DRLD2D: Deep Reinforcement Learning-Based Optimized Routing for 5G Device-to-Device Networks

Anu Mangal
Research Scholar, Department of Electrical and Electronics Engineering, National Institute of Technical Teachers Training and Research, Bhopal, India
anu.mangal001@gmail.com

Anjali Potnis
Associate Professor, Department of Electrical and Electronics Engineering, National Institute of Technical Teachers Training and Research, Bhopal, India
apotnis@nitttrbpl.ac.in

Murtaza Abbas Rizvi
Professor, Department of Computer Science and Engineering, National Institute of Technical Teachers Training and Research, Bhopal, India
marizvi@nitttrbpl.ac.in

Abstract: *In context of 5G network, having an efficient communication between the devices in Device to Device (D2D) communication network is a challenging task. The optimised routing in this dynamic and decentralized environment of D2D is crucial for maintaining efficient communication. To overcome these limitations this paper presents Deep Reinforcement Learning (DRL) based routing for D2D 5G network. DRLD2D uses DRL to dynamically optimize the route choice in the network. The problem of routing in DRLD2D is modelled as Markov Decision Process (MDP) and uses the approach of DRL to enable intelligent, adaptive and efficient routing decisions. The optimal paths are learnt by Reinforcement Learning (RL) by analysing network state, node load, Signal-to-Noise Ratio (SNR) and not relying on the static routing tables which is done in the conventional routing protocols. Extensive simulations are conducted using Network Simulator 3 Artificial Intelligence (NS3 AI) model for the proposed DRLD2D protocol which out performs the prevailing routing protocols such as Ad hoc On-Demand Distance Vector (AODV) and Multipath Energy and Quality of Service-Aware Optimized Link State Routing Protocol Version 2 (MEQSA-OLSRv2) when the performance is measured in terms of End-to-End (E2E) delay, Packet Delivery Ratio (PDR), Throughput and Energy Efficiency. The obtained results validate the effectiveness of RL in addressing the routing complexities of D2D communication network which makes it a promising solution for future 5G networks and beyond.*

Keywords: *5G (D2D) networks, (MDP), next hop selection, network performance enhancement, throughput, end to end delay, optimised routing, (DRL), intelligent routing protocol.*

Received September 25, 2025; accepted January 26, 2026
https://doi.org/10.34028/iajit/23/5/15

1. Introduction

The intense rise and elevated demand in mobile data traffic have led to a demand of effective communication paradigms. 5G networks propose Device to Device (D2D) communication, which upgrades overall network performance by downscaling latency, upscaling throughput, and mitigating excess traffic from base stations Malathy *et al*. [21], and Mangal *et al*. [24] Efficient routing in D2D networks is a crucial task with network setup witnessing frequent changes in topology, link failures, and changing mobility of user. Tilwari *et al*. [39], and Pandey *et al*. [27]. Prevailing routing methods fail to accommodate these issues, and henceforth, Reinforcement Learning (RL) is a smart option for effective routing decisions Gottam *et al*. [11].

D2D communication systems are an inherent part of 5G and Beyond (B5G) mobile telecommunication networks Bunu *et al*. [5]. In contrast to conventional and preexisting cellular networks, D2D communication enables and allows nodes to directly interact with one another without or with the intervention of the base station Elleuch *et al*. [9]. The increase in throughput and reduction in latency is observed when the close devices directly communicate without the intervention of eNodeB which would be heavily congested because of traffic loads Yu *et al*. [44] thus, the potential of D2D communication has transformed several fields such as public safety networks, proximity-based networks, vehicular networks, cellular offloading and many more such applications which incorporates it Wang *et al*. [40]. The performance of D2D communication is greatly dependent upon the protocol of routing which it selects. The routing protocol determines the effectiveness and efficiency of the entire network performance. Hence, picking up and the application of the accurate routing protocol for achieving appropriate Quality of Service (QoS) parameters is of utmost importance Zhi *et al*. [46], and Long *et al*. [19]. Figure 1 presents the application of D2D communication in a 5G network.

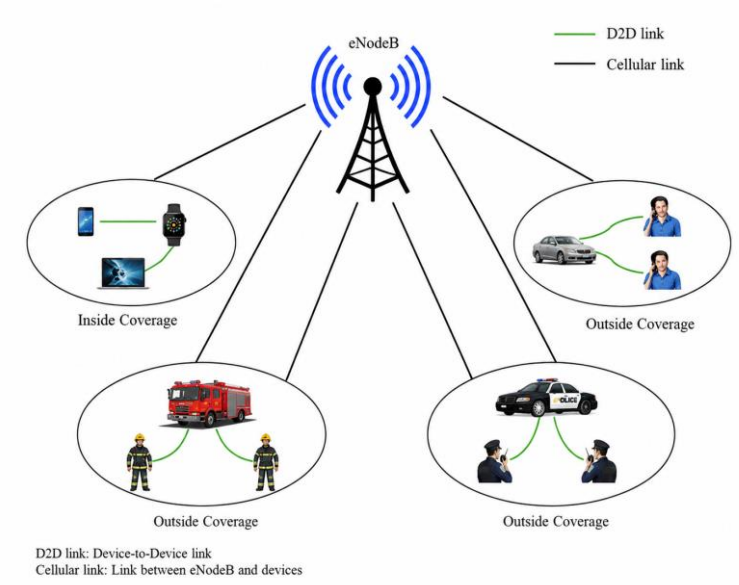

Figure 1. The application of D2D communication scenario in 5G network.

As a consequence of the above problem, various optimization techniques of preexisting and prevailing routing protocols have been investigated by the researchers taking into account several crucial QoS parameters, reliability of the system and the effective utilisation of existing resources Abou *et al*. [1]. One of the analytical challenges with D2D communication is the deployment of wireless device relays with limited energy reserves. So, for the improvement in network performance and maintaining the sustainability in the communication; energy efficient and low overhead protocols for routing is the need of the day Chen *et al*. [6].

For any communication network, Machine Learning (ML) will provide the essentials for optimization of various parameters Huang *et al*. [15]. A subset of ML that is RL provides a promising solution in this context. The complete routing problem is framed as Markov of Decision Process (MDP) and network nodes learn the optimal routing strategies with the continuous interaction by environment Huang *et al*. [13]. Among the various techniques this paper illustrates the Deep Reinforcement Learning (DRL) which is the integration of neural networks with RL and it shows the promising results in determining the efficient QoS in dynamic network conditions Obour *et al*. [3]. DRL makes the decisions which are based on firstly current network condition and secondly the node load, its link quality and traffic patterns without the intervention of the Global network knowledge Moghaddasi *et al*. [25]. The algorithm DRLD2D provides an efficient routing technique that enable our network to constantly learn and adapt itself dynamically to the devices attached to it Huang *et al*. [14]. DRLD2D provides wonderful response for the network topological changes node mobility variation and provides the optimal path for data transmission Yu *et al*. [43]. DRLD2D helps in the enhancement of communication stability and making the system more reliable, providing a smooth and effective data exchange between the devices in a 5G network environment.

1.1. Contributions of the Paper

Initially DRLD2D routing protocol for D2D communication in 5G network has been proposed which is based on DRL. The proposed DRLD2D routing technique redefines hello message structure which is being used in the traditional networks. It optimises it by attaching load and link estate information, transmission error counts along with the conventional message format from source to the destination device.

Secondly the decision of routing is designed as Markov Decision Process (MDP) which specifies agent, state and action spaces; also based on the information received from links and nodes reward function is formulated to meet the requirements of efficient D2D communication. The input of link and node state information is fed in Feed-Forward Neural Network (FNN), the output of which determines which should be the next hop. The entire model is conditioned and trained using DQL algorithm which provides a quick respond ecosystem for the variations in the network condition.

Consequently, NS3 is blended with Network Simulator 3 Artificial Intelligence (NS3 AI) model, the proposed DRLD2D routing strategy is incorporated amongst the devices in D2D which provides insightful routing decisions during the inception of routes. Comparative study of the proposed DRLD2D is provided with conventional protocols which provides significant improvement in PDR, Throughput, End-to-End (E2E) delay and Energy Efficiency showcasing its effectiveness in assurance of liableness and stability of data transmission.

We have presented the remaining of the paper and the paper is organised according to the following: Section 2. illustrates the already existing literature in the related

field and its analysis based on various categories; Section 3. proposes the system model and framework and explanation of system flow of DRLD2D. Illustration of the designing its mathematical Equation framing in addition with the algorithm design is mentioned in Section 4. Section 5 provides simulation environment and the simulation results of DRLD2D protocol and its comparison with already existing protocols for PDR, E2E delay, throughput, and Energy Efficiency. Finally, the summarisation of the entire paper and future work are discussed in section 6.

2. Background and Literature Review

Various researchers have comprehensively studied the routing problems inherent in 5G D2D communication and suggested a wide range of intelligent and adaptive solutions. Based on a detailed study of the related literature, we have observed prevailing thematic areas which broadly classify current trends of research Sahu *et al*. [33]. This survey not only enhanced our understanding of the key concepts but also enabled us to identify research gaps, refine our methodology, and design strategies to improve QoS in 5G D2D communication networks effectively. The literature surveyed belongs to the following categories:

- **Routing Protocols and 5G and D2D Network Challenges**

D2D communication serves as the foundation of B5G networks and aids collaborative communication between close devices without any intervention from the core network. Desauw *et al*. [18]. and Malathy *et al*. [21] provides a high-level and enlarged overview of D2D routing ideas with emphasis on barriers and advancements for IoT and 5G use cases. Duong *et al*. [8] extends Ad hoc On-Demand Distance Vector (AODV) routing by the employment RL for strict QoS support in Mobile Ad hoc Network (MANET) which utilizes 5G. Bunu *et al*. [5] proposes a node status metric-driven method by employing machine learning to optimize relay node selection for B5G networks. In delay-constrained NB-IoT service. Naumen *et al*. [26] employs Q-learning to maximise D2D communication, lowering the energy consumed and reducing the transmission delay. Sahu *et al*. [31]. and Jayakumar *et al*. [16] proposes a distributed Q-learning solution for scalable and adaptive resource allocation in dynamic D2D networks Sahu *et al*. [30]. These works reinforce the necessity of smart, self-calibrating routing protocols to handle changing network needs.

- **Energy-Efficient Routing and Base Station Optimization**

Energy efficiency is a critical issue in high-density 5G and sensor network scenarios where constant communication is prone to consume power resources quickly Adu-Manu *et al*. [2] and Malta *et al*. [23] an SARSA-based framework is introduced to dynamically put base stations into sleep mode, gaining significant energy-saving in Ultra-Dense Networks (UDNs). Yang *et al*. [42] an amalgamated approach of swarm intelligence and DRL, WOAD3QN-RP, is adopted to promote cluster head selection and route optimization in WSNs. Arya *et al*. [4] combines RL, Deep Belief Networks, and nature-inspired optimization algorithms for energy-efficient routing in IoT settings. In addition, Kiran *et al*. [37] introduces a secure routing protocol for smart homes in 5G networks based on Sailfish optimization for efficient IP mobility and data security. Collectively, these papers show the potential of learning-based methods to minimize power consumption without network reliability and security degradation.

- **Reinforcement Learning and Deep Learning in Routing**

RL and deep learning are most likely employed for tackling the routing and resource provisioning complexity in 5G and Internet of Things (IoT) systems. Zhang *et al*. [45] unveils DRL based approach for a Proximal Policy Optimization (PPO) for effective task offloading in Mobile Edge Computing (MEC), minimizing latency and energy expenditure. Ron *et al*. [29] applies a deep neural network (DNN) to simultaneously optimize power control and mode selection in D2D networks, solving a traditionally NP hard issue. Malik *et al*. [22], an RL model is introduced for routing in Cognitive Radio-Enabled Internet of Things (CR-IoT) networks, which incorporates channel selection within the learning process, thus enhancing throughput and delay performance. In the same vein, Lu *et al*. [20] presents an energy, cache status, and node location-aware Q-learning-based routing mechanism for Vehicle-to-Vehicle (V2V) communication to enhance overall routing decisions in vehicular networks. These articles demonstrate the increasing presence of Artificial Intelligence (AI) in optimizing and automating the process of routing.

- **QoS and Privacy-Aware Routing Mechanisms**

QoS with data confidentiality is a requirement in heterogeneous and dynamic networks. Kumari *et al*. [17] presents a Hybrid Metaheuristic Algorithm (HAPM) based on ACO and PSO to enhance multicast routing efficiency in MANETs with a considerable improvement in some QoS parameters. Wang *et al*. [41], an instance of a federated learning-based protocol QoS and Data Privacy-Aware Routing Protocol (QoSPR) for the 5G-connected Industrial IoT (IIoT) is proposed that facilitates privacy and QoS simultaneously through distributed gateway placement and deep RL. Quy *et al*. [28] proposes routing protocol which is efficient in energy and awareness in delay carefully tailored for self-organizing networks in Intelligent Transportation Systems (ITS), with improved throughput and reduced latency. These papers reveal encouraging directions for balancing QoS optimization with growing demand for privacy-aware communication protocols.

- **Vehicular and Aerial Network Routing**

Vehicular Ad Hoc Networks (VANETs) and Flying Ad Hoc Networks (FANETs) have distinctive routing requirements with high mobility, these high rates networks have unique routing demands of dynamic movement, topology volatility, and hard real-time performance. Cui *et al*. [7], presents TARRAQ, a Q-learning-based routing protocol adapted for communication in FANETs using queuing theory for adaptation to network changes. Sahu *et al*. [12], presents the 6G roadmap for vehicular intelligence with emphasis on synergies between edge computing, AI, and green communication in future transportation networks. of topology change, and hard real-time performance. Cui *et al*. [7], proposes TARRAQ, a Q-learning-based routing algorithm specific to UAV communication in FANETs with queuing theory-adaptive network dynamism. Sahu *et al*. [12], addresses the 6G roadmap of vehicular intelligence with emphasis on synergistic effects of AI, edge computing, and green communications for future transportation networks. Lu *et al*. [20], also examines RL-based routing in V2V scenarios, showing enhanced decision making with the application of dynamic environmental factors Sahu *et al*. [34], These papers are important in highlighting the importance of AI and machine learning towards enabling strong and adaptive routing for highly mobile networks Sahu *et al*. [36].

- **Network Slicing and Architecture for 5G and B5G**

To cope with the growing heterogeneity of services and devices in 5G/B5G networks, network slicing has come forth as a critical architectural solution. Ssengonzi *et al*. [38] discusses the use of DRL for controlling the complete lifecycle of network slices, providing smart automation for service orchestration, allocation, and optimization. In complement, Goswami *et al*. [10] presents a technical and historical survey of the growth and evolution of mobile networks from 1G to 6G with a special emphasis on developments in routing and machine learning technologies. These seminal papers offer critical insights into the architectural and strategic frameworks underpinning scalable, flexible, and intelligent wireless next-generation systems Sahu *et al*. [32], A critical analysis of existing routing protocols, D2D communication networks and RL strategies is presented in Table 1, which highlights the key limitations of current approaches and identifies major research gaps that the proposed model aims to address.

Table 1. Literature review analysis.

S. No	Reference	Study	Protocol	Objective	Methodology	Advantages	Limitations/Remarks
1	Duong *et al.* [8]		Enhanced AODV with reinforcement learning	To ensure QoS in 5G-based MANETs with high traffic loads	Reinforcement learning is used for maintaining a state information database at each node, which helps in selecting QoS-guaranteed routes	Improved throughput, reduced end-to-end delay, and enhanced SNR	Limitations of existing protocols in handling stringent QoS in 5G-MANETs are addressed, but scalability in ultra-dense scenarios may still be a concern
2	Malta *et al.* [23]	Energy-efficient base station management using reinforcement learning in 5G ultra-dense networks	SARSA-based sleep mode management	To optimize energy consumption of base stations while maintaining QoS in 5G UDNs	Reinforcement learning (SARSA) is used to dynamically manage sleep, inactive, and active states of base station components, using QoS-aware metrics	Achieves up to 80% energy savings under low traffic; adaptable trade-off between energy and QoS	Limited energy savings when ultra-low latency (e.g., 1 ms) is prioritized; performance depends on traffic scenarios
3	Bunu *et al.* [5]	Machine Learning-based Optimization of OLSRv2 for Efficient Multi-hop D2D Communication in 5G and B5G Networks	Modified OLSRv2 with Supervised ML	To improve routing efficiency and energy usage in out-of-coverage multi-hop D2D communication	Introduces Node's Status (NS) metric based on battery level, mobility, node degree, and BS connection; uses supervised ML (K-NN, RF, MLP, GBC) to combine metrics into one routing decision	Reduces routing overhead and energy consumption; Random Forest achieves 100% accuracy in simulations	Dependent on training data quality and simulation setup; real-world adaptability and generalization may vary
4	Zhang *et al.* [45]	PPO-Based D2D-Assisted Computation Offloading and Resource Allocation in MEC-Enabled 5G and Beyond Networks	PPO-based D2D Offloading Framework	To optimize computational offloading and resource allocation in MEC networks for reduced energy and latency	Markov Decision Process (MDP) model solved using Proximal Policy Optimization (PPO); D2D communication enables collaborative task offloading	Enhanced latency reduction, energy efficiency, and convergence compared to benchmarks	Complexity in real-world deployment; performance highly dependent on accurate MDP modelling and dynamic network conditions

5	Ron *et al.* [29]	Deep Neural Network-Based Joint Mode Selection and Power Allocation in Underlaid D2D Communication	DNN-Based Mode Selection and Power Control	To jointly optimize mode selection and transmit power allocation in D2D pairs for interference mitigation	Formulates NP-hard joint optimization problem with linear and nonlinear constraints; solved using a custom-trained Deep Neural Network based on Lagrange duality function	Achieves near-optimal performance with lower complexity than exhaustive search; improves interference management	May require large training datasets; performance depends on accuracy of channel gain estimation and dynamic interference conditions
6	Cui *et al.* [7]	Topology-Aware Resilient Routing Strategy Using Adaptive Q-Learning for FANETs	TARRAQ (Topology-Aware Resilient Routing using Adaptive Q-learning)	To design an energy-efficient, low-delay, and high-PDR routing protocol for highly dynamic Flying Ad-Hoc Networks (FANETs)	Uses queuing theory to model UAV dynamics; derives NCR and NCIT for sensing interval calculation; applies adaptive Q-learning for distributed and autonomous routing	Achieves significantly lower overhead and energy consumption; improves PDR over QTAR, MPVR, and EE-Hello protocols	Complexity in modelling dynamic behaviour and tuning learning parameters; dependent on accurate estimation of NCR/NCIT
7	Kumari *et al.* [17]	Hybrid ACO-PSO Meta-Heuristic (HAPM) for QoS-Aware Multicast Routing in MANETs	HAPM (Hybrid ACO-PSO Meta-Heuristic)	To improve multicast routing in MANETs with QoS constraints using a biologically inspired hybrid approach	Combines Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and dynamic queuing; simulation in NS2; evaluates metrics like PDR, delay, hop count, overhead, throughput	Significant improvements: PDR ↑increase by up to 20%, delay decreases by up to 54%, overhead reduced by up to 49%, throughput increased by up to 40%, hop count reduction by up to 90% compared to baseline protocols	
8	Nauman *et al.* [26]	RL-Intelligent-D2D (RL-ID2D) Communication for NB-IoT	RL-ID2D (Q-Learning-based D2D Relay Selection)	To improve energy efficiency and delivery ratio in NB-IoT for delay-sensitive applications (e.g., healthcare) using D2D communication	Formulates an optimization problem to maximize end-to-end delivery ratio (EDR); uses Q-learning to select the optimal cellular relay for two-hop D2D uplink instead of direct base station communication	Reduces repetition overhead and energy consumption, especially for edge NB-IoT devices; improves EDR using intelligent relay selection	Focused on two-hop relay; scalability and adaptability to high-mobility or multi-hop scenarios may need further study
9	Arya *et al.* [4]	Energy-Efficient DBN-Based Routing for IoT-WSNs	RL-MRFO-DBN Routing Protocol	To enhance network lifetime, energy efficiency, and PDR in IoT-based constrained WSNs using intelligent clustering and CH selection	Clustering via Reinforcement Learning (RL); Cluster Head (CH) selection using Mantaray Foraging Optimization (MRFO); Routing through Deep Belief Network (DBN)	Achieves higher packet delivery ratio, improved network lifetime, better node reachability, and energy-efficient routing	Depends on accurate model training; computational complexity of DBN and MRFO may be high for resource-constrained WSN nodes
10	Jayakumar *et al.* [16]	Distributed Resource Allocation for D2D in B5G using Q-Learning	Q-learning-based Distributed Resource Allocation	To achieve high throughput, improved spectrum and energy efficiency, and fairness in dynamic D2D environments	Distributed reinforcement learning with Q-learning; agents (devices) learn optimal resource allocation through real-time interaction and reward feedback	Improves throughput (6-8%), fairness, latency, and efficiency; scalable with low overhead	Performance gain depends on training time; may need careful tuning for fast convergence in highly mobile or dense networks
11	Wang *et al.* [41]	QoS and Privacy-Aware Routing in 5G-IIoT	QoSPR	To ensure QoS and data privacy in 5G-IIoT routing with reduced latency and balanced load	Info-map-based community detection; Deep RL for gateway deployment; Federated RL for privacy-preserving universal model; QoS-aware routing	Reduces average and maximum latency; ensures data privacy; improves load balancing	Computational overhead due to federated learning; dependent on community detection accuracy and real-time data availability

12	Malik *et al.* [22]	Reinforcement Learning-Based Routing in CR-IoT	RL-IoT	To achieve higher data rate and reduced end-to-end delay in CR-enabled IoT networks	RL-based channel-aware routing; Integrated channel selection at network layer; Simulated using CRCN	Improved throughput, reduced packet collisions and EED; outperforms AODV-IoT, ELD-CRN, and SpEED-IoT	Dependent on simulation environment (CRCN); may require real-time channel state info for deployment
13	Lu *et al.* [20]	Reinforcement Learning-Based V2V Routing in 5G IoT	RLbR	To reduce load on 5G cellular networks and improve V2V routing efficiency	RL-based routing; Q-values computed using Cache Factor (CF), Energy Factor (EF), and Position Factor (PF); Convergence accelerated by environmental model	Highest traffic offload rate; improved V2V network lifetime; better delivery ratio and average delay	Primarily benefits non-real-time traffic; effectiveness depends on accurate neighbour vehicle evaluation and environment modelling
14	Sharma *et al.* [37]	Secure Routing in 5G Smart Home using SbDMM	Sailfish-based Distributed IP Mobility Management (SbDMM)	To ensure secure and efficient data routing in 5G-enabled smart homes	Sailfish Optimization for path fitness; uses session, cipher, and authenticated keys for security; implemented in Python	Improved execution time, confidentiality, efficiency, and task completion; optimized secure routing	Focused on smart home IoT; performance outside such environments (e.g., industrial or vehicular networks) not discussed
15	Quy *et al.* [28]	Efficient Routing in Self-Organizing Networks (SONs) for Smart Cities	Not explicitly named; SON-based routing scheme	To enhance routing performance and reduce energy consumption in SONs for Intelligent Transportation Systems (ITS)	Develops a routing scheme for SONs with self-configuration and autonomous communication tailored to low-mobility ITS scenarios	Improved network lifetime, packet delivery ratio, and latency in low-speed conditions	Limited performance evaluation under low mobility; effectiveness in high-speed scenarios or heterogeneous environments not discussed

Table 1 provides the literature review analysis and finding the notable contributions and also the research gaps in the concerned field. This comparative review of the previous works presents several critical research gaps: there is a notable inadequacy of routing protocols that efficiently integrates real-time adaptability and improves QoS by effectively choosing the route based on parameters and learning the path so as to improve overall efficiency Sahu *et al*. [32]. Moreover, there is a lack of unified framework that combines D2D communication, RL in 5G network for proper optimization of QoS parameters.

3. System Model

For addressing the challenges in the routing of 5G D2D communication DRLD2D is proposed which is based on RL and is designed to enhance the performance of the network and improves its QoS parameters. The method is based on adaptive and self-learning capabilities of RL to make the routing decisions real time under various network conditions without relying on predefined routes. DRLD2D is suitable for heterogeneous and dynamic nature of 5G D2D networks where the traditional routing protocols are inefficient.

This section presents the system framework, design flow and complete description of the proposed methodology DRLD2D. The objective of the section is to provide a crisp and clear understanding of the proposed approach.

3.1. DRL-D2D Framework

A potent solution is outlined by DRL for routing in D2D networks, where fluid network structure, limited resources, and decentralized control present significant challenges to traditional routing protocols. By combining the decision-making capabilities of RL with the feature extraction power of deep neural networks, DRL enables D2D nodes to learn optimal routing strategies through continuous interaction with their environment. n this approach, each node operates as an intelligent agent that observes its local state-such as neighbour availability, link quality, and buffer occupancy-and selects the best next-hop node to maximize long-term communication performance. Figure 2 provides the System Framework of DRLD2D which comprises several key components that enhance routing efficiency and adaptability in a dynamic 5G network environment.

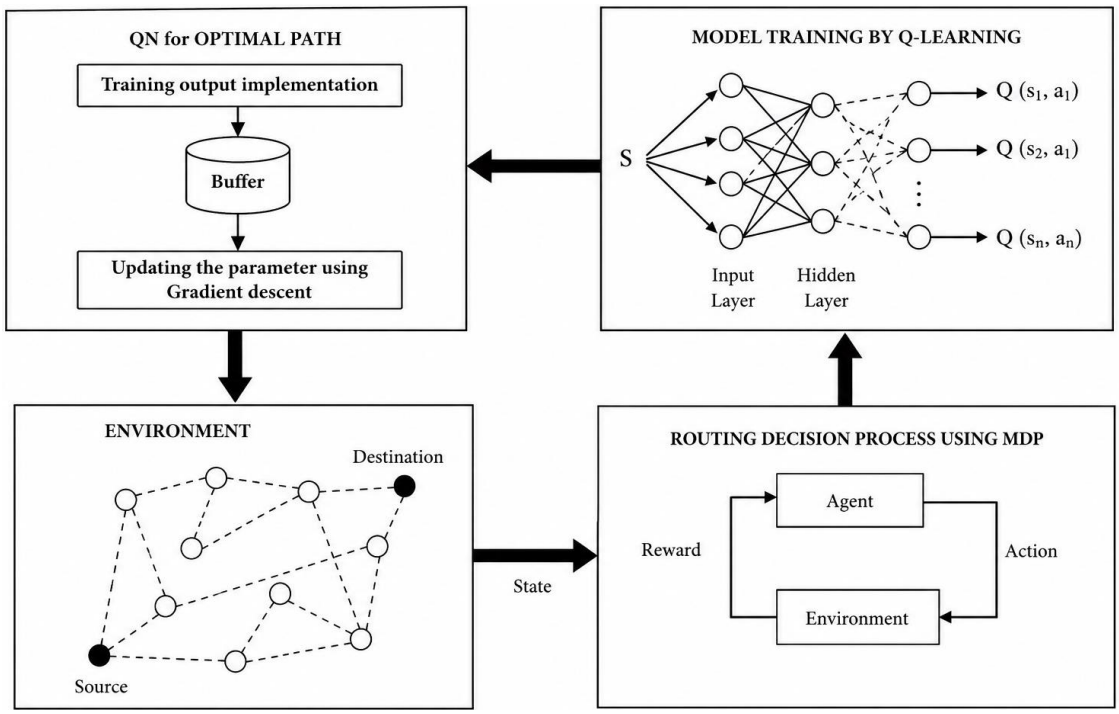

Figure 2. Diagram of system framework.

The system framework of DRLD2D algorithm can be broadly classified in four sections as follows:

- **System Environment:**

It starts with the environment in which various devices are there which are interconnected or they can communicate with each other without the need of central architecture that is D2D communication and one device wants to transfer the data to another device. The device which wants to send the data is considered as the source and the device to whom the data is being transmitted this considered as destination. The environment provides state information to the routing decision module. The first component is the enhanced hello message exchange, where the conventional hello message format is extended to include additional node and link state information. This allows for more informed routing decisions. Each node periodically shares critical metrics with its neighbours, such as residual energy levels for energy-aware routing, link quality and stability to reduce failures, traffic load to prevent bottlenecks, and relative distance and mobility patterns to enhance route reliability. This real-time exchange of network state information improves route selection efficiency.

- **Markov Decision Process (MDP):**

The routing decision is formulated as a MDP to enable intelligent routing. The State Space (S) represents the dynamic network environment, incorporating factors like node mobility, link quality, energy levels of relay nodes, and traffic load on potential paths. The Action Space (A) defines possible routing actions, where each action corresponds to selecting an optimal next-hop relay node. This selection is driven by RL optimization, ensuring energy efficiency, load balancing, and stable connectivity. It involves an agent that interacts with the environment, selecting actions based on received states. The agent takes routing actions, and in return, receives rewards or penalties depending on the quality of the selected path. This process continuously updates routing decisions to improve network performance. The Reward Function Design further refines the routing process by reinforcing optimal paths and penalizing inefficient ones. Positive rewards are assigned to routes with stable links, high residual energy, low congestion, and minimal hop count, ensuring network longevity and efficiency. Conversely, negative rewards are given to paths with frequent link failures, congestion-prone nodes, and excessive energy consumption, discouraging suboptimal routing choices.

- **Model Training by DRL:**

This block handles the training of the routing policy using Deep Q-Learning (DQL). A Deep Neural Network is used to approximate the action-value function Q (s, a; θ) which predicts the expected cumulative reward for taking action an in states. The training involves input layer which receives the state vector, hidden layers which process the input using non-linear activations and output layer which provides Q-values for all possible actions (i.e., next-hop candidates).
The model is trained using the Bellman Equation to update the Q-values as Equation (1):

$$Q(s_t a_t) \leftarrow Q(s_t, a_t) + \alpha[rt_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)] \quad (1)$$

Where α is Learning Rate and γ is the Discount Factor.

During training, ε-greedy exploration is used to balance exploration of new paths with exploitation of known good routes. Training is performed offline, using

large datasets generated through simulation, allowing resource-constrained nodes to use only lightweight inference during actual routing.

- **QN for Optimal Path:**

The QN for Optimal Path module stores training outputs in a buffer and updates parameters using gradient descent. This step ensures that learned routing decisions are refined over time, leading to progressively better routing performance. The trained model then feeds back into the environment, applying its optimized routing policies to improve packet forwarding decisions.

Together, these four blocks constitute a robust system framework for enabling distributed, intelligent routing in 5G D2D networks. The network environment models real-time D2D conditions, the MDP formalizes decision-making, the DQL model learns optimal routing behaviour offline, and the replay buffer ensures efficient and stable training. Once deployed, devices use their pre-trained models to make real-time routing decisions based only on local observations, enabling scalable, energy-aware, and adaptive communication.

3.2. Neural Network Model

Due to high memory necessities and poor scalability, the conventional Q-learning techniques based on tabular representations become unfeasible when provided with high-dimensional and continuous nature of the state space. To get around this restriction, a FNN is used to approximate the Q-function, turning the Q-table update process into a function approximation problem. The proposed Deep Q-Network (DQN) architecture consists of an input layer, two hidden layers, and an output layer. The job of the input layer is to receive the state data, the hidden layers does the job of extracting the feature from the provided data and updating the weight. Each neuron provides a weighted summation succeeded by a nonlinear activation function. In correspondence with all the possible actions, the output layer generates the Q-values. The network uses state information which is derived from the current node and its neighbour nodes in the D2D network environment.

- **Input Layer:**

The dimension of the input layer is fixed at 4+6N (by design) where N represents the maximum number of neighbouring nodes in the present network topology. If number of neighbours is less than N, the unused slots of input are padded with a default value of-1 for the maintenance of constant input size for the given neural network. This design enables the model to accommodate dynamic network topologies with changing counts of neighbour.

The first four input features represent the destination node's position (x_d', y_d) and the current node's position (x_C', y_C). The remaining $6N$ features correspond to the information of the neighboring nodes. For each neighbor, six parameters are included: its position (x', y), node load n_l, transmission error count c, link duration l_d, and the distance d to the destination node.

- **Hidden Layers:**

The network comprises of two hidden layers. Each layer consists of 64 neurons. Both layers employ the Rectified Linear Unit (ReLU) activation function for introducing non-linearity and enhancement of the network's capability to learn complex relationships within the state space. For the updating of network weights efficiently during training, the Adam optimization algorithm is utilized, providing faster convergence and improved learning stability.

- **Output Layer:**

The output layer has a dimensionality of N which corresponds to the maximum number of neighbouring nodes. Each output neuron represents the Q-value associated with selecting a particular neighbour as the next hop. For nodes with fewer than Nneighbors, the additional output values are assigned a predefined exceptional value to approximately handle the given cases and prevent invalid action selection.

3.3. Design Flow of DRLD2D

In the highly dynamic and decentralized environment of 5G D2D communication, traditional routing protocols struggle with scalability, mobility, and real-time adaptability. DQN addresses these limitations by enabling each device to learn optimal routing decisions based on experience, without requiring a central controller or static routing tables. Flowchart of the proposed methodology i.e. DRLD2D is described in two flowcharts, first flowchart figure 3 presents how the routing of packet is done in 5G D2D Communication Network using the RL and second flowchart Figure 4 provides the flow of working of RL block i.e. DDQN for routing. Figure 3 provides the flowchart of DRLD2D Routing in 5G Network. The flowchart illustrates a process for determining an efficient path using RL in a D2D 5G network. It begins at the "Source," which is the device which wants to send the data where the system initializes path efficiency and gathers information about the topology of the available choices. Next, network devices share topological information to facilitate decision-making as provided in Figure 3. The RL algorithm is then applied to find the optimal route based on value of the reward function. It is an iterative process involving observation, learning, and taking action. The selected path is then projected, and its performance is assessed. If the required performance is not achieved (based on the reward obtained), the process loops back for further refinement. If the performance criteria are met, the optimal path with high rewards is chosen. Finally, the data reaches the "Destination" completing the process through the path chosen by RL.

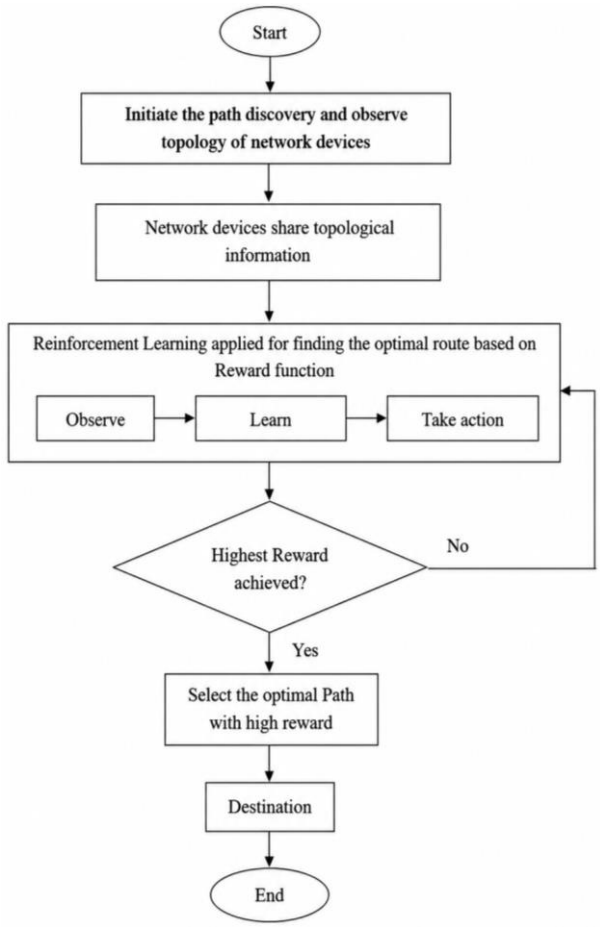

Figure 3. Flowchart demonstrating DRLD2D routing in 5G network.

The flowchart in figure 3 provides a structured approach for dynamic path optimization in 5G D2D network using RL.

In the DQN-based routing framework as mentioned in Figure 4, firstly the DQN Agent is initialized and the parameters of neural networks are set. At each stage of decision making, the agent observes the current state of D2D environment comprising features like neighbour availability, link quality, Signal-To-Noise Ratio (SNR), buffer occupancy, and residual energy. Based on the observed state the agent selects the action for the next neighbouring hop for the forwarding of packet based on the reward obtained containing the required parameters. The routing objective is modelled as a MDP, where the agent learns a policy to maximize a cumulative reward-which can be defined to promote high throughput, low latency, reduced energy consumption, and reliable delivery. The agent checks whether the selected action obtains the maximal reward. If not, the agent is looped back in the environment for continuous learning process. If the maximum reward is achieved, then the action is beneficial and moved to the next stage. The experience is stored in the replay buffer. The DQN model is updated periodically after every N steps. The process then checks whether all packets have reached their destination or not. If no, then the process is looped back into training and learning. If yes, the episode concludes and routing process comes to end.

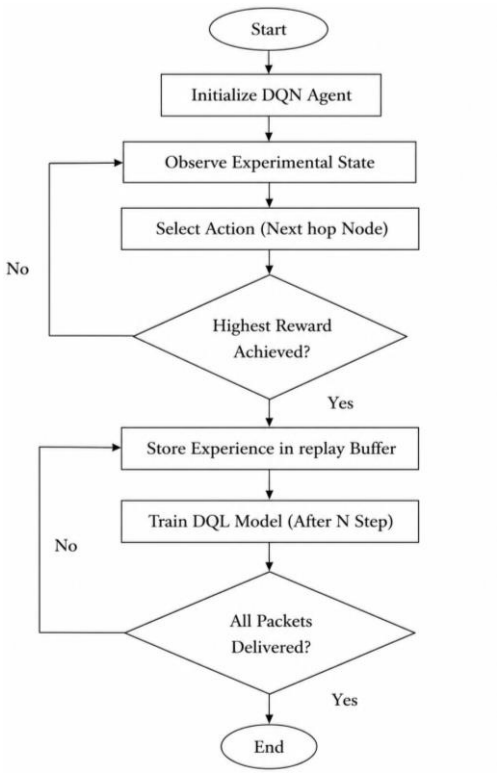

Figure 4. Flowchart showing DQN operation flow for routing in D2D communication in 5G network.

Figure 4 demonstrates DQL Operation flow for Routing in D2D Communication in 5G Network, the details of which are already mentioned in the previous paragraphs.

4. Design and Implementation of DRL-D2D

This section presents the design and implementation of the proposed DRL-D2D communication framework tailored for 5G D2D networks. The framework is structured around three core design components to address the limitations of conventional routing approaches in dynamic and resource-constrained environments. First, the protocol content design involves refining the structure of routing messages and routing tables to enable more efficient and accurate data dissemination. This includes integrating additional state information, such as node energy levels, link quality, and mobility patterns, into control messages to facilitate informed routing decisions.

Second, the model design represents the routing problem as a MDP, capturing the sequential nature of decision-making in dynamic network scenarios. This abstraction allows the agent to learn optimal routing policies based on the state of the network and its historical interactions.

Third, the algorithm design employs a DQN to enable the agent to select optimal forwarding paths while avoiding the overestimation bias typically encountered in standard DQN approaches. DQN enhances the stability and convergence of the learning process, particularly in environments with large and continuous state spaces.

4.1. Protocol Content Design

4.1.1. Enhanced HELLO Message Format:

Traditional routing protocols utilize HELLO messages for neighbour discovery and link status updates. However, in highly dynamic D2D environments, conventional HELLO messages lack critical information about node performance, energy levels, and mobility.

To address this, the proposed DRL-D2D protocol introduces an enhanced HELLO message format, integrating additional parameters that enable efficient and stable routing decisions. Table 2 presents the structure of the enhanced HELLO message, incorporating new fields to facilitate energy-aware and mobility-aware routing.

Table 2. Enhanced HELLO message format.

Message type	Flag	Re-served	Prefix size	Hop count	Destina-tion IP Address	Destination sequence number	Originator IP address	Life-time	Node position X	Node position Y	Node velocity X	Node velocity Y	Remaining energy level	Link quality	Node load	Transmis-sion error count

Table 3. Modified routing table entry fields.

Destination IP address	Sequence number	Verify sequence number flag	Interface	Hop count	Next hop	Lifetime	Node load	Remaining energy	Transmission error count	Link duration

Each HELLO message contains essential information that aids in efficient network communication and routing decisions. It includes the node's position and velocity, which help predict mobility patterns and anticipate potential link disruptions. Additionally, the remaining energy level is shared to prevent routing through nodes with low energy, thereby extending the overall network lifetime. The message also conveys link quality by measuring signal strength and stability, ensuring the selection of reliable communication paths. Furthermore, node load information is included to distribute traffic efficiently and avoid congested nodes. Lastly, the transmission error count is reported to assess communication reliability and steer data away from high-error links, enhancing the overall robustness of the network. These enhancements improve the protocol's adaptability and robustness, enabling efficient routing in dynamic D2D networks.

4.1.2. Modified Routing Table Design:

To complement the enhanced HELLO messages, the routing table structure is modified to integrate energy-aware parameters for better decision-making as mentioned in Table 3. Each routing entry contains critical parameters that enhance network efficiency and reliability. It includes the remaining energy of nodes, ensuring that paths prioritize those with sufficient energy reserves to prolong network operation. Additionally, the transmission error count is recorded to help avoid low-quality links, maintaining robust and stable communication. Link duration is also considered, estimating the stability of connections to prevent frequent route failures. By integrating these parameters, the routing mechanism optimizes network longevity and reliability, reducing disruptions and enhancing overall performance.

4.2. Design Matrices

A scenario involving a D2D communication comprised of mobile communication devices within a 5G communication network environment is considered in this paper. Each node in this network can serve as either a source or destination. Data packets are transmitted over wireless links between nodes/devices, but node mobility or damage can lead to link disconnections and subsequent communication disruptions. The object is to identify an optimal path in this dynamic setting to ensure reliable and stable network communications. To achieve this, a DQL based routing strategy is developed that dynamically adapts to changes in network topology and effectively addresses the challenges associated with node movements or damages. This study aims to provide an efficient routing solution for D2D communication in 5G network.

This study considers a 5G D2D communication network, where mobile devices communicate directly without relying on centralized infrastructure. Each device in the network functions as a node, serving as either a source or a destination for data transmission. All nodes are assumed to be randomly deployed in a 2D plane, where they automatically establish connectivity. The D2D network can be modelled as a connected Graph (G) The model is trained using the Bellman Equation to update the Q-values as Equation (2):

$$G = (Z, E) \tag{2}$$

Where:

- $Z = \{z_1, z_2, z_3, \ldots, z_i, \ldots, z_N\}$ represents all nodes/devices
- E represents available communication links, where each $e_{ij} \in E$ denotes the communication link between node i and node j.

Additionally, each node maintains the following information:

- Positional coordinates (x_i, y_i)
- Velocity vector (v_{ix}, v_{iy})
- Remaining energy level E_i

4.3. Mathematical Model of Protocol Design DRLD2D

4.3.1. Markov decision process:

In this paper, a MDP is employed to model the routing decision process for devices. A MDP comprises a set of interacting components, namely, agents and the environment to interact, which include states, actions, and rewards. Based on the MDP, the agent perceives the current state according to the strategy to implement the action on the environment, thereby altering the state and receiving the reward. Subsequently, detailed definitions of the agent, state space, action space, and reward function are described.

- **Agent:**

Given that route establishment fundamentally relies on per hop node/device selection, and each node independently selects its next-hop neighbour, each device within the network is treated as an agent.

- **State Space (s_i^t):**

At time step t, the state observed by a node *i* is denoted by the model is trained using the bellman Equation to update the Q-values as Equation (3):

$$s_i^t = \{D_i^t, N_i^t, NL_i^t, C_i^t, LD_i^t\}) \tag{3}$$

where:

- D_i^t: Destination node ID.
- N_i^t: Set of current one-hop neighbors of node *i*.
- N_i^t: Queue load at node i's neighbours.
- C_i^t: Transmission error count of neighbours.
- LD_i^t: Link duration between node *i* and each neighbour.
- Action Space *A*: The action $A_{i,j}{}^t$ corresponds to selecting a neighbouring node $j \in N_i^t$ for forwarding the data packet at time *t*.

4.3.2. Reward Function R:

The reward function evaluates the suitability of selecting neighbour *j* from node *i*, considering load balancing, link reliability, stability, and path optimality The model is trained using the Bellman Equation to update the Q-values as Equation (4):

$$r_i^t(s_i^t, a_{i,j}^t) = \begin{cases} \max, j \in D_i^t \\ \alpha 1 \cdot (1 - \frac{nl_j}{P_{max}}) + \alpha 2 \cdot \left(1 - \frac{c_j}{E_{max}}\right) + \alpha 3 \cdot \frac{ld_{i,j}}{L_{max}} + \alpha 4 \cdot (1 - d_{rel}), j \in D_i^t \end{cases} \tag{4}$$

where:

- nl_j: Queue length at node *j*.
- P_{max}: Max queue capacity.
- c_j: Transmission errors at node *j*.
- E_{max}: Max transmission errors allowed.
- $ld_{i,j}$: Link duration between *i* and *j*.
- L_{max}: Maximum link duration threshold.
- d_{rel} Relative distance metric to the destination (defined below).
- α1, α2, α3, α4: Weighting factors.

The values in the Equation (2) of reward function can be plugged in from Equations (3, 4, 5, 6, 7 and 8) respectively. On the basis of the reward achieved the next hop routing decision will be done.

4.3.3. Link Duration Estimation

Let

- (x_i, y_i) and (x_j, y_i): positions of node *i* and node *j*
- (v_{ix}, v_{iy}) and (v_{jx}, v_{jy}) be the velocities

The Euclidean distance $(d_{i,j}{}^t)$as Equation (5):

$$d_{i,j}^t = (x_i \ , x_j \ )^2 + \left(y_i - y_j\right)^2 \tag{5}$$

Relative velocity (v_{rel}) as Equation (6):

$$v_{rel} = \sqrt{\left(v_{ix} - v_{jx}\right)^2 + (v_{iy} - v_{jy})^2} \tag{6}$$

Estimated link duration ($ld_{i,j}$) as Equation (7):

$$ld_{i,j} = \frac{R - d_{i,j}}{v_{rel}} (capped \ \ at \ \ L_{max} \tag{7}$$

Where *R* is the communication range.

4.3.4. Relative Distance to Destination (d_{rel})

Given:

- Destination coordinates: (x_d, y_d)
- Next-hop coordinates: (x_j, y_j)

Then the distances are defined as follows, as Equation (8) and (9):

$$d_{i,d} = \sqrt{(x_i - x_d)^2 + (y_i - y_d)^2} \tag{8}$$

$$d_{j,d} = \sqrt{\left(x_j - x_d\right)^2 + \left(y_j - y_d\right)^2} \tag{9}$$

The relative progress metric is as Equation (10):

$$d_{rel} = \begin{cases} \frac{d_{j,d}}{d_{i,d}} & , if \ d_{j,d} \le d_{i,d} \\ 0, & otherwise \end{cases} \tag{10}$$

This encourages selection of nodes moving closer to the destination.

- **Q-Learning Update Rule**

Using the DQL approach, the Q-value update is as Equation (11):

$$Q(s_i^t, a_{i,j}^t) \leftarrow Q(s_i^t, a_{i,j}^t) + \alpha[r_i^t + \gamma \max_{a'} Q(s_j^{t+1}, a') - Q(s_i^t, a_{i,j}^t)] \tag{11}$$

Where:

- α: Learning rate
- γ: Discount factor
- s_j^{t+1}: New state after taking action $a_{i,j}^t$

D2D networks.

4.4. Algorithm Design

The DQN-based routing decision framework is designed to enhance network efficiency and stability through intelligent decision-making. It begins with network initialization, where nodes exchange HELLO messages to collect vital link information. Following this, each node updates its routing table with key metrics such as link quality, energy levels, congestion, and transmission errors. The routing decision is then made using the DQN model, where the agent selects the optimal next-hop node based on computed Q-values. In cases where multiple nodes have similar rewards, the selection prioritizes nodes with better link stability and higher energy levels. To improve learning and adaptability, experience replay is used to train the DQN model, allowing for dynamic decision-making. Finally, the performance of DRLD2D strategy is evaluated against AODV and OLSRv2 based on metrics such as throughput, latency, Packet Delivery Ratio (PDR), and energy consumption. By integrating these intelligent routing mechanisms, the Energy-Aware Routing Algorithm Using DQN ensures optimal route selection while maintaining energy efficiency and network stability, making it highly suitable for dynamic 5G D2D networks. The training of the system for the achievement of reward and generation of Q value for next hop is done using Algorithm 1 and complete sending and receiving of data packets from source to destination is provided as a workflow in Algorithm 2

Algorithm 1. DQN-based routing optimization.

***Input**: Source node S, Destination Node D, Network Graph G (Z, E)*
***Output**: Trained Weights* w *of the DQN and optimal path selection from source to destination*
1. *Initialize the **DQN parameters** w with random weights.*
2. *Set the **target network parameters** $w' = w$, where w' is the target network weight.*
3. ***For each iteration** $i = 1$ **to** **N**:*
   - *If i is a multiple of target network update interval (T), update the target network: $w' = w$.*
   - *Sample a minibatch of experiences (s_j, a_j, r_j, s_{j+1}) from buffer D where s_j represents current state, a_j represents action state, r_j represent reward of state j and s_{j+1} is the next state.*
   - *For each tuple in the minibatch:*
     a. ***Compute Q-values** for all actions using the DQN: $Q(s_j, \cdot; w)$*
     b. ***Predict next-state Q-values** using the target network: $Q(s_{j+1}, \cdot; w')$.*
     c. ***Select the action** a* with the highest Q-value in the next state:*

$$a^* = \arg\max_a Q(s_{j+1}, a; w)$$

, $\max_a$ defines the best possible action in the next state

     d. *Forward packet to node j = a**
     e. *Observe reward r_i^t and new state s_j^{t+1}*
     f. *Store experience $(s_i^t, a_{ij}^t, r_i^t, s_j^{t+1})$ in replay buffer*
     g. *Update $i \leftarrow j$*
4. *Reward Calculation:*

   4.1. *For each neighbor j, compute reward:*

$$r_i^t = \begin{cases} \max, & \text{if } j = D \\ \alpha_1(1 - nl_j / P_{max}) + \alpha_2(1 - c_j / E_{max}) + \alpha_3(ld_{ij} / L_{max}) + \alpha_4(1 - d_{rel}), & \text{otherwise} \end{cases}$$

(use equations (3), (4), (5), (6), (7))
5. *Determine destination proximity*
6. ***Compute loss function(L)**:*

$$L = \frac{1}{2}\left(Q(s_j, a_j; w) - qtarget\right)^2$$

$$\text{where } q_{target} = r_j + \gamma \max_a Q(S_{j+1}, a; w')$$

Where:
- *$r_j \rightarrow$ reward received at the current step*
- *$Q(S_{j+1}, a; w') \rightarrow$ next-state Q-value predicted by the target network*

*Perform **gradient descent** on L to update the DQN parameters w.*
7. *Repeat until **convergence** is reached and return trained weights w.*

Algorithm 2. DRL-D2D.

***Input:** A graph network G= (Z, E), in which Z represents the nodes and E the edges; source node S, destination node D*
***Output:** The most efficient from S to D*
***Initialization:** All the nodes keep sending HELLO messages periodically to neighborhood nodes to update and keep their routing tables up to date.*

Step 1: Route Discovery
1. *Source node S sends out a route request (RREQ) and determines the next hop based on a Deep Q-Network (DQN) model with the highest Q-value.*
2. *Iterate until RREQ packet reaches destination node D: Every internal node passes the RREQ to the node of the highest Q-value via its trained DQN.*
3. *Upon receiving the RREQ, D sends a route reply (RREP) back to S along the reverse path of discovery.*
4. *Upon receipt of the RREP within the time limit by S: Update routing table with new path.*
5. *Else: Start new route discovery process (repeat steps 1-4).*

Step 2: Route monitoring and repair
6. *Whenever a node i does not get a HELLO message from its neighboring node j for a certain time interval, then T: i sends a route error (RERR) message and transmits it to the source node S.*
7. *When it gets the RERR, the source node S: Removes the existing route and initiates the route discovery process again from Step 1.*

***Return:** The ultimate and most efficient routing pathway from the source to the destination.*

The DQN offers significant advantages for 5G D2D routing, ensuring enhanced stability, efficiency, and scalability. By minimizing estimation errors, DQN improves routing stability, making it highly effective in dynamic 5G environments. Its intelligent path selection prioritizes low-energy, high-stability routes, leading to efficient resource utilization and optimized energy consumption. Additionally, faster Q-value convergence accelerates the learning and decision-making process, reducing overhead and improving adaptability. Designed to handle large-scale networks with high node mobility, DQN efficiently manages expansive state-action spaces,

making it an ideal solution for 5G D2D applications requiring robust, real-time routing optimization. For DQN setup, the Hyperparameters are illustrated in Table 4.

Table 4. DDQN hyperparameter setup.

Hyperparameters	Values
Discount Factor	0.95
Learning Rate	0.001
Batch Size	1.28
Replay Buffer Capacity	5000
Exploration Rate	1.0-0.01
Training Episodes	2500

5. Results and Discussion

5.1. Simulation Setup

Extensive simulations were conducted for the evaluation and analysis for the performance of the proposed DRL-D2D in 5G D2D networks. The simulation parameters are summarized in Table 5. The simulation was conducted and the simulation environment is created using the NS-3 simulator, integrated with Python for RL support and implementation. The environment simulates a 5G cellular network with D2D communication underlying the cellular infrastructure. The scenario presented in figure 5 consists of a fixed number of User Equipment's (UEs), direct D2D links enabled based on proximity and channel conditions. To model mobility and wireless dynamics realistically, the Random Waypoint Mobility model and 3GPP 5G channel propagation models were employed.

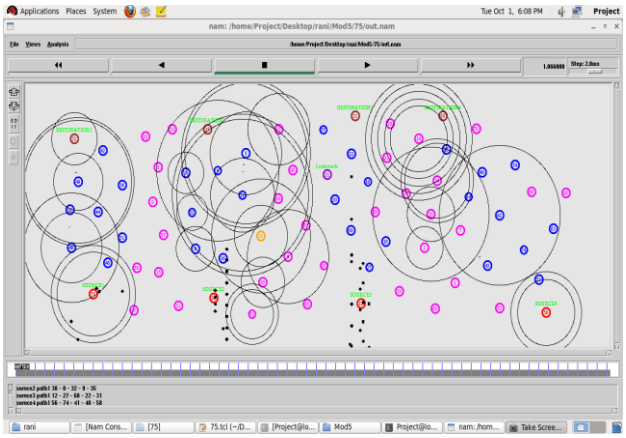

Figure 5. Network simulation windows.

Table 5. Simulation parameters.

Simulation parameters	Values
Routing protocol	Proposed DRLD2D
Simulation time	200 s with 200 iterations
Number of devices	49
Simulation area	500m×500m
Mobility model	Random waypoint (RWP)
Device speed range	0-30 m/s
Application traffic model	Constant bit rate (CBR)
Packet size	512 bytes
Data rate	11 Mbps
Wireless channel frequency	2.4 GHz
Transmission range	270 meters
Path loss model	Two-ray ground
Transmission power	31.623 mW
Energy model	Generic energy model
Learning rate	0.001
Discount factor (γ)	0.95
Exploration rate (ε) Decayed from 1.0 to 0.01	1.0 to 0.01
Communication model	3GPP TR 38.901 channel model

The state space includes link quality indicators SNR, RSSI, buffer occupancy, residual energy, and current position of the node. The action space consists of the possible next-hop selections and transmission parameters. The reward function was designed to maximize throughput while minimizing delay, energy consumption, and interference.

5.2. Performance Evaluation Metrics

To assess the efficiency of the DRLD2D, various performance evaluation metrics were considered, comparing it with conventional routing protocols. The following metrics were used:

1. **Throughput**: Measures the successful data packet delivery rate at the destination over the total transmission time. It is computed as Equation (12):

$$Throughput = \frac{\sum N_{Rec} \times 8}{\sum T_{simulation}} \times 1000 \ \ kbps \tag{12}$$

where N_{Rec} is the total number of received packets, and $T_{simulation}$ is the total simulation time.

2. **End-to-End Delay (EED):** Represents the average time taken for a data packet to travel from the source to the destination, including route discovery delay, queuing delay, and transmission delay as Equation (13):

$$EED = \frac{1}{N_{Rec}} \sum_{c \in N} Delay(c) \tag{13}$$

where Delay(c) is the delay of each packet and N_{Rec} is the total number of received packets.

3. **Packet Delivery Ratio (PDR):** The percentage of successfully delivered packets compared to the total sent packets, reflecting the efficiency of the routing protocol as Equation (14):

$$PDR = \frac{N_{Rec}}{N_{Sent}} \times 100 \quad (14)$$

Where N_{Sent} is the total number of packets sent in the network.

4. **Energy Consumption:** Computes the total energy consumption of all devices throughout the simulation as Equation (15):

$$Energy\ \ Consumption = \frac{1}{N}\sum_{c \in N} E_{Total}(c) \quad (15)$$

Where $E_{Total}(c)$ is the total energy consumed by each device, and N is the total number of devices. These metrics provide a comprehensive analysis of the proposed DRLD2D performance, ensuring efficient route selection, energy optimization, and scalability in 5G D2D networks.

5.3. Comparative Analysis

5.3.1. Throughput Evaluation

The performance of throughput parameter of proposed DRLD2D protocol was evaluated against AODV and Multipath Energy and QoS-Aware Optimized Link State Routing Protocol Version 2 (MEQSA-OLSRv2) under dynamic node densities[35]. As shown in Tables 6 and 7, DRLD2D consistently achieves efficient throughput across varying node densities. On average, DRLD2D delivers 35.56% better performance than AODV and 27.27% better performance as compared to MEQSA-OLSRv2 respectively. This enhancement in the performance of DRLD2D is attributed to the decision-making capabilities of the DRL model, which dynamically chooses optimal routes by leveraging both previous experiences and real-time network feedback. In contrast with the traditional protocols that are based on static or reactive routing strategies, DRL-D2D proactively changes itself in network topology, consequently reducing route discovery latency and packet loss which directly contributes to improved data throughput.

Table 6. Throughput (in kbps) of DRL-D2D, AODV, and MEQSA-OLSRv2 at different node density

Number of nodes	20	30	40	50	60
AODV	1586.20	2683.45	3304.32	5122.33	6784.71
OLSRv2	1680.85	2887.57	3591.56	5394.33	7134.23
DRL-D2D	2014.50	3499.36	4520.32	7193.74	9715.84

Table 7. Throughput comparison (in %) of DRL-D2D with AODV and MEQSA-OLSRv2.

Number of nodes	DRLD2D compared to AODV (%)	DRLD2D compared to MEQSA-OLSRv2 (%)
20	26.92	19.82
30	30.45	21.22
40	36.81	25.86
50	40.42	33.32
60	43.23	36.14
Average	35.56	27.27

By rigorous simulations it is found that DRL-D2D consistently delivers the highest throughput at every node as seen in figure. The graph of DRLD2D as shown in figure 6 begins at 2014.50 kbps at 20 nodes and steadily increases up to 9715.84 kbps till reaching towards 60 nodes, reflecting efficient data transmission efficiency and proper utilization of network resources. DRL-D2D's growth rate is significantly higher as compared to both the protocols, indicating its robustness and effectiveness under increasing network load. These results clearly establish DRL-D2D as a highly effective protocol for maximizing network throughput, particularly in dynamic and high-traffic scenarios.

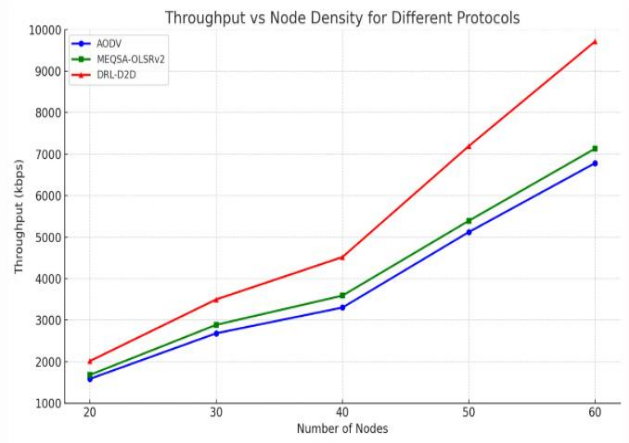

Figure 6. Comparison of throughput with varying node density.

5.3.2. End to End Delay Evaluation

The E2E delay evaluation of the proposed DRLD2D protocol was calculated and compared with AODV and MEQSA-OLSRv2 across varying node densities as presented in Tables 8 and 9, DRLD2D consistently provides significantly lower delay values at varying network densities. On average, DRLD2D provides 23.6% reduction

in delay compared to AODV and 19.94% reduction compared to MEQSA-OLSRv2.

This reduction in delay can be attributed to the learning (proactive) and predictive decision-making capabilities of the DRL agent. Unlike traditional protocols that reactively respond to change in topology, the DRL-D2D model anticipates network behaviour and dynamically makes a selection of more stable and less congested routes. Additionally, the integration of D2D communication within the RL framework reduces the dependency on intermediate nodes. This leads to lesser hops and retransmission, the two major factors which contributes to high delay in conventional routing protocols.

Table 8. End to end delay of DRL-D2D, AODV, and MEQSA-OLSRv2 at varying node densities.

Number of nodes	20	30	40	50	60
AODV	52.42	36.24	48.24	62.86	87.24
MEQSA-OLSRv2	49.55	34.07	45.22	59.47	82.58
DRL-D2D	42.75	28.32	34.86	43.23	60.29

Table 9. Percentage reduction in E2E delay of DRL-D2D in comparison with AODV and OLSRv2.

Number of nodes	DRLD2D compared to AODV (%)	DRLD2D compared to MEQSA-OLSRv2 (%)
20	18.45	13.33
30	21.84	15.21
40	27.77	17.22
50	31.23	24.53
60	30.86	29.42
Average	26.03	19.94

From a performance perspective, the delay for DRL-D2D starts at 42.75 ms at 20 nodes and increases only moderately to 60.29 ms at 60 nodes, illustrating excellent scalability and adaptability as shown in Figure 7. In contrast, AODV and MEQSA-OLSRv2 exhibit a sharper increase in delay over time. While MEQSA-OLSRv2 performs better than AODV, it still lags behind DRL-D2D in every instance. These results clearly establish that DRL-D2D not only reduces latency but also maintains consistent performance under varying node densities, making it highly suitable for delay-sensitive applications.

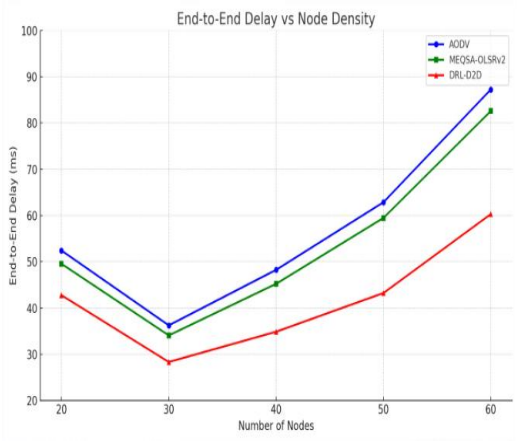

Figure 7. Variation of end-to-end delay at varying node density.

5.3.3. PDR Evaluation

In the simulation results presented below, the DRL-D2D protocol illustrates a significant improvement in PDR compared to traditional routing protocols such as AODV and OLSRv2 as shown in Table 10 and Table 11, especially when there is an increase in number of nodes. At 20 nodes, a PDR of 50.86% is achieved in case of DRLD2D, which is considerably higher as compared to AODV (38.28%) and MEQSA-OLSRv2 (40.24%). At 60 nodes, DRL-D2D reaches a PDR as high as 68.48%, while AODV and MEQSA-OLSRv2 records PDR of 52.32% and 59.72%, respectively. The average increase in PDR is 30.33% as compared to AODV and 24.94% as compared to MEQSA-OLSRv2 respectively as illustrated in Figure 8. The improved performance of DRL-D2D can be attributed to its incorporation of DRL, which enables nodes to make intelligent routing decisions based on real-time feedback.

Table 10. PDR (in %) of DRL-D2D, AODV, and MEQSA-OLSRv2 at different node density.

Number of nodes	20	30	40	50	60
AODV	38.28	42.46	45.47	50.64	52.32
MEQSA-OLSRv2	40.24	43.45	45.48	51.22	59.72
DRL-D2D	50.86	54.24	59.72	65.35	68.48

Table 11. Comparison of PDR of DRL-D2D with AODV and OLSRv2 (in%).

Number of nodes	DRLD2D compared to AODV (%)	DRLD2D compared to MEQSA-OLSRv2 (%)
20	32.8	26.39
30	27.74	24.83
40	31.33	31.31
50	29	27.58
60	30.8	14.6
Average	30.33	24.94

Unlike AODV, which is based on reactive route discovery, and MEQSA-OLSRv2, which makes use of periodic updates for topology information, DRL-D2D dynamically learns and analyses optimal paths by evaluating various factors such as link stability, congestion levels, and node mobility. This results in reduction in packet loss, least retransmissions, and better handling of network dynamics. In summary, DRL-D2D consistently outperforms both the protocols viz. AODV and MEQSA-OLSRv2 in terms of PDR due to its adaptive nature, self-learning-based routing strategy, making it a more efficient solution for densely populated and highly dynamic wireless networks.

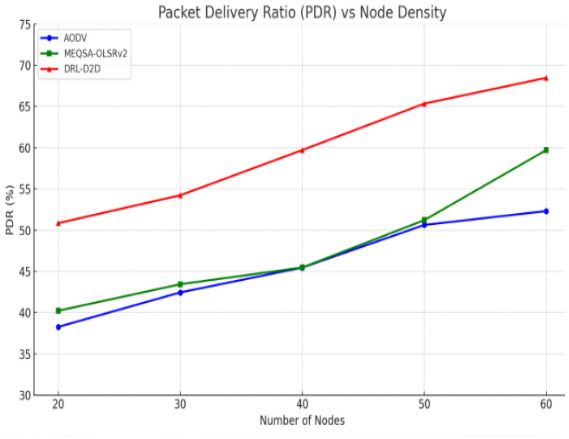

Figure 8. Comparison of PDR of various protocols with varying node density.

5.3.4. Energy Efficiency Evaluation

The analysis of energy consumption across DRL-D2D, AODV, and MEQSA-OLSRv2 protocols provides that DRL-D2D consistently has less energy consumption as compared to the traditional routing protocols as illustrated in Table 12 and Table 13, making it predominantly more energy-efficient. At 20 nodes, DRL-D2D has an energy consumption of 49.24 units, which is substantially lesser as compared to AODV (63.32 units) and MEQSA-OLSRv2 (60.37 units), reflecting a reduction of approximately 22% and 18% respectively. This efficiency persists as the network scales, with DRL-D2D maintaining lower energy usage even at 60 nodes, where it consumes 147.36 units compared to 152.45 units for AODV and 150.55 units for OLSRv2. Although the performance gap is narrower at higher node counts which can be seen in graph, DRL-D2D still is more energy efficient and have an edge in terms of energy savings.

Table 12. Energy consumption (in Joules) of DRL-D2D, AODV, and MEQSA-OLSRv2 at different node density.

Number of nodes	20	30	40	50	60
AODV	63.32	79.73	83.21	96.42	152.45
MEQSA-OLSRv2	60.37	78.28	81.44	96.60	150.55
DRL-D2D	49.24	67.41	74.71	93.30	147.36

Table 13. Energy efficiency comparison of DRL-D2D with AODV and MEQSA-OLSRv2.

Number of nodes	DRLD2D compared to AODV (%)	DRLD2D compared to MEQSA-OLSRv2 (%)
20	22.24	18.44
30	15.45	13.89
40	10.21	8.26
50	3.24	3.42
60	3.34	2.12
Average	10.86	9.2

This improvement in energy consumption in DRL-D2D is largely attributed to the adaptive and its learning-based methodology, which reduces the requirement for frequent route discoveries and control message exchanges that are common in AODV and MEQSA-OLSRv2 as shown in Figure 9. While AODV relies on reactive routing and MEQSA-OLSRv2 on periodic link-state updates, both of which lead to increase in overhead, DRL-D2D utilizes real-time feedback and past learning to make optimized routing decisions. In nutshell, it minimizes unnecessary transmissions and conserves energy more effectively.

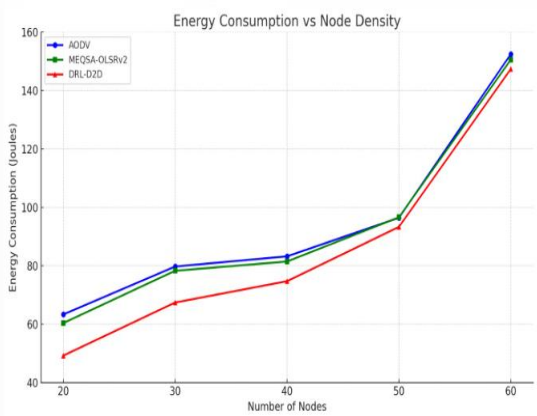

Figure 9 Variation in energy consumption with varying node density.

Overall, the DRL-D2D protocol offers enhanced scalability and making it more energy efficient, creating it particularly suitable for energy-constrained and dense network environments such as IoT, sensor networks, and 5G D2D communications.

6. Conclusions

For addressing the ever-increasing problem of efficient communication in D2D communication in 5G networks, a DRL based routing strategy DRLD2D is proposed. The proposed routing strategy captures the topology change between the network nodes (devices) by the improvement in message format and designs an efficient reward function by integrating the node and link state information for efficient and reliable routing. DRLD2D out performs the performance of the existing protocols AODV and MEQSA-OLSRv2 when tested upon various parameters such as throughput, E2E delay, PDR and energy consumption. The throughput of DRL-D2D is 35.56% higher as compared to AODV and it is 27.2% more efficient as compared to. MEQSA-OLSRv2 respectively. The E2E delay is 26.03% less as compared to AODV and it is 19.94% less as compared to MEQSA-OLSRv2 respectively. The PDR is 30.33% more as compared to AODV and it is 24.94% more as compared to MEQSA-OLSRv2 respectively. DRLD2D is 10.86% more energy efficient as compared to AODV and 9.2% more energy efficient as compared to MEQSA-OLSRv2 respectively. These results illustrate that DRLD2D is highly efficient in routing of packet from source to destination in D2D 5G networks and they provide a robust and efficient communication.

6.1. Limitations and Future Scope

Although the suggested DRLD2D approach has shown some notable gains in D2D routing efficiency, there are still some practical issues that must be resolved for the massive, large-scale deployment into ultra-dense real scenarios. Neural network training and ongoing policy updates result in non-negligible computational complexity when DRL is integrated. Direct deployment on resource-constrained D2D devices may be limited without hardware acceleration or distributed processing support, although it is possible in simulation and edge-assisted environments. There are still unanswered questions about scalability and signalling overhead in large deployments because the behaviour of DRLD2D under ultra-dense 5G and beyond-5G scenarios with massive D2D connectivity has not yet been thoroughly examined. The learning and routing framework of the current DRLD2D design does not specifically incorporate security, privacy preservation, or adversarial robustness. Since 5G D2D networks are susceptible to malicious activity and routing attacks, implementing secure and trust-aware DRL mechanisms is still a crucial step toward practical deployment.

The research papers provide various insights of reinforcement-based routing in 5G D2D networks for overall performance enhancement. However, this research can be extended to DRL based routing algorithms in 6G scenarios. Another variant could be integration of DRLD2D with edge computing and network slicing approaches will provide more effective and flexible network management. These combined approaches will lead to reduction in latency and improvement in the performance of D2D communication. Moreover, research can be done in application of DRL to provide adaptive security so that system can detect and eliminate the security threats in real time. By exploring these future directions DRLD2D can be more refined and polished for more effective, secure and robust D2D communication network for Gen Next communication networks.

References

[1] Abou-Rjeily C and El-Zahr S., "Deep Reinforcement Learning Based Relaying for Buffer-Aided Co-Operative Communications," *Physical Communication*, vol. 59, pp. 1-33., 2023. https://doi.org/10.1016/j.phycom.2023.102086

[2] Adu-Manu K., Amoako E., and Engmann F., "Advancements in Machine Learning-Enhanced Green Wireless Sensor Networks: A Comprehensive Survey on Energy Efficiency, Network Performance, and Future Directions," *Journal of Sensors*, vol. 2025, no. 1, pp. 1-25, 2025. https://doi.org/10.1155/2025/5242517

[3] Agyekum K, Boakye A., Opoku J., Agyemang J., and et al.," Resource Allocation in D2D-Enabled 5G Networks Using Multiagent Reinforcement Learning," *Journal of Computer Networks and Communications*, vol. 2024, no. 1, pp. 1-14, 2024. https://doi.org/10.1155/2024/2780845

[4] Arya G., Bagwari A., and Chauhan D., "Performance Analysis of Deep Learning-Based Routing Protocol for an Efficient Data Transmission in 5G WSN Communication," *IEEE Access*, vol. 10, pp. 9340-9356, 2022. DOI: 10.1109/ACCESS.2022.3142082

[5] Bunu S., Saraee M., and Alani O., "Machine Learning-Based Optimized Link State Routing Protocol for D2D Communication in 5G/B5G," *in Proceedings of the International Conference on Electrical Engineering and Informatics*, Banda Aceh, pp. 7-12, 2022. https://doi.org/10.1109/ICELTICs56128.2022.9932126

[6] Chen P., Liu S., Wang X. and Kamwa I., "Physics-Guided Multi-Agent Deep Reinforcement Learning for Robust Active Voltage Control in Electrical Distribution Systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 2, pp. 922-933, 2023. DOI: 10.1109/TCSI.2023.3340691

[7] Cui Y., Zhang Q., Feng Z., Wei Z., and et al., "Topology-Aware Resilient Routing Protocol for FANETs: An Adaptive Q-Learning Approach," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18632-18649, 2022. DOI: 10.1109/JIOT.2022.3162849

[8] Duong T and Binh L., "An improved method of AODV routing protocol using reinforcement learning for en-suring QoS in 5G-based mobile ad-hoc networks,". *ICT Express*, Vol. 10, no. 1, pp. 97-103, 2024. https://doi.org/10.1016/j.icte.2023.07.002

[9] Elleuch W., Sondi P., Meddahi A., and Lecomte S., "Evaluation of 5G Relay-Empowered and Device-to-Device Communications for Rescue Mission," *IEEE Access*, vol. 13, pp. 104614-104629, 2025. https://doi.org/10.1109/ACCESS.2025.3579872

[10] Goswami M., Panda N., Mohanty S., and Pattnaik P., "Machine Learning Tech-niques and Routing Protocols in 5G and 6G Mobile Network Communication System-An Overview," *in Proceedings of the 7th International Conference on Trends in Electronics and Informatics*, Tirunelveli, pp. 1094-1101, 2023. DOI: 10.1109/ICOEI56765.2023.10125697

[11] Gottam S and Kar U., "Graph Attention Transformer-Based Meta-Reinforcement Learning for Secure and Low-Latency D2D Routing in 6G Networks," *TechRxiv*, 2025. https://doi.org/10.36227/techrxiv.174803075.57648343/v1

[12] Guo H., Zhou X., Liu J., and Zhang Y., "Vehicular Intelligence in 6G: Networking, Communications, and Computing," *Vehicular Communications*, vol. 33, no. c, pp. 100399, 2022. https://doi.org/10.1016/j.vehcom.2021.100399

[13] Huang J., Yang C., Zhang S., Yang F., and et al., "Reinforcement Learning Based Resource Management for 6G-Enabled mIoT with Hypergraph Interference Model," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4179-4192, 2024. DOI: 10.1109/TCOMM.2024.3372892

[14] Huang Z., Li T, Song C, Li Z, and et al., "Joint Spectrum and Power Al-Location Scheme Based on Value Decomposition Networks in D2D Communication Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2024, no. 1, pp. 1-16, 2024. https://doi.org/10.1186/s13638-024-02393-1

[15] Huang, J., Yang Y., Gao Z., He D., and Ng D., "Dynamic Spectrum Access for D2D-Enabled Internet of Things: A Deep Reinforcement Learning Approach," *IEEE Internet of Things Journal*, vol. 9 no. 18, pp. 17793-17807, 2022. DOI: 10.1109/JIOT.2022.3160197

[16] Jayakumar S and Nandakumar S., "Distributed Resource Optimisation Using the Q-Learning Algorithm, in Device-to-Device Communication: A Reinforcement Learning Paradigm" *Results in Engineering*, vol. 23, pp. 1-14, 2024. https://doi.org/10.1016/j.rineng.2024.102462

[17] Kumari P and Sahana S., "Swarm Based Hybrid ACO-PSO Meta-Heuristic (HAPM) for QoS Multicast Routing Optimization in MANETs," *Wireless Personal Communications*, vol. 123, pp. 1145-1167, 2022. https://doi.org/10.1007/s11277-021-09174-9

[18] Lauric D, Luxey-Bitri A, Raes R, Rouvoy R, and et al., "A Critical Review of Mobile Device-To-Device Communication," *in proceedings of the 25th IFIP WG 6.1 International Conference, held as Part of the 20th International Federated Conference on Distributed Computing Techniques*, Lille, pp. 1-24, 2025. https://doi.org/10.1007/978-3-031-95728-4_1

[19] Long D., Wu Q, Fan Q, Fan P, and et al., "A Power Allocation Scheme for MIMO-NOMA and D2D Vehicular Edge Computing Based on Decentralized DRL," *Sensors*, vol. 23, no. 7, pp. 1-19, 2023. https://doi.org/10.3390/s23073449

[20] Lu Y., Wang X., Li F., and Huang M., "RLbR: A Reinforcement Learning Based V2V Routing Framework for Offloading 5G Cellular IoT," *IET Communications*, vol. 16, no. 4, pp. 303-313, 2022. https://doi.org/10.1049/cmu2.12346

[21] Malathy S., Jayarajan P., Hindia M., Tilwari V., and et al., "Routing Constraints in The Device-To-Device Communication for Beyond Iot 5G Networks: A Review," *Wireless Networks*, vol. 27, no. 5, pp. 3207-3231, 2021. https://doi.org/10.1007/s11276-021-02641-y

[22] Malik T., Malik K., Afzal A., Wang L., and et al., "RL-IoT: Reinforcement Learning-Based Routing Approach for Cognitive Radio-Enabled IoT Communications," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1836-1847, 2023. DOI: 10.1109/JIOT.2022.3210703

[23] Malta S., Pinto P., and Fernandez-Veiga M., "Using Reinforcement Learning to Reduce Energy Consumption of Ultra-Dense Networks with 5G Use Cases Requirements," *IEEE Access*, vol. 11, pp. 5417-5428, 2023. DOI: 10.1109/ACCESS.2023.3236980

[24] Mangal A., Rizvi M., and Khan S., Advanced Wireless Sensing Techniques for 5G Networks, Chapman and Hall/CRC. 2018. DOI:10.1201/9781351021746-1

[25] Moghaddasi K., Rajabi S., Gharehchopogh F., and Ghaffari A., "An Advanced Deep Rein-Forcement Learning Algorithm for Three-Layer D2D-Edge-Cloud Computing Architecture for Efficient Task Offloading in The Internet of Things," *Sustainable Computing: Informatics and Systems*, vol. 43, pp. 100992. 2024. https://doi.org/10.1016/j.suscom.2024.100992

[26] Nauman A., Jamshed M., Qadri Y., Ali R., and

Kim S., "Reliability Optimization in Narrowband Device-to-Device Communication for 5G and Beyond-5G Networks,". *IEEE Access*, vol. 9, pp. 157584-157596, 2021. DOI: 10.1109/ACCESS.2021.3129896

[27] Pandey S., Vyas M., and Mangal A., "Exploring 5G: Technologies, Challenges, and 5G for Enhancement of IoT," *Grenze International Journal of Engineering and Technology*, vol. 6, no. 2., pp. 92-100, 2020. https://research.ebsco.com/c/z3bf3p/search/details/xfim7gowtr?request-context=plink&db=aci

[28] Quy V., Chehri A., Quy N., Nguyen V., and Ban N., "An Efficient Routing Algorithm for Self-Organizing Networks in 5G-Based Intelligent Transportation Systems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1757-1765, 2024. DOI: 10.1109/TCE.2023.3329390

[29] Ron D and Lee J., "Learning-Based Joint Optimization of Mode Selection and Transmit Power Control for D2D Communication Underlaid Cellular Networks," *Expert Systems with Applications*, vol. 198, pp. 116725, 2022. https://doi.org/10.1016/j.eswa.2022.116725

[30] Sahu R., Rizvi M., and Mishra M., "Routing Overhead Performance Study and Evaluation of Zone Based Energy Efficient Multi-Path Routing Protocols in Manets by Using NS2. 35" *in proceedings of the International Conference on Advances in Technology, Management and Education*, Bhopal, pp. 88-93, 2021. DOI: 10.1109/ICATME50232.2021.9732762.

[31] Sahu R., Sharma S., and Rizvi M., "Energy Consumption Evaluation of ZBLE, AOMDV and AODV Routing Protocols in Mobile Ad-hoc Networks," *Advances in Wireless Communications and Net-works*, vol. 5, no. 2 pp. 41-51, 2019. https://doi.org/10.11648/j.awcn.20190502.11

[32] Sahu R., Sharma S., and Rizvi M., "ZBLE: Energy Efficient Zone-Based Leader Election Multipath Routing Protocol for MANETs," *International Journal Innovative Technological Exploration Engineering*, vol. 8, no. 9, pp. 2231-2237, 2019. https://doi.org/10.35940/ijitee.H6916.078919

[33] Sahu R., Sharma S., and Rizvi M., "ZBLE: Zone Based Efficient Energy Multipath Protocol for Routing in Mobile Ad Hoc Networks," *Wireless Personal Com-munications*, vol. 113, pp. 2641-2659., 2020. https://doi.org/10.1007/s11277-020-07345-8

[34] Sahu R., Sharma S., and Rizvi M., "ZBLE: Zone based Leader Selection Energy Constrained AOMDV Routing Protocol," *International Journal of Computer Networks and Applications*, vol. 6, no. 3, pp. 56-68, 2019. https://doi.org/10.5815/ijwmt.2019.05.05

[35] Sahu R., Sharma S., and Rizvi M., "ZBLE: Zone Based Leader Selection Protocol," *International Journal of Wireless and Microwave Technologies*, vol. 9, no. 4, pp.11-25, 2019. https://doi.org/10.5815/ijwmt.2019.04.02

[36] Sahu V and Sahu R., "Energy Efficient Multipath Routing in Zone-based Mobile Ad-hoc Networks: Mathematical Formulation," *International Journal of Mathematical Sciences and Computing*, vol. 8, no. 3, pp. 37-48, 2022. https://doi.org/10.5815/ijmsc.2022.03.04

[37] Sharma V., Mohapatra S., Shitharth S., Yonbawi S., and et al., "An Optimization-Based Machine Learning Technique for Smart Home Security Using 5G," *Computers and Electrical Engineering*, vol. 104, pp. 10843, 2022, https://doi.org/10.1016/j.compeleceng.2022.108434

[38] Ssengonzi C., Kogeda O., and Olwal T., "A Survey of Deep Reinforcement Learning Application in 5G and Beyond Network Slicing and Virtualization," *Array*, vol. 14, pp. 1-27, 2022. https://doi.org/10.1016/j.array.2022.100142

[39] Tilwari V., Dimyati K., Hindia M., Amiri I., and Mohmed Noor Izam T., "EMBLR: A High-Performance Optimal Routing Approach for D2D Communications in Large-Scale Iot 5G Network," *Symmetry*, vol. 12, no. 3, pp.1-23, 2020. https://doi.org/10.3390/sym12030438

[40] Wang C., Chai X., Peng S., Yuna Y, and Li G., "Deep Reinforcement Learning with Entropy and Attention Mechanism for D2D-Assisted Task Offloading in Edge Computing," *IEEE Transactions on Services Computing*, vol. 17, no. 6, pp. 3317-3329, 2024. DOI: 10.1109/TSC.2024.3495503

[41] Wang X., Hu J., Lin H., Garg S., and et al., "QoS and Privacy-Aware Routing for 5G-Enabled Industrial Internet of Things: A Federated Reinforcement Learning Approach," I*EEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 4189-4197, 2022. DOI: 10.1109/TII.2021.3124848

[42] Yang X., Yan J., Wang D., Xu Y., and Hua G., "WOAD3QN-RP: An Intelligent Routing Protocol in Wireless Sensor Networks-A Swarm Intelligence and Deep Reinforcement Learning Based Ap-Proach," *Expert Systems with Applications*, vol. 246, pp.123089, 2024. https://doi.org/10.1016/j.eswa.2023.123089

[43] Yu S and Lee J., "Deep Reinforcement Learning Based Resource Allocation for D2D Communications Underlay Cellular Networks," *Sensors*, vol. 22, no. 23, pp. 1-19, 2022. https://doi.org/10.3390/s22239459

[44] Yu Y and Tang X., "Hybrid Centralized-Distributed Resource Allocation Based on Deep Reinforcement Learning for Cooperative D2D Communications,". *IEEE Access*, vol. 12, pp. 196609-196623, 2024. https://doi.org/10.1109/ACCESS.2024.3521590

[45] Zhang C., Wu C., Lin M., Lin Y., and Liu W.,

"Proximal Policy Optimization for Efficient D2D-Assisted Computation Offloading and Resource Allocation in Multi-Access Edge Computing," *Future Internet*, vol. 16, no. 1, pp. 1-17, 2024. https://doi.org/10.3390/fi16010019

[46] Zhi Y., Deng X., Tian J., Qiao J., and Lu D., "Deep Reinforcement Learning-Based Resource Allocation for D2D Communications in Heterogeneous Cellular Networks," *Digital Communications and Networks*, vol. 8, no. 5, pp. 834-842. 2022. https://doi.org/10.1016/j.dcan.2021.09.013

Anu Mangal is a Research Scholar in the department of Electrical and Electronics Engineering, National Institute of Technical Teachers Training and Research Bhopal. She has published many research papers. Her area of interest includes 5G, Mobile Adhoc Network, Artificial Intelligence and Machine Learning.

Anjali Potnis is a Professor at Department of Electrical and Electronics Engineering, National Institute of Technical Teachers Training and Research Bhopal. She has got a total of 18 years of teaching experience. She has published many research papers. Her area of interest includes Digital Image Processing, Digital Signal Processing, Artificial Intelligence and Machine Learning.

Murtaza Abbas Rizvi is a Professor at Department of Computer Science and Engineering, National Institute of Technical Teachers Training and Research Bhopal. He has got a total of 25 years of teaching experience. He has published many research papers. His areas of interest include Adhoc Network, Cryptography, Big Data, Artificial Intelligence and Machine Learning.