi(\tilde{S}')$
terminal_flag = false
else if $a \in A_c$:

```

 $\hat{y} = \text{decision\_from\_current\_state}(s)$ 
 $r = 0$  if  $\hat{y} = y$  else  $-1$ 
 $\tilde{s}' = \tau, s' = \tau$ 
terminal_flag = true
end
// 3) Store transition
Store  $(s, a, r, s')$  in  $D_{\text{replay}}$ ; keep at most  $B$  samples
// 4) Q-update
Sample minibatch  $\mathfrak{B} \subset D_{\text{replay}}$ 
For each  $(s_k, a_k, r_k, s'_k) \in \mathfrak{B}$ :
    if  $s'_k = \tau$ :
         $y_k = r_k$ 
    else:
         $y_k = r_k + \gamma \cdot \max_{a'} Q\theta^-(s'_k, a')$ 
    end
    Update  $\theta$  by minimizing:
 $L(\theta) = (1/|\mathfrak{B}|) \sum_k [y_k - Q\theta(s_k, a_k)]^2$ 
    Periodically set  $\theta^- \leftarrow \theta$ 
     $s \leftarrow s'$ 
if terminal_flag: done  $\leftarrow$  true
end while
 $\epsilon \leftarrow \max(\epsilon_{\text{min}}, \epsilon \cdot \text{decay})$ 
end for
Output: policy  $\pi^*(s) = \operatorname{argmax}_a Q\theta(s, a)$ 
End Algorithm

```

4-Results

Table (1) shows the simulated network configuration for evaluating the proposed method; the number of mobile users is considered to be between 200 and 2000 users and the radio access structure consists of one macro base station and 12 GSM micro cells. The coverage radius of the micro cells is set to be between 100 and 300 meters and the radius of the macro cell is set to 500 meters to create a realistic scenario of a multilayer cellular network. The data generation rate of each user is considered to be between 64 kbps and 1 Mbps and the radio standard is assumed to be based on GSM/EDGE in the 900 and 1800 MHz bands; also, the total downlink capacity of the macro cell is considered to be about 20 Mbps and the average payload size of each packet is considered to be 1500 bytes. Table (2) specifies the configuration of the Deep Reinforcement Learning (DRL) agent, in which the Deep Q-Learning algorithm is used with a neural network consisting of two hidden layers with 64 neurons per layer, a playback buffer size of 10,000 samples, a learning rate of 0.001, a discount factor of 0.95, an initial discovery rate ϵ of 0.9 with gradual reduction, and the number of training episodes is set between 500 and 1000 with 100 to 200 steps per episode. Finally, Table (3) describes the design of the state and action space; such that the state space includes variables such as SNR, RSSI, aggregation index (CG), and user density and is modeled as continuous and normalized, while the action space is designed as discrete and includes different access/path options such as LTE link selection, Wi-Fi direct connection, and Wi-Fi connection with relay R_m ; This structure enables intelligent and adaptive decision-making of the DRL agent in the cellular network environment.

Table (1) Details of simulated network settings

Parameter	Value
Number of Mobile Users (MSs)	200 – 2000
Number of GSM Microcells (BTSs)	12
Number of GSM Macro Base Stations	1
Microcell Coverage Radius	100 – 300 meters
Macrocell Coverage Radius	500 meters
Data Generation Rate per MS	64 kbps – 1 Mbps
GSM Standard	GSM/EDGE (900 / 1800 MHz)
Total Downlink Capacity (Macrocell)	~ 20 Mbps (aggregate GSM/EDGE capacity)
Average Packet Payload	1500 bytes

Table (2) DRL agent configuration

Parameter	Suggested Value
RL Algorithm Type	Deep Q-Learning (DQN)
Neural Network Structure	2 hidden layers, 64 neurons each
Replay Buffer Size	10,000 samples
Learning Rate (α)	0.001
Discount Factor (γ)	0.95
Initial ϵ (Exploration Rate)	0.9 with decay
Number of Episodes	500 – 1000
Steps per Episode	100 – 200

Table (3) State and action space design

Parameter	Description
State Space	SNR, RSSI, CG, user density
State Space Type	Continuous / normalized
Action Space	LTE, Wi-Fi direct, Wi-Fi with Relay R_m
Action Space Type	Discrete with M+2 options

In the comparison of methods, the proposed Deep Reinforcement Learning (DRL) algorithm shows the most stable and efficient behavior in the dynamic conditions of the GSM network. Relying on the approximation of the Q-value function based on the neural network, this algorithm is able to learn the complex patterns of dependence between environmental parameters - such as cell traffic load, channel quality, Watts-Strugatz graph structure and cell power/state status - well and update its control policy adaptively. The result of this ability is to obtain a higher cumulative reward during the training process and to achieve a suitable balance between throughput, energy consumption and compliance with QoS constraints.

In contrast, the classical Q-learning algorithm, although it offers significant improvements over non-intelligent methods, is inferior to DRL in terms of stability and final reward value due to its reliance on the Q-table and inherent inability to generalize in a large and continuous state space, and in complex cellular network scenarios, its convergence is slower and more sensitive to environmental changes. The heuristic-based heuristic method, although it provides limited improvement over the baseline method in the early stages, stops around a reward level close to a constant value due to its inability to adapt to changing traffic patterns and cell states and does not exceed a certain limit. Finally, the baseline method without learning (e.g., fixed power configuration and serving cell selection based on classical criteria) shows the weakest

performance and records negative rewards with high fluctuations over time; This behavior indicates the inability of this approach to deal with dynamic and changing network conditions and the inability to strike a proper balance between energy consumption and quality of service.

Figure 5 shows the average cumulative reward obtained by the four algorithms – the base method, the heuristic method, Q-learning and the proposed DRL – in the whole training process as a bar graph. These results clearly confirm the significant superiority of the proposed DRL algorithm, which achieves the best performance in terms of the combination of throughput, energy and QoS with an average reward of about 3.41; while Q-learning records an average reward of about 2.28, the heuristic method records an average reward of about 0.7 and the base method records a negative average reward of about -3.55. The analysis of Figure 6 also shows that the proposed DRL algorithm is superior to the other methods not only in terms of the final reward amount, but also in terms of convergence speed and stability of control behavior. These results confirm the suitability of the DRL framework based on graph modeling in real cellular networks with variable environments and demonstrate its generalizability to more advanced applications such as next-generation networks, the Internet of Things, and smart industrial systems.

Figure (5) Comparison of cumulative rewards of different algorithms over time for choosing data upload paths in the network

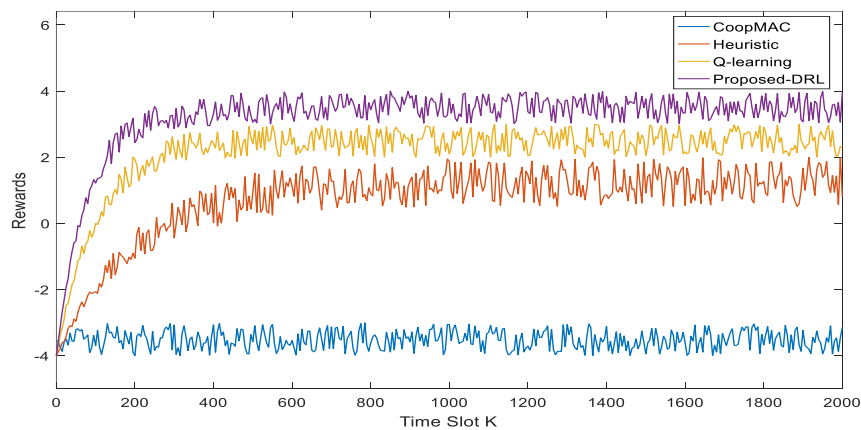
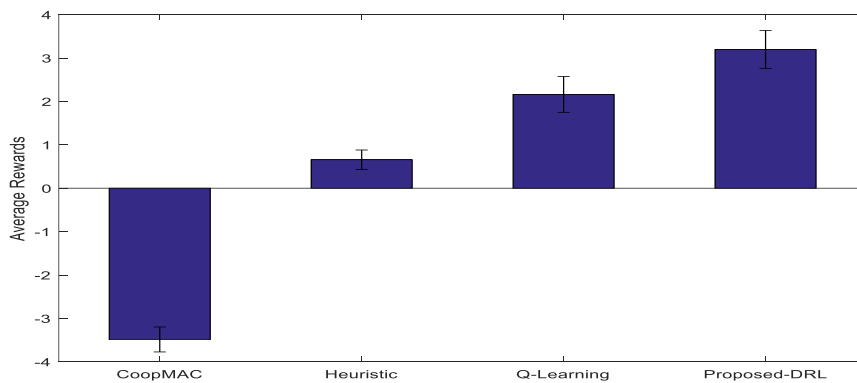


Figure (6) Comparison of average rewards of different algorithms for releasing data in the desired network



Throughput

Figure 7 shows the evolution of the energy efficiency of the GSM network for four different algorithms. The results show that the proposed DRL algorithm reaches the highest energy efficiency level very quickly and maintains this value stably at around 2.5; while the classical Q-learning stops at a lower level (around 2.05) despite improving over the non-intelligent methods. The heuristic method with an efficiency of around 1.56 offers only a limited improvement over the baseline method, and the baseline method without learning with an efficiency of around 1.21 has the weakest performance, indicating the inefficiency of the power constant adjustment and the absence of an adaptive decision-making mechanism. Overall, Figure 7 confirms the superiority of the proposed DRL algorithm in terms of convergence speed and energy efficiency.

Figure 8 also compares the average throughput of the same four algorithms and shows that the proposed DRL algorithm, achieving an average throughput efficiency of about 2.24, makes the best use of the radio resources of the GSM network while respecting energy constraints. This advantage is the result of utilizing the deep learning structure and using network graph features (degree and centrality of cells, neighborhood pattern and traffic load) in the decision-making process, which establishes a good balance between delay, throughput and energy consumption.

Figure (7) Comparison of the average throughput efficiency of different algorithms in the data upload process in the target network.

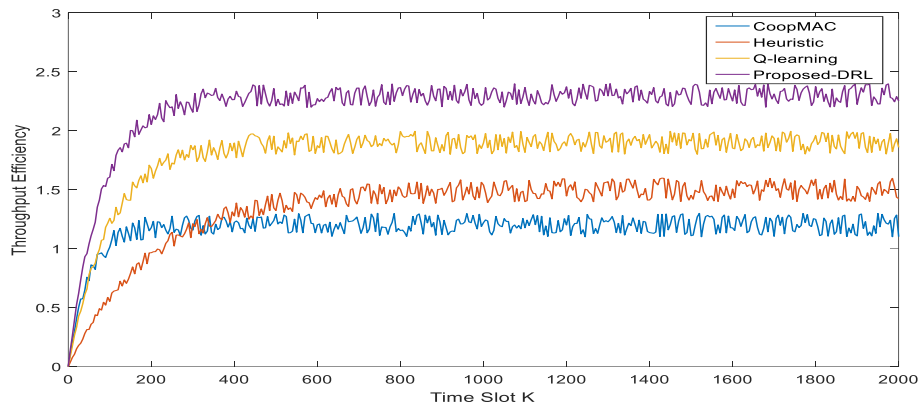


Figure (8) Comparison of the average throughput efficiency of different algorithms in the target network

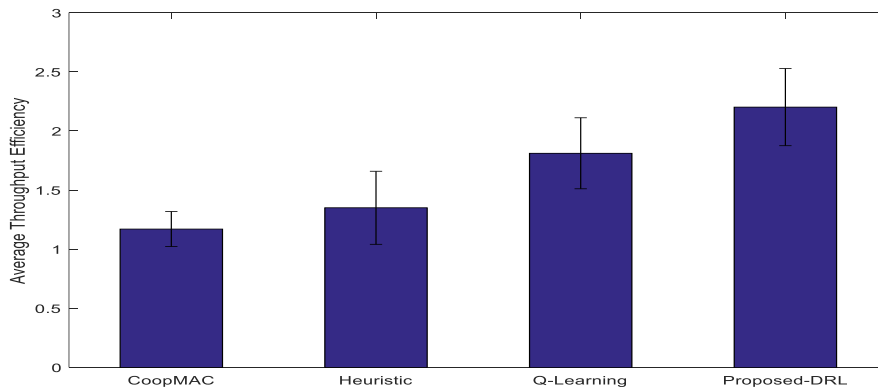
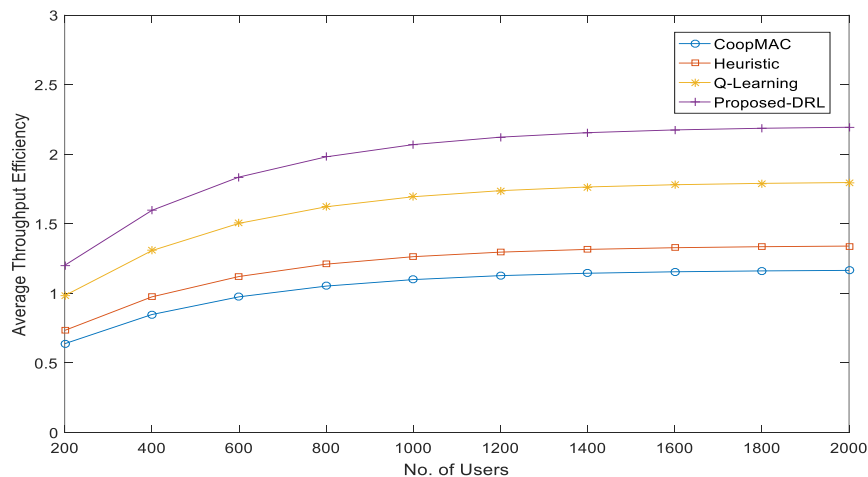


Figure 9 shows the comparison of the performance of the four algorithms against different numbers of users in throughput. In this figure, the proposed DRL algorithm consistently outperforms Q-learning, heuristics and baseline methods in the entire range of user numbers, from 200 to 2000 users. In particular, the DRL approach, due to the use of graphical modeling (Watts-Strogatz), the instantaneous load and channel quality of cells, and adaptive learning through a deep neural network, is able to adapt online to changes in user density and radio conditions and continuously increase the network throughput; so that in conditions of maximum user density, it reaches a value of about 2.23. In contrast, the Q-learning algorithm, despite improving compared to non-intelligent methods (up to about 1.8), does not reach the performance level of DRL due to the structural limitation in modeling a large and continuous state space. The heuristic and basic methods remain at about 1.4 and 1.2, respectively, and due to the lack of a learning and adaptation mechanism, they are unable to effectively use the network capacity in high-density scenarios. Overall, the analysis of Figure 10 confirms that the use of DRL in combination with cell graph modeling leads to maximizing network capacity utilization and maintaining QoS in dense GSM network scenarios.

Figure (9) Average throughput efficiency with different number of users



Energy efficiency

Figure 10 shows the energy efficiency trend for four different algorithms – the basic method, the heuristic method, Q-Learning and the proposed DRL algorithm – over a time horizon of 2000 control steps in the GSM network. In this figure, the energy efficiency is measured as the ratio of “total network throughput to base station energy consumption”. The results show that the proposed DRL algorithm reaches a stability level close to 2.5 after about 200 steps and maintains this level almost constant over time. This stable behavior and high energy efficiency value indicate that the DRL agent is able to select appropriate configurations of power adjustment, cell on/off and traffic routing between neighboring cells using network graph information (Watts-Strogatz graph), cell traffic load, link quality and cell power/state status; In such a way that while maintaining the QoS of users, the energy consumption of the network is effectively reduced. In comparison, the classical Q-Learning algorithm, although it provides acceptable energy efficiency, lags behind the DRL algorithm due to its reliance on the Q table and its weakness in generalizing to a large and continuous state space and converges to a lower level of energy efficiency. The heuristic method with an average efficiency of about 1.61 is unable to continuously and optimally adjust the power and

state of the cells due to the lack of a learning and adaptation mechanism, and as a result, it remains at a lower level than the reinforcement methods. The basic method without learning also records the lowest energy efficiency of about 1.31 and high fluctuations, indicating instability in decision-making and inefficiency in allocating energy resources in dynamic network conditions.

Figure (10) The trend of changing energy efficiency of different algorithms over time in the studied network

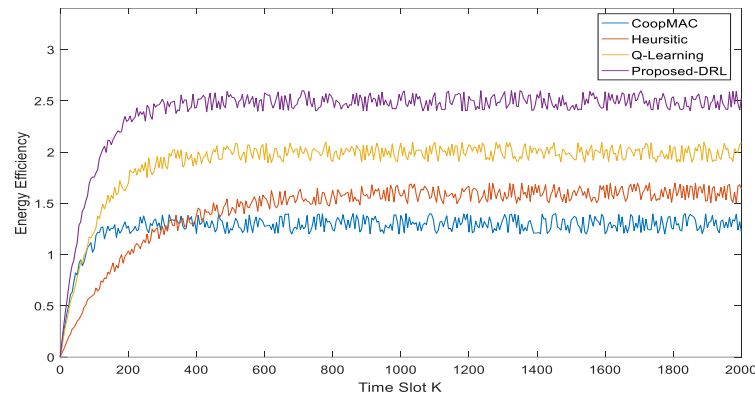


Figure 11 compares the average throughput of four different algorithms in a bar graph. This metric indicates the algorithm's ability to effectively utilize data transmission resources to achieve maximum network efficiency in data forwarding. The proposed DRL algorithm, with an average throughput of 2.23, performed much better than other algorithms. This success is the result of intelligent path selection and the use of the POMDP structure to adapt to changing network conditions. The Q-learning algorithm ranks second (average 1.91), while the Heuristic (average 1.3) and CoopMAC (average 1.2) methods rank next due to their lack of learning and adaptation. As a result, Figure 4-11 clearly shows that the proposed DRL algorithm offers the highest level of efficiency in terms of throughput due to its deep learning structure and intelligent decision-making capability at low network levels. This feature establishes it as a top choice for use in applications with high QoS and network efficiency requirements, such as 5G networks and industrial communications.

Figure (11) Comparison of average energy efficiency of different algorithms

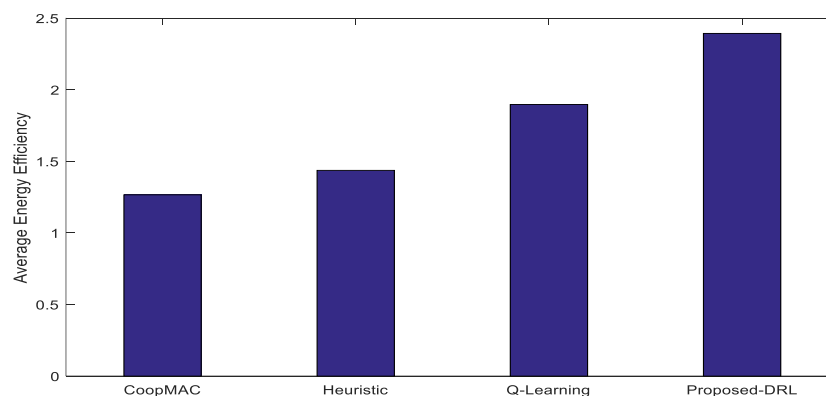
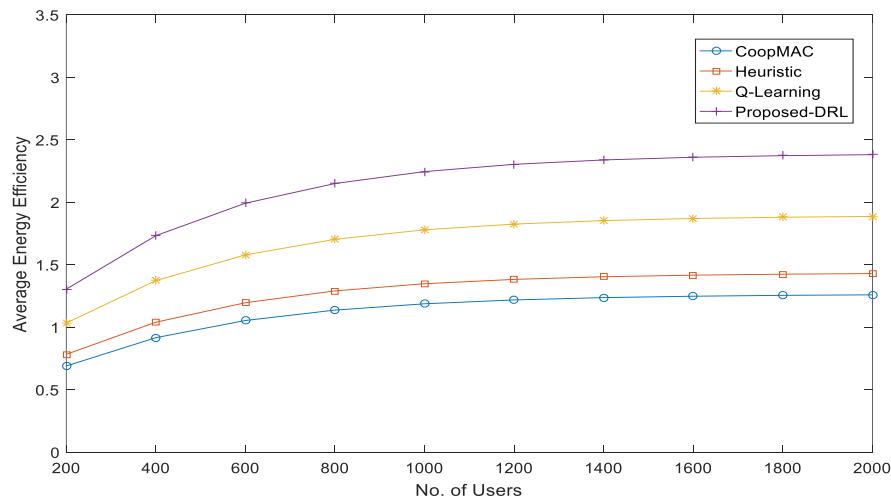


Figure 12 “Comparison of average energy efficiency at different numbers of users” shows the change trend of GSM network energy efficiency for four algorithms in the range from low to very high user density. As the number of users increases, all curves have an upward trend, but their slopes and final levels are different. The proposed DRL algorithm forms the highest curve at all points and continues to maintain its distance from other methods as the number of users increases; that is, even in very high-density scenarios, it can learn a suitable combination of power adjustment, cell on/off, and traffic routing between neighboring cells to consume less energy than the transmitted data. The Q-Learning curve is located below the DRL curve, indicating that classical reinforcement learning provides some improvement, but at higher densities it lags behind the proposed method in terms of energy efficiency. The heuristic method and then the basic method (CoopMAC) are at the lowest levels and, despite a slow increase in efficiency with the growth of the number of users, never reach the efficiency limit of the two learning methods; this shows their inability to adapt to the dynamic conditions of the network and to optimally use the graph structure of the cells.

Figure (12) Comparison of the average energy efficiency of different numbers of users

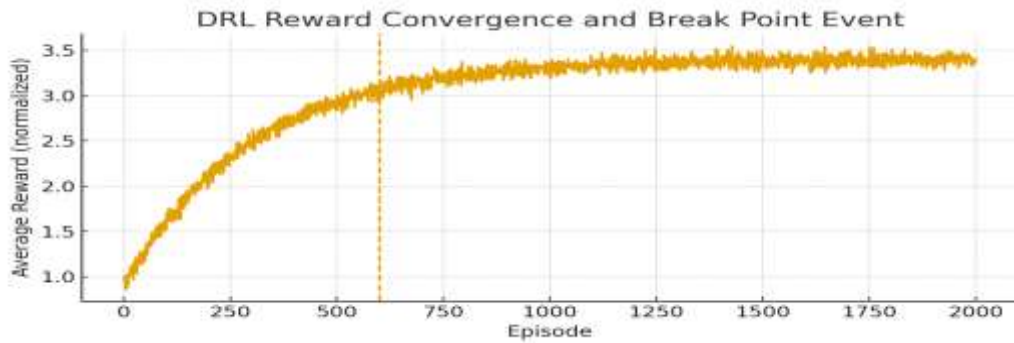


Reward curve analysis

Figure 13 shows the convergence behavior of the proposed DRL algorithm during 2000 training episodes. The horizontal axis shows the episode number and the vertical axis shows the “normalized average reward”; this reward represents the net performance gain (a combination of throughput and energy efficiency improvements) after deducting the penalties related to QoS violations and control overhead. At the beginning of the training process, the agent is in the exploration phase and by testing different values of transmitted power, on/off states of cells and traffic routing patterns between neighboring cells, it gradually learns to avoid actions that lead to high energy consumption or QoS degradation; for this reason, the average reward increases with a relatively steep slope. After a certain number of episodes, the curve enters a region where the increase in reward becomes very slow and oscillates around an almost constant value. The vertical dotted line in the figure marks the break point; that is, the point where further improvement in throughput or energy reduction is possible only by paying an additional penalty (in the form of increased control overhead or aggravated QoS violations) and is not very justified from the perspective of the total benefit-penalty.

In other words, after this point, the learned policy establishes a stable balance between increasing the efficiency of the GSM network and the energy and QoS costs, and subsequent changes in the curve mainly appear as small fluctuations around the optimal reward value.

Figure (13) Reward convergence curve of the DRL algorithm and the Break Point event for 2000 episodes



5-Conclusion

This study presents a comprehensive framework for simultaneous optimization of throughput and energy consumption in GSM networks, which consists of three main components: network topological modeling as a small-world graph based on the Watts–Strogatz model, structural and traffic feature extraction using graph exploration, and the design of a deep reinforcement learning agent for adaptive control of network parameters. In the modeling section, it was shown that appropriate adjustment of parameters k and p in the Watts–Strogatz model produces a graph with a high clustering coefficient and a short average path length that simultaneously reflects the neighborhood, interference, and handover relationships between cells and provides a suitable platform for learning on the network structure. Next, by defining the state space based on indicators such as cell degree and centrality, traffic load, link quality and power status, and designing the action space including power adjustment, cell on/off and traffic allocation policies, the problem was formulated as a Markov decision process with a multi-objective reward function. The simulation results showed that the proposed DRL algorithm has a significant advantage over the basic, heuristic and classical Q-learning algorithms in terms of both the final reward amount and the convergence speed; such that its average cumulative reward reaches about 3.41, while the best competing method (Q-learning) stops at 2.28. From the perspective of network metrics, the proposed method was able to maintain energy efficiency at a stable level of close to 2.5 and at the same time significantly increase the average network throughput compared to other methods. Analysis of the results for a wide range of user numbers (200 to 2000 users) showed that DRL was able to support traffic growth without significant QoS degradation even in high-density conditions by adapting online to the load and channel conditions. Overall, it can be concluded that combining the structural insight from small-world graphs with the flexibility of deep reinforcement learning provides an effective solution for energy and capacity management in GSM networks and can be used as a general approach in designing self-organizing policies for next-generation cellular networks, the Internet of Things, and smart industrial systems. From a future research perspective, developing multi-agent versions of the present approach, considering user mobility patterns, and incorporating real-world operational constraints of radio equipment could provide attractive avenues for extending this work.

References

- [1] M. Marwani, and G. Kaddoum, "Graph neural networks approach for joint wireless power control and spectrum allocation," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 717-732, 2024.
- [2] Y. Lu, Z. Zhang, X. Xu, L. Liu, Q. Fu, J. Chen, and C. Chen, "GTD3-NET: A deep reinforcement learning-based routing optimization algorithm for wireless networks," *Peer-to-Peer Networking and Applications*, vol. 18, no. 1, pp. 23, 2025.
- [3] S. Mclaughlin, P. M. Grant, J. S. Thompson, H. Haas, D. I. Laurenson, C. Khirallah, Y. Hou, and R. Wang, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 10-17, 2011.
- [4] C. A. Chan, W. Li, S. Bian, I. Chih-Lin, A. F. Gygax, C. Leckie, M. Yan, and K. Hinton, "Assessing network energy consumption of mobile applications," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 182-191, 2015.
- [5] L. Chen, Y. Xu, F. Xu, Q. Hu, and Z. Tang, "Balancing the trade-off between cost and reliability for wireless sensor networks: a multi-objective optimized deployment method," *Applied Intelligence*, vol. 53, no. 8, pp. 9148-9173, 2023.
- [6] T. Zhang, J. Cao, Y. Chen, L. Cuthbert, and M. El Kashlan, "A small world network model for energy efficient wireless networks," *IEEE Communications Letters*, vol. 17, no. 10, pp. 1928-1931, 2013.
- [7] M. Vasuki, A. D. Kumar, and R. Prabhakaran, "A Study on GSM Mobile Phone Network in Graph Theory," *International Journal of Current Research and Modern Education*, vol. 1, no. 1, pp. 772-783, 2016.
- [8] H. F. Song, and X.-J. Wang, "Simple, distance-dependent formulation of the Watts-Strogatz model for directed and undirected small-world networks," *Physical Review E*, vol. 90, no. 6, pp. 062801, 2014.
- [9] Z. Zhu, "Particle swarm optimization with Watts-Strogatz model." pp. 506-513.
- [10] M. Karimi, M. Harouni, E. I. Jazi, A. Nasr, and N. Azizi, "Improving monitoring and controlling parameters for Alzheimer's patients based on IoMT," *Prognostic models in healthcare: Ai and statistical approaches*, pp. 213-237: Springer, 2022.
- [11] A. A. Alaidany, and A. Lakizadeh, "Improving the Accuracy of Cancer Driver Gene Identification based on Dimensionality Reduction Using Deep AutoEncoders," *International Journal of Intelligent Engineering&Systems*, vol. 18, no. 9, 2025.
- [12] A. Ali A, M. Ali K, M. Marwah M, and F. Tibah, "A REVIEW OF MACHINE LEARNING IN BANKING RISK MANAGEMENT AND POSSIBLE RESEARCH TOPICS," *Journal of Engineering, Mechanics and Modern Architecture*, vol. 4, no. . ,1pp. 50-57, 2025.
- [13] M. Stasiak, M. Glabowski, A. Wisniewski, and P. Zwierzykowski, *Modeling and Dimensioning of Mobile Wireless Networks: From GSM to LTE*: John Wiley&Sons, 2010.
- [14] Y. Liu, and L. Li, "Efficient Graph Sequence Reinforcement Learning for Traveling Salesman Problem." pp. 256-267.
- [15] M. Z. Rafique, and M. Abulaish, "Graph-based learning model for detection of SMS spam on smart phones." pp. 1046-1051.
- [16] M. Karimi, Z. Karimi, M. Khosravi, Z. Delaram, M. H. Dehsheikhim, S. A. Najafabadi, M. A. Aliabadi, and N. Tavakoli, "Feature Selection Methods in Big Medical Databases: A Comprehensive Survey," *International Journal of Theoretical&Applied Computational Intelligence*, pp. 181-209, 2025.
- [17] S. Georgousis, M. P. Kenning, and X. Xie, "Graph deep learning: State of the art and challenges," *IEEE Access*, vol. 9, pp. 22106-22140, 2021.

- [18] O. Jouini, K. Sethom, A. Namoun, N. Aljohani, M. H. Alanazi, and M. N. Alanazi, "A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions," *Technologies*, vol. 12, no. 6, pp. 81, 2024.
- [19] A. Rahman, T. Debnath, D. Kundu, M. S. I. Khan, A. A. Aishi, S. Sazzad, M. Sayduzzaman, and S. S. Band, "Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities," *AIMS Public Health*, vol. 11, no. 1, pp. 58, 2024.
- [20] M. Harouni, M. Karimi, A. Nasr, H. Mahmoudi, and Z. Arab Najafabadi, "Health monitoring methods in heart diseases based on data mining approach: A directional review," *Prognostic models in healthcare: Ai and statistical approaches*, pp. 115-159: Springer, 2022.
- [21] S. Wen, B. Hu, and H.-K. Lam, "Reinforcement learning optimization for base station sleeping strategy in coordinated multipoint (CoMP) communications," *Neurocomputing*, vol. 167, pp. 443-450, 2015.
- [22] M. Ali, S. Qaisar, M. Naeem, and S. Mumtaz, "Energy efficient resource allocation in D2D-assisted heterogeneous networks with relays," *IEEE Access*, vol. 4 ,pp. 4902-4911, 2016.
- [23] A. Shahid, K. S. Kim, E. De Poorter, and I. Moerman, "Self-organized energy-efficient cross-layer optimization for device to device communication in heterogeneous cellular networks," *IEEE Access*, vol. 5, pp. 1117-1128, 2017.
- [24] J. Li, H. Wang, X. Wang, and Z. Li, "Optimized sleep strategy based on clustering in dense heterogeneous networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 290, 2018.
- [25] Y. S. Nasir, and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE Journal on selected areas in communications*, vol. 37, no. 10, pp. 2239-2250, 2019.
- [26] Z. Zhang, H. Qu, J. Zhao, and W. Wang, "Deep reinforcement learning method for energy efficient resource allocation in next generation wireless networks." pp. 18-24.
- [27] C.-K. Hsieh, K.-L. Chan, and F.-T. Chien, "Energy-efficient power allocation and user association in heterogeneous networks with deep reinforcement learning," *Applied Sciences*, vol. 11, no. 9, pp. 4135, 2021.
- [28] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Energy-efficient deep reinforcement learning assisted resource allocation for 5G-RAN slicing," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp.2021 ,871-856 .
- [29] A. Mughees, M. Tahir, M. A. Sheikh, A. Amphawan, Y. K. Meng, A. Ahad, and K. Chamran, "Energy-efficient joint resource allocation in 5G HetNet using Multi-Agent Parameterized Deep Reinforcement learning," *Physical Communication*, vol. 6 ,1pp. 102206, 2023.
- [30] H. Choi, T. Kim, S. Lee, H.-S. Choi, and N. Yoo, "Energy-Efficient Dynamic Enhanced Inter-Cell Interference Coordination Scheme Based on Deep Reinforcement Learning in H-CRAN," *Sensors (Basel, Switzerland)*, vol. 24, no. 24, pp. 7.2024 ,980
- [31] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Mobility aware and energy-efficient federated deep reinforcement learning assisted resource allocation for 5G-RAN slicing," *Computer Communications*, vol. 217, pp. 166-182, 2024.
- [32] V. Shilpa, and R. Ranjan, "Radio Resource Scheduling in 5G Networks Based on Adaptive Golden Eagle Optimization Enabled Deep Q-Net," *SN Computer Science*, vol. 5, no. 5, pp. 517, 2024.

- [33] H. Kim, and J. So, "Distributed Multi-Agent Deep Reinforcement Learning-Based Transmit Power Control in Cellular Networks," *Sensors*, vol. 25, no. 13, pp. 4017, 2025.
- [34] H. QU, Z. ZHU, J. ZHAO, R. TANG, L. WANG, and Z. CAO, "Energy-efficient joint relay selection and resource allocation scheme for mobile relay aided device-to-device communication," *电子与信息学报*, vol. 39, no. 10, pp. 2464-2471, 2017.