International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 310-327
https://doi.org/10.70917/ijcisim-2026-2709
Article

Enhancing Throughput and Energy Efficiency in GSM Networks Through Watts–Strogatz Graph Modeling and Reinforcement Learning

Imam Qasim Mousa, Kheirolah Rahsepar Fard

Department of Computer Engineering and Information Technology, University of Qom, Ghadir boulevard, Qom, Iran,imamqasim888@gmail.com, rahseparfard@gmail.com

Abstract: The problem of simultaneous optimization of both throughput and energy consumption in GSM network is considered as a multi-objective problem and studied in this paper. The cell topology is modeled as a "small world" graph based on the stochastic graph model of Watts and Strogatz, and properties of the graph such as degree, centrality, traffic load and QoS indices of each cell are extracted through exploration. With these features along with the channel information of each channel and the power level of the base stations, a deep reinforcement learning agent has the state space. The DRL agent learns the control actions (transmit power, ON/OFF of cells, routing traffic between adjacent cells) so that a reward function defined as normalized throughput, normalized energy consumption and penalty on QoS violation is maximized by using a neural network based Q-value function approximation. 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. Moreover, the energy efficiency of the network, i.e., ratio of total throughput to total power at the base-station, for the proposed method is kept at a stable value of 2.5, whereas Q-learning has a value of 2.05, and the two non-intelligent methods have values of 1.56 and 1.21, respectively. The obtained results show that the small-world graph modeling and deep reinforcement learning can be an efficient framework for intelligent energy and capacity management in GSM networks.
Keywords: GSM cellular networks, energy efficiency, throughput optimization, Watts–Strotz small-world graph, deep reinforcement learning

1. Introduction

Since the last few decades, cellular networks using the GSM standard have been the primary infrastructure used to deliver mobile telecommunications services in many countries and still are the key infrastructure for a large share of voice and data traffic [1]. As the number of users grows, the range of the services grows and the need for quality of service (QoS) grows, the problem of optimal utilization of radio and hardware resources of these networks becomes a major challenge. However, the rise in energy use within base stations entails considerable environmental effects and may contradict the objective of green and sustainable networks [2] as well as high costs for the operators. According to estimates, a significant portion of energy use in cellular networks is dedicated to base stations, so any energy saving solution in base stations can directly benefit in terms of cost reduction and in reduction of greenhouse gas emissions [3].

In such cases, the problem of optimizing throughput and energy consumption in a GSM network has become a multi-objective problem, with particular importance. The optimization problem in the GSM networks has been addressed in most of the classical studies with a single objective or with a focus on the major metric, say, capacity maximization, call drop minimization or interference minimization [4]. In these methods they typically assume that other criteria are given and/or implicitly and indirectly taken into account. In real life, however, the operator must come up with a dynamic solution to balance the conflicting goals, such as higher throughput and lower energy usage. For instance, raising the transmission power will increase the signal quality and consequently the throughput, however, it will increase both energy consumption and inter-cell interference. But decreasing power and turning off some equipment may also save energy; if not managed properly it will compromise the QOS of users. For this reason, multi-objective optimization techniques which can systematically and intelligently handle this trade-off are required [5].

A proper model of the network structure and dynamics is one of important prerequisites for the design of intelligent algorithms in cellular networks. Paralleling the network of the GSM in this respect, the network can be modeled as a graph having “small-world” characteristics, which would present a more realistic representation of the relations between the cells [6]. A number of studies have demonstrated that many real networks, such as social or communication networks, share similar structural characteristics with small-world networks, high clustering coefficient and short average path length between network nodes [7]. Such features are also seen in cellular networks: each cell has a number of close neighbors, and a few cells further away can be connected to the former ones by a few close cells. Therefore the application of stochastic graph models like the Watts–Strotz model of a small-world network to approximate the topological structure of cellular networks is very attractive [8].

The Watts–Strogatz model is a popular example of how to create a small-world random graph by first wiring each node to a small number of nearby nodes in a regular manner, and then rewiring links with a certain probability to connect them to other nodes at a larger distance away. The outcome of this process is a graph that has relatively high local clustering, while the average path length is small. Assuming cells or base stations of a GSM network as nodes of this graph, the edges may be interpreted as neighborhood relationships, the possibility of handover or dominant paths that users traverse between cells. By this method, the Watts-Strogatz model can be used to realistically simulate cell-to-cell communication and investigate network dynamics [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. These features can be analyzed on the Watts–Stromata graph of the GSM network to identify critical cells, high traffic areas and energy-sensitive points[12]. This information can be extremely useful in the design of resource managements policies and parameters as input to a learning or optimization algorithm. Recently reinforcement learning (RL) and particularly deep RL has been proposed as one of the new tools for solving complex sequential decision making problems in telecommunication systems. Within this paradigm, the learning agent learns control policies by direct interaction with the environment with the goal of maximizing a long term reward function [13].

One of the most salient aspects of RL is that it does not need to have a detailed analytical model of the environment and can learn optimal or quasi-optimal policies from experience and observed data in a scenario where explicit system modeling is challenging or costly. For cellular networks with high complexity, uncertainty, complex graph structure, and time varying traffic, this feature makes reinforcement learning very appealing for application [14-16.[

Although the application of reinforcement learning to resource management in next generation networks has grown significantly, there are still some important gaps in the area of GSM networks [17]. First, most of the previous works are concentrated on one of the objectives, for example capacity maximization or coverage enhancement, and the issue of simultaneous optimization of throughput and energy consumption for a multi-objective framework has not been studied. Secondly, the network structure in many studies is highly simple and regular or the general graph models are applied without taking account of the small world properties. Third, there is less research on the coherent combination of graph modeling based on the Watts–Strogatz model, graph exploration and reinforcement learning to extract structural features which can be used in designing optimal .[control policies[18-20

The present paper, entitled “Simultaneous Optimization of Throughput and Energy Consumption in GSM Networks Using Graph Modeling, Graph Exploration and Reinforcement Learning”, aims to fill these gaps. In the proposed approach, first, the GSM network is modeled as a small-world stochastic graph that is defined by the number of local neighbors of each cell and the probability for re-drawing edges of the network, according to the Watts–Strogatz model. Then, via graph exploration techniques, a set of structural and traffic characteristics is extracted from this graph that will well describe the state of each cell, and its function in the network. These, in addition to traffic information and current control parameters, constitute the state space of the reinforcement learning agent.

A multi-objective reinforcement learning (RL) framework is then designed and the agent attempts to maximize a reward function consisting of maximization of network throughput and minimization of energy consumption by dynamically and intelligently varying the parameters of the network, such as the base station transmission power and the radio resource allocation pattern. The proposed reward function is designed to act as a combination of throughput and energy criteria and also take into account the user QoS constraints requirements as a penalty; such that if the user QoS constraints are not met, the reward is decreased, which steers the agent towards acceptable policies as per the service provision. The learning agent learns to set up the right trade-off between conflicting goals and the policies that will enhance the performance of the network in the long run.

The key 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 enrichment of the state space by using graph exploration, second, the design of a multi-objective reinforcement learning framework that can at the same time optimize throughput and energy use, and third, an extensive evaluation of the proposed method in simulated scenarios and comparison with conventional methods of fixed tuning and simple control rules. It is hoped that the results will demonstrate that a fusion of the structural information from the small-world graph and the adaptability of reinforcement learning will result in network performance improvements in both throughput and energy efficiency.

The organization of the rest of the paper is as follows. The next section gives a detailed survey of the previous works on energy and throughput optimization in cellular networks, graph modelling using small-world networks and reinforcement learning applications for resource management. In the third part, the system model and the modeling method for the GSM network that is based on the Watts-Strogatz model are introduced, as well as the methods of exploring the graph. The fourth section focuses on setting up the multi-objective optimization problem and clarifying the key ingredients of the reinforcement learning framework. The results of a simulation and comparison of the performance of the suggested method will be given in the fifth section and the last section will be dedicated to a conclusion and directions for further research in this field.

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 select the sleep strategy of cells by multi-stage Q-learning. The objective: save BS energy and ensure downstream performance and quality of service. The learned policy is shown to simultaneously yield energy efficiency and throughput efficiency gains over simple threshold policies through simulations.

For HetNet equipped with relays and D2D communication, Ali et al. (2016) [22] solve the problem of maximizing the ratio of network throughput to power consumption as a nonlinear fractional programming. They apply the Charnes–Cooper transform to the problem and the Outer Approximation algorithm to get the ε-optimal solution for power, spectrum allocation, and cell selection, and demonstrate that their EE is much better than the reference schemes.

Shahid et al. (2017) [23] propose a self-organizing mechanism for various layers in a heterogeneous D2D based network for energy saving and spectral efficiency, which regulates the physical and MAC layers jointly. They demonstrate that the energy per bit can be lowered while keeping capacity and QoS by modelling the user-cell interactions and developing adaptive policies in a game-play.

In dense heterogeneous networks, Li et al. (2018) suggest an optimal sleep schedule, in which small cells are grouped into clumps in accordance with an interference graph. The nodes of the graph represent BSs and edges indicate strong interference. The energy can be saved by shutting clusters which are under-loaded and reconfiguring the coverage under QoS constraints in order to maintain the total throughput and user coverage [24.[

In the multi-cell network, Nasir&Guo (2019) describe dynamic power control as a multi-agent reinforcement learning problem where each link is an agent that learns reinforcement. The proposed algorithm optimizes the total network rate in the presence of interference with the use of a joint DNN architecture and simultaneous updating. The results demonstrate that this method can reach higher throughput with the lower power consumption for the targeted throughput than classical optimization methods like 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 abstracted as an MDP and the DQN agent is trained using the current load, channel and power state as an observation. It is seen that the proposed scheme exhibits higher EE with same or better aggregate rate than the convex optimization-based schemes [26].

Hsieh et al. (2021) introduce a DRL framework to optimize the power allocation and user association in heterogeneous networks. The state consists of the load of each BS, link quality and channel state, and the DQN agent chooses the BS and power level for each user to maximize the overall energy efficiency. Based on simulations the proposed method is capable of achieving a higher EE and a satisfactory user rate at the same time in comparison to heuristic algorithms [27.[

The resource allocation problem in 5G RAN slicing is studied in (Azimi et al., 2022), and EE-DRL-RA algorithm is proposed. In this framework, coarse-scale DL network specifies the amount of the resources for each slice, while a fine-scale DRL agent allocates power and radio resources to the 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.[

To tackle the joint user association and power allocation problem in 5G HetNet, Mughees et al. (2023) propose a multi-agent parametric DRL (MA-PDRL) approach. The power and user association action space is modeled as continuous and discrete respectively and represented using parametric actions. The results demonstrate that the MA-PDRL method provides better energy efficiency, faster aggregation rate, and more users with acceptable QoS compared to WMMSE, non-cooperative games and traditional DRL [29.[

Choi et al. (2024) propose a dynamic and energy-efficient eICIC scheme for H-CRAN networks, where the DRL agent optimizes scheduling and power parameters for ABSFs based on the interference level and network load. Reward function is the merit of energy efficiency and quality of service. According to simulations, energy consumption is lowered, while spectral efficiency and user throughput are enhanced for the energy-efficient solution, which is implemented with simulations compared to a static eICIC [30

Azimi et al. (2024) present a federated deep reinforcement learning (f-DRL) approach for learning resource allocation policies in 6G networks, where a base station (BS) trains resource allocation policies in a decentralized manner and only shares gradients with the central server. Reward function design takes into account energy efficiency and quality of service on mobile users, and the results indicate that this design can decrease signaling overhead and energy consumption and retain the accuracy of the model [31

In this paper, the radio resource scheduling problem in 5G networks is explored by applying DRL. Traffic queue, channel conditions, and service priority are included in the state of the system and the DRL agent chooses users and distributes RBs to maximize the overall rate while satisfying the delay constraints. The study demonstrates that the proposed method has better throughput and latency than the classical schedulers like PF and MLWDF [32

Kim&So (2025) suggest an algorithm for power control in cellular networks based on distributed multi-agent deep reinforcement learning (DMDRL), where each BS is considered a local DRL agent. The agents learn a policy with limited information exchange which reduces inter-cell interference, ensures link quality and lowers power consumption. The proposed technique is seen to improve the EE of the network for multi-cell scenarios, while not substantially impacting the user throughput [33

In a relay-based D2D network with the goal of maximizing energy efficiency while achieving throughput, QU et al. (2025) study the relay selection and resource allocation problem. They demonstrate that network EE and D2D user capacity can be greatly enhanced by jointly optimising power and bandwidth allocation together with intelligent relay selection [34] using analytical modelling and optimization algorithms.

3. Proposed method

The proposed method's system model is shown in figure (1). The physical network, comprised of macro cells, small cells and mobile users linked via wireless connections, is illustrated on left, while the abstract graph representation of the same network, with the BTSs represented as nodes and interference and handover relationships represented as edges, is shown on the right. The middle arrow is the one of "graph modeling", in which the complex physical network is converted into a simpler graph in order to perform structural analysis, and also to feed the reinforcement learning module for simultaneous optimization of throughput and energy .

The overall process of the proposed method of simultaneous optimization of throughput and energy consumption in GSM networks is presented in figure (2). 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. Following that, from this graph, the characteristics of every cell, degree, centrality, traffic load, and transmission power and quality of service (QoS) indices are derived and provided as a state vector for the reinforcement learning module. The reinforcement learning agent decides on control actions (power regulation, carrier on/off and user acceptance threshold adjustment) based on the observed states and eventually reaches a policy for a balanced throughput/energy tradeoff using a reward function designed as a combination of normalized throughput, normalized energy consumption and penalty for QOS violation. Lastly, it is evaluated by several metrics including the total throughput, energy efficiency, percentage of satisfied users whose QoS is acceptable, and the call drop rate, and the results are employed to fine-tune the parameters of the graph model and the reward function.

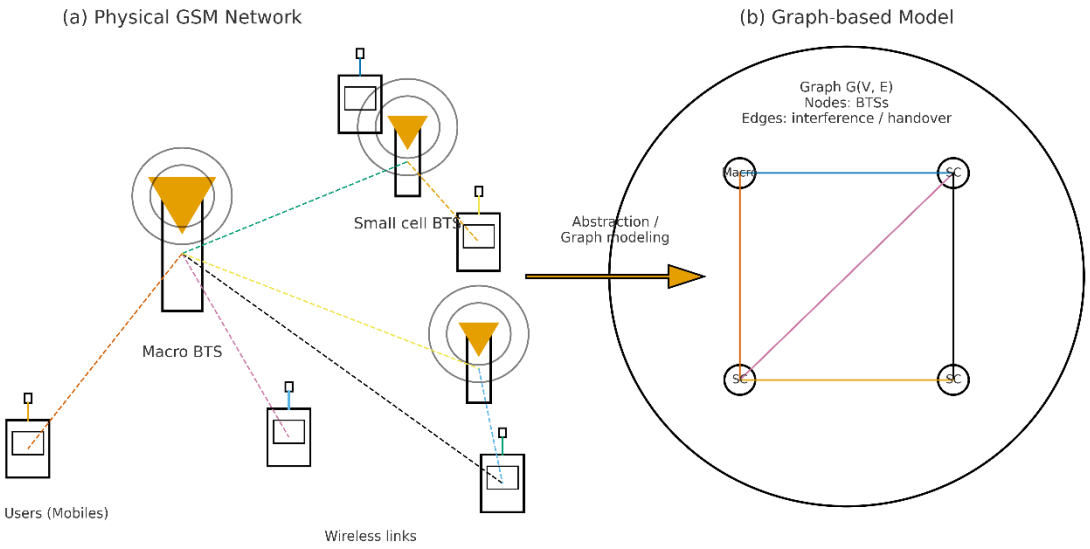

Figure (1) System model of the proposed method

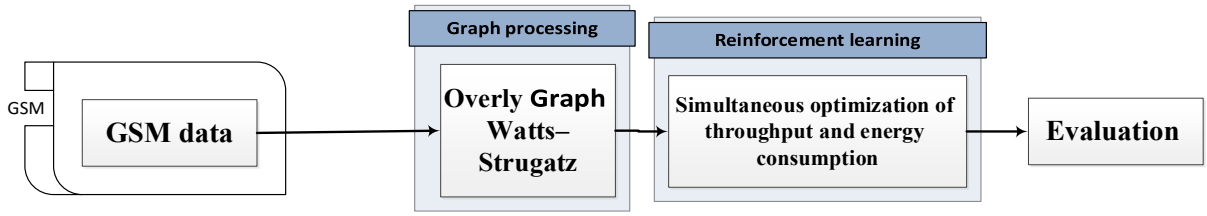

Figure (2) Block diagram of the proposed method

In the proposed method, the communication between the cells of the GSM network is considered as a small-world random graph as per the Watts–Strogatz model. In this model, each cell is first connected to its k nearest neighbors (in a regular circular graph), before some of the edges are randomly drawn again with probability p, resulting in a high clustering coefficient and short paths between more distant cells. The network topology adapts to the density of the cells in the neighborhood and how often the cells communicate with each other, resulting in a topology that is realistic and represents the neighborhood relationship and interference between the cells, while at the same time offering a suitable platform for reinforcement learning. Through this structure, learning based on the low-energy control strategy for power adjustment, cell management and serving cell selection is enabled for the learning agent, and thereby, the network throughput and energy efficiency of the base stations is simultaneously improved. This graph is built as in Figure 3.

.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.

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

3.1 Dynamic Learning with Cellular Automata

The network goes into dynamic learning mode after the modeling of the communication between the cells from the Watts–Strogatz graph. The purpose of this phase is to modify the behaviour of cells over time to trade off energy consumption with communication stability and to minimize the amount of unneeded communication traffic. During this stage each cell (or cell cluster) is represented as a "learning agent", with a set of control actions including direct transmission to the base station, routing through neighboring cells, transmission power adjustments and functional state changes (e.g., from active to semi-active). Each cell will choose at each time step from one of these actions, depending on the state of the network, and adjust their policy on how to make the decision in the next step based on the feedback they receive from the network – including reward and penalty. The reward function is defined in accordance with the criteria of energy consumption, link quality and the effective distance to the BS and the local traffic load such that actions that result in less energy consumption and better quality data transmission are given 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. This allows cells to learn the best communication and operating strategies as a self-organized and distributed mechanism in the context 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 $fi \in F = \{f_1 \dots f_n\}$ and n is the total number of cells. Also, $y \in Y$. Let $c: F \to R$ be the mapping function of the chosen path f to the desired search cost c(f). Consider the states $\tilde{S} = (\mathrm{x}, \mathrm{y}, \tilde{F})$ where

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

In (1), the power of the set F is indicated with the power notation ψ(F), $\tilde{F}$ denotes 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 set S contains the observed state s, so s ∈ S is a set of (x(i), f_i) that we will be given for each cell that we select.

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

One of the possible learning activities of the chosen cell is the activity a∈A such that a a∈A= Ac∪A_f, i.e., activity A_f=F or activity Ac=y. If the control action Ac is taken, the part ends and the agent gets a reward of zero. If it is classified correctly, or in the later cases, it is -1, then Af is selected (released) with the corresponding value, and the agent is rewarded with a negative reward -λ c (f_i). Only the existing states, not yet explored before, are included in the set of present control states or stability control activities in the present set. In these definitions, the reward function, r: $\tilde{S}$ ×A →R is defined as (3)

$$r\left((x, y, \tilde{F}), \alpha\right) = \begin{cases} -\lambda\, \mathrm{C}\,(f_i) & if\ \ a \in A_f, a = f_i \\ 0 & \text{if } a \in A_c \text{ and a} = \mathrm{y} \\ -1 & \text{if } a \in A_c \text{ and a} \neq \mathrm{y} \end{cases} \tag{3}$$

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 S0, 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):

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

This property makes the desired environment for reinforcement learning an inherently segment with a maximum segment length of |F|+1. Now the task is to find a policy π and select the available suitable paths from F sequentially and finally choose the desired cells from y. This operation must be undertaken so that the desired reward and expectation is maximised.

The location of an agent will only affect the observed states s, and will not be a good indicator of which state S˜ has been chosen. This means that for each agent, not only their reward function is chosen at random, but they also select a transfer function at random despite the environment being deterministic. To overcome this weakness, the action of each agent will be based on probabilities. Thus, optimization will be carried out 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 ×A →Δ(SUT). The reward function r(s,a) will also tell us the value of the reward we will receive, if we go to location a, when we are in state s.

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)

$$Q^{\pi}(s, a) = r(s, a) + E\left[\gamma\, Q^{\pi}\left(S, \pi(S)\right)\right] \tag{6}$$
$$s \sim t(s, a)$$

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.

$$\pi^{*}(s) = argmax_a\, Q^{*}(s, a) \tag{7}$$

Which in this regard

$$Q^{*}(s_1 a) = r(s_1 a) + E\,(Y\, max\, Q^{*}\ (s', a') \tag{8}$$
$$S \sim t\,(s, a$$

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

$$\pi^{\theta}(S) = argmax_a Q^{\theta}(s, a) \tag{9}$$

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) s~ 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.

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

In (10) q is the estimate of the objective function Q, a constant. If values of w, r, and t: it is constant. In fact, the parameters r, w and t do not change in the optimization step.

q(r,s)=r+y max⟦Q^θ ⟧ (s,a) ((11

The convergence of the expected function Q^θ to the expected function Q* will be stronger as the error value grows.

To summarize, at this stage the model is presented such that reinforcement learning enables the most optimal decisions about the cells within the network, based on experience and feedback from the environment. Each node (Learning Agent) is aware of the current state of the network and considers the different routes to take and selects the action that will generate the maximum anticipated reward. This process repeats until the algorithm eventually reaches the optimum policy π*; then, there will be more and more paths chosen with a lower cost and more and more stable. The value function Q(s,a) and its continuous improvements based on the rewards that have been received make each node's decisions more optimal and give a balance between accuracy, cost and stability. The pseudocode for implementing the steps of the reinforcement learning process for selecting the best path is given in the Figure 4:

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

Input:

D = (x, y) // samples: features x, true label/target y

$fi \in F = \{f_1 \dots f_n\}$ // candidate cells/routes

c: F → ℝ // cost of candidate fi

λ > 0 // cost–accuracy trade-off

γ, α_lr, ε // discount, learning rate, exploration

R, B // number of episodes, replay buffer size

Qθ(s,a), Qθ¯(s,a) // Q-network and target Q-network

State&Actions:

Latent state: S̃ = (x, y, F̃) ∈ X × Y × ψ(F)

Observable state: s = Φ(S̃) // features of selected candidates

Actions: A = A_f ∪ A_c

A_f: select candidate fi ∈ F̃ // continue exploration

A_c: commit final decision ŷ ∈ Y // terminate

Reward:

if a ∈ A_f and a = fi: r = − λ · c(fi)

if a ∈ A_c and ŷ = y: r = 0

if a ∈ A_c and ŷ ≠ y: r = −1

Procedure RL_Training

Initialize θ (and $\theta^{-} \leftarrow \theta$), replay buffer $D_{replay} \leftarrow \emptyset$

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 ← false

while not done do

// 1) ε-greedy action selection

with prob ε: a ← random feasible action

else: a ← argmax_a $Q\theta(s,a)$

// 2) Environment response

if a ∈ A_f (select fi ∈ $\tilde{F}$):

$r = -\lambda \cdot c(fi)$

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 ∈ A_c:

$\hat{y}$ = 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_replay; keep at most B samples

// 4) Q-update

Sample minibatch $\mathfrak{B} \subset D_{replay}$

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 θ by minimizing:

$L(\theta) = (1/|\mathfrak{B}|) \Sigma_k [y_k - Q\theta(s_k, a_k)]^2$

Periodically set $\theta^{-} \leftarrow \theta$

$s \leftarrow s'$

if terminal_flag: done ← true

end while

$\varepsilon \leftarrow \max(\varepsilon_{min}, \varepsilon \cdot decay)$

end for

Output: policy $\pi^*(s) = \text{argmax\_a } Q\theta(s,a)$

End Algorithm

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

4. Results

The network configuration which is simulated for the purpose of evaluating the proposed method is shown in table (1), with 12 GSM micro cells and one macro base station and number of mobile users is set to be 200 and 2000. The range of the micro cells is designed to be 100-300m, while the range of macro cell is designed to be 500m, which will form 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. The Deep Reinforcement Learning (DRL) agent configuration is set with the Deep Q-Learning algorithm and a neural network with two hidden layers each having 64 neurons, a playback buffer of 10,000 samples, a learning rate of 0.001, a discount factor of 0.95, an initial discovery rate of $\varepsilon = 0.9$, which decreases over time, and a number of training episodes between 500 and 1000, each with an episode length of 100-200 steps. Lastly, the state space is defined, such that it contains the parameters SNR, RSSI, aggregation index (CG) and user density, and is modeled as continuous and normalized; the action space is designed as discrete, and consists of different access/path options, including LTE link selection, Wi-Fi direct connection, and Wi-Fi connection with relay R_m; with the above structure, it will be possible to develop an intelligent and adaptive decision-making process for 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

The proposed Deep Reinforcement Learning (DRL) algorithm performs the most stable and efficient behavior in the dynamic environment of the GSM network compared with other methods. 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. This ability leads to the cumulative reward being achieved during the training process and to achieving a proper trade-off in the throughput, energy consumed and compliance to the QoS constraints.

In comparison with the classical Q-learning algorithm, which is a very powerful method that is much better than non-intelligent methods, but still weaker than DRL in terms of the stability and final reward value, is less stable and less sensitive to changes of the environment, especially in complex cellular network scenarios, convergence to optimal solutions is slower and less robust. The heuristic based heuristic method only marginally improves over the baseline method in the initial stages but eventually reaches an approximately fixed reward level because it cannot adjust to the evolution of traffic and cell states and cannot surpass a certain limit. Finally, the baseline method without learning (e.g., fixed power configuration, and serving cell selection based on classical criteria) exhibits the poorest performance, and has the largest rewards with fluctuations over time, suggesting that the approach is not suitable to handle dynamic and changing network conditions and to find a suitable balance between energy saving and quality of service.

The average cumulative rewards over the whole training process of the four algorithms (the base method, the heuristic method, Q-learning and the proposed DRL) are plotted as a bar graph in Figure 5. The results are clearly the significant superiority of the proposed DRL algorithm that achieves the best performance in terms of the combination of energy, QoS, throughput and 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 value of about -3.55. From the analysis in Figure 6, it is also concluded that the proposed DRL algorithm has better convergence speed and better stability of control behavior than the other algorithms, in addition to being better than other algorithms in terms of the amount of reward received at the end. The results validate the appropriateness of the DRL framework based on graph modeling for real cellular networks with varying environments, and its applicability to more complex systems such as the Internet of Things and smart industrial systems in the future.

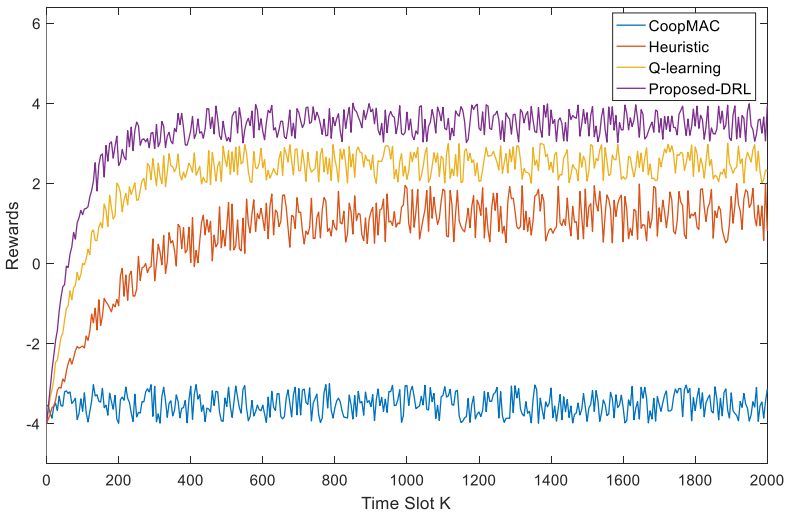

Figure (5) Comparison of cumulative rewards of different algorithms over time for choosing data upload paths in the network

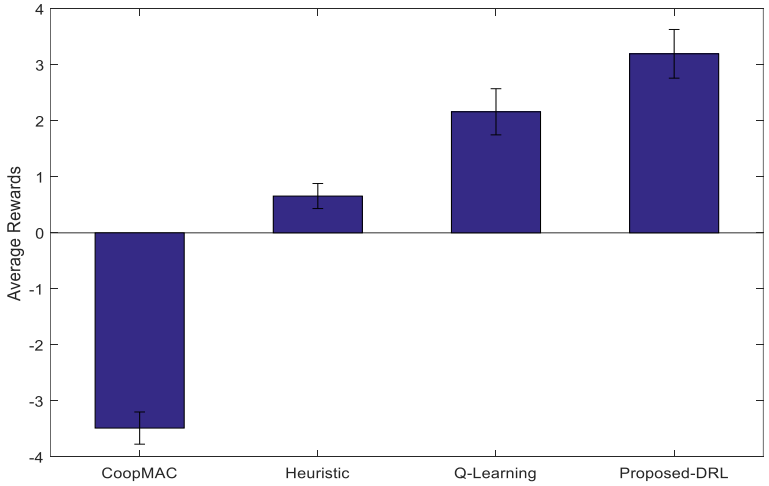

Figure (6) Comparison of average rewards of different algorithms for releasing data in the desired network

Throughput

Figure 7 illustrates the evolution of energy efficiency of the GSM network for four different algorithms. In terms of results, the proposed DRL algorithm achieves the best level of energy efficiency in a few trials and remains at this level quite stable (around 2.5), whereas the classical Q-learning algorithm improves over the non-intelligent ones, but it does not reach the same level of energy efficiency (around 2.05). The heuristic method with an efficiency of approximately 1.56 does not greatly improve upon the baseline method and the baseline method without learning with an efficiency of approximately 1.21 has the lowest performance, suggesting the inefficiency of the power constant adjustment and the lack of an adaptive decision making mechanism. Overall, as shown in Figure 7, the proposed DRL algorithm is superior in terms of the speed of convergence and energy saving.

Finally, the average throughput of the same four algorithms is shown in figure 8, and the proposed DRL algorithm can be seen to be most efficient in using the radio resources of the GSM network with respect to 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.

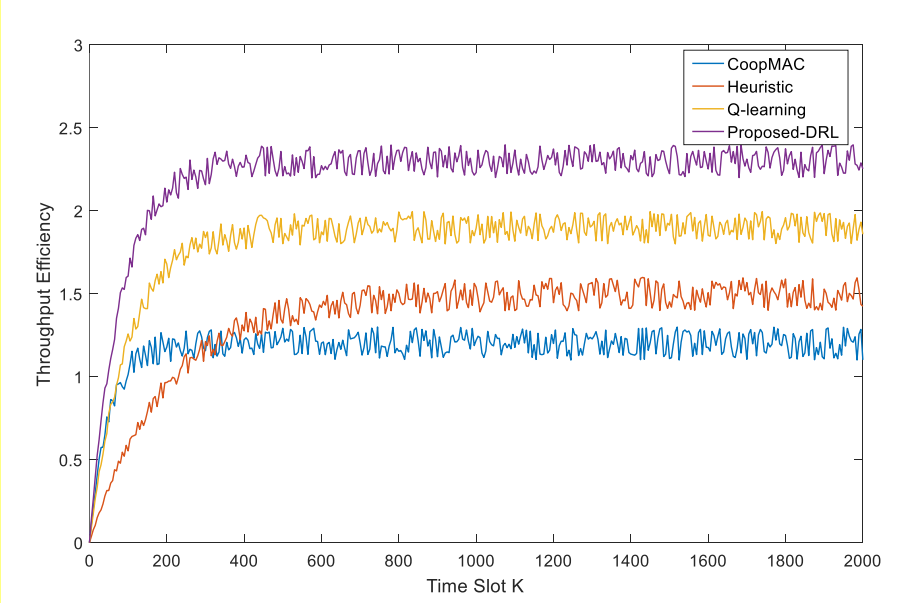

Figure (7) Comparison of the average throughput efficiency of different algorithms in the data upload process in the target network.

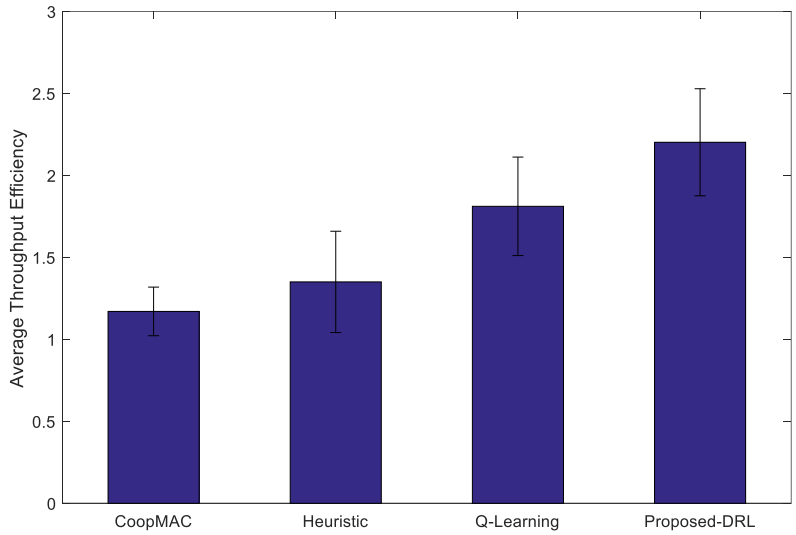

Figure (8) Comparison of the average throughput efficiency of different algorithms in the target network

The results in terms of throughput are shown in Figure 9 where the performance of the four algorithms are compared with varying number of users. As shown in this figure, the proposed DRL algorithm consistently outperforms Q-learning, heuristics and baseline methods, across the full range of users, from 200 to 2000 users. Specifically, the DRL approach, which employs graphical modeling (Watts-Strogatz) and instantaneous load and channel quality of cells, and adaptive learning via deep neural network, is capable of adapting online to changes in user density and radio condition, and continuously boosting network throughput up to 2.23 in the maximum user density condition. The Q-learning algorithm, which has been enhanced over non-intelligent approaches (up to approximately 1.8), has not yet attainable DRL performance because of the modeling constraint of a large and continuous state space. The heuristic and basic methods are still approximately 1.4 and 1.2 respectively and as there is no learning and adaptation mechanism, they cannot effectively utilize network capacity in high-density scenarios. However, on a general basis, the results of the analysis in Figure 10 indicate that the use of DRL along with cell graph modeling results in the maximization of network capacity utilization and maintenance of QoS in dense GSM network scenarios.

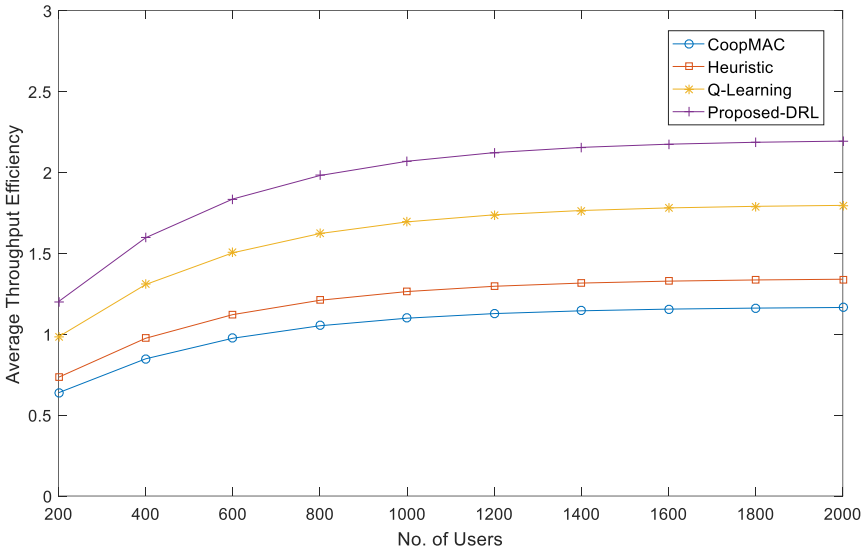

Figure (9) Average throughput efficiency with different number of users

Energy efficiency

The energy efficiency trend for the four different algorithms – the basic method, heuristic method, Q-Learning and the proposed DRL algorithm – is displayed in Figure 10 within a time horizon of 2000 control steps

in the GSM network. The energy efficiency is defined as the ratio of "total network throughput to base station energy consumption" in this figure. The results indicate that the proposed DRL algorithm attains almost 2.5 level of stability starting at around 200 steps and then stays nearly constant in time. This stable behavior and high energy efficiency value shows that the DRL agent can select an energy efficient and proper power adjustment configuration, cell on/off and traffic routing between adjacent cells with the information on the network graph (Watts–Strogatz graph), cell traffic load, link quality and cell power/state status; In such a way that energy consumption of the network is effectively reduced while maintaining the QoS of users. In contrast, the classical Q-Learning algorithm, which uses the Q table, has acceptable energy efficiency but it is less powerful than the DRL algorithm in its ability to generalize to a large and continuous state space and it has lower energy efficiency. Heuristic method with average efficiency of approximately 1.61 cannot continuously adapt the power and state of the cells optimally as it does not have a learning and adaptation mechanism, hence, it is not at its optimal level as compared to the reinforcement methods. Without learning the basic method records very low energy efficiency of approximately 1.31 and high fluctuations, which shows that the method is unstable in the decision making process and inefficient in allocating energy resources in dynamic network situations.

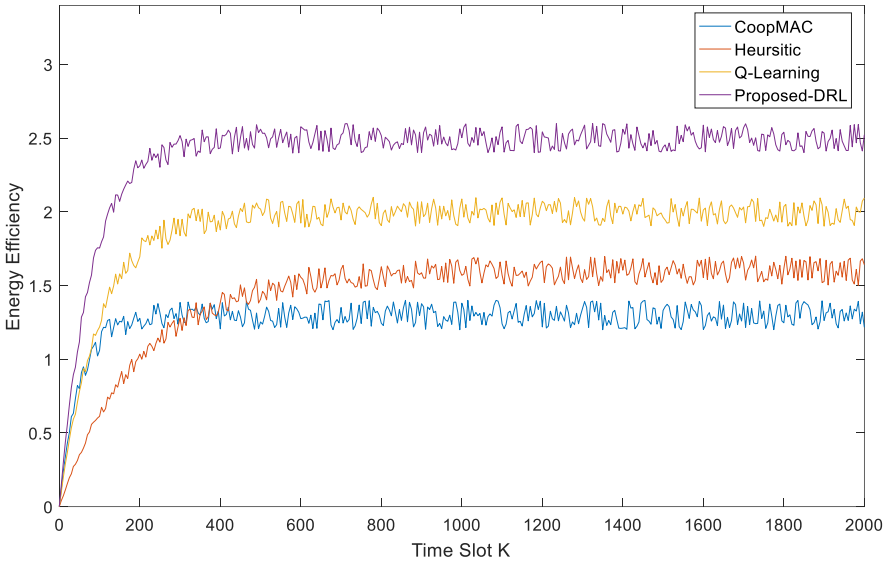

Figure (10) The trend of changing energy efficiency of different algorithms over time in the studied network

The average throughput of four different algorithms is compared in a bar graph as shown in Figure 11. This is a measure of how well the algorithm can make use of the data transmission resources to achieve maximum network efficiency in data forwarding. The proposed DRL algorithm outperform other algorithms with the average throughput of 2.23. This success is achieved by intelligent path selection and adapting to change by using the POMDP structure. The Q-learning algorithm ranks 2nd (average 1.91), followed by the Heuristic (average 1.3) and CoopMAC (average 1.2) methods because they have no learning and adaptation. From this, it is evident from the Figure 4-11 that the proposed DRL algorithm is the most efficient algorithm based on the throughput performance as it is built on a deep learning architecture and intelligent decision making process at low network levels. With this functionality, it becomes a premier option for use in applications that demand high QoS and network efficiency, like 5G networks and industrial communications.

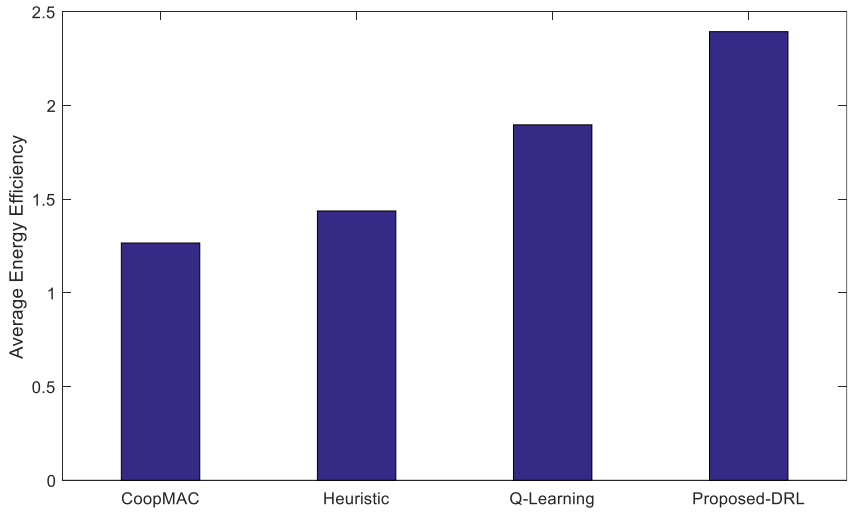

Figure (11) Comparison of average energy efficiency of different algorithms

The change trend of GSM network energy efficiency for 4 algorithms from low to very high user density is illustrated in Figure 12 "Comparison of average energy efficiency at different numbers of users". All the curves are increasing to the right, but the rates and end values are unique depending on the number of users. The proposed DRL algorithm achieves the maximum curve behavior at all times, and remains separated from the other methods as the number of users grows, that is, even for very high density situations, it can learn a good combination of power adjustment, cell on/off, and traffic routing between neighboring cells to use less energy than the energy needed to send the data. The Q-Learning curve is below the DRL curve, which shows classical reinforcement learning has a bit of progress, but it is less energy efficient than the proposed method at higher densities. 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.

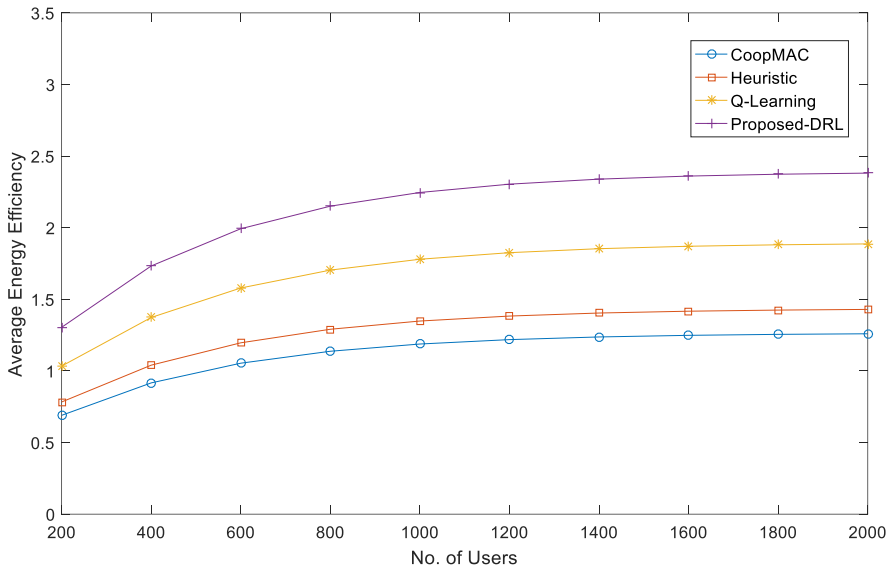

Figure (12) Comparison of the average energy efficiency of different numbers of users

Reward curve analysis

The convergence behavior of the proposed DRL algorithm is given in Figure 13 after 2000 training episodes. Note that the horizontal axis is defined by the number of episodes and the vertical axis is the "normalized average reward", which is the result of the performance gain (throughput and energy efficiency improvements)

minus the penalties that result from QoS violations and control overhead. The agent starts with zero knowledge of its surrounding cells and tries various power levels, cell on/off states and traffic routing patterns between neighbouring cells and slowly learns to avoid actions that would consume high energy or degrade the QoS; therefore the average reward goes up with a fairly large rate during the initial part of training. After some number of episodes, the reward curve enters an area where the reward gain is very small and fluctuates around a nearly steady value. The vertical dotted line in the figure represents the break point, meaning that the improvements in the throughput or energy savings that could be achieved by increasing the energy penalty (increased control overhead or worsened QoS violations) would not be warranted by the total energy benefit/penalty. In other words, the learned policy after this point, creates a stable trade-off between improving efficiency of the GSM network and energy and QoS costs, and future curve changes are predominantly small variations around the optimal reward value.

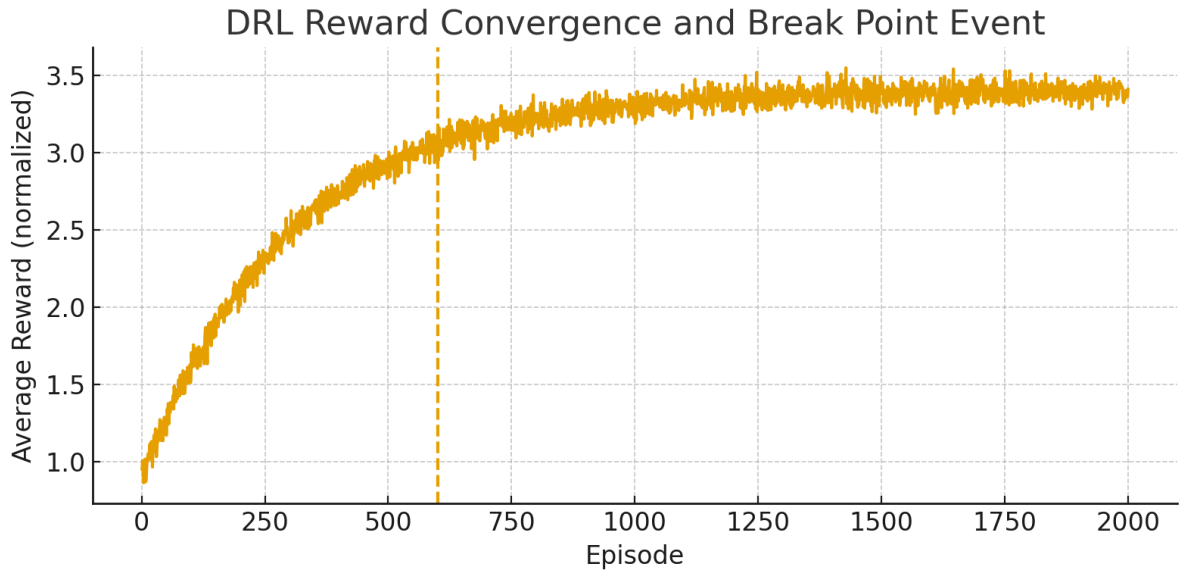

Figure (13) Reward convergence curve of the DRL algorithm and the Break Point event for 2000 episodes

5. Conclusion

This work provides a complete framework to optimize throughput and energy use simultaneously in GSM networks, including a network topological modeling as a small-world graph according to the Watts–Strotz model, a structural and traffic feature extraction through graph exploration, and a design of a deep reinforcement learning agent to perform adaptive control of network parameters. In the modeling part it was demonstrated that the Watts–Strotz model can be tuned with appropriate parameters k and p to generate the graph with a high clustering coefficient and a short average path length, that displays neighborhood and interference relationship between cells as well as handover relationship between them, and that serves to good learning of the structure of the network. Then, based on cell degree and centrality, traffic load, link quality and power status, and action space (power adjustment, cell on/off and traffic allocation policies), the problem was cast as a multi-objective reward function MDP. The results of the simulation demonstrated that the proposed DRL algorithm has significant advantage over the basic, heuristic and classical Q-learning algorithms in terms of the amount of the final reward as well as in the speed of convergence; the average cumulative reward for the proposed algorithm is approximately 3.41 while for the best competing algorithm (Q-learning) is limited to 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. This set of results was analysed for a wide range of user numbers (200 to 2000 users) and found that DRL was able to support traffic growth without significant degradation of QoS even in high-density situations by adapting online to the traffic and channel conditions. Overall, it can be concluded that the structural information from small-world graphs combined with the flexibility of deep reinforcement learning is a promising solution for energy and capacity management in GSM networks and can be applied as a general approach for designing self-organizing policies in the next generation of cellular networks, Internet of Things, and smart industrial systems. The future research opportunities include extension of the present approach by developing multi-agent versions; incorporating real-world operational constraints for radio equipment, and considering user mobility patterns, all of which could make for an exciting extension of the current research.

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. Elkashlan, “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 ,1 .pp. 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.