International Journal of Computer Networks and Applications (IJCNA)
DOI: 10.22247/ijcna/2025/28 Volume 12, Issue 3, May – June (2025)

REVIEW ARTICLE

A Comprehensive Review of Reinforcement Learning Based Routing Algorithms for Wireless Sensor Networks

Archana Chaudhari
Instrumentation Engineering Department, Vishwakarma Institute of Technology, Pune, Maharashtra, India.
archana.kchaudhari@gmail.com

Vivek Deshpande
Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India.
vsd.deshpande@gmail.com

Divya Midhunchakkaravarthy
Department of Postgraduate Studies, Lincoln University College, Petaling Jaya, Malaysia.
divya@lincoln.edu.my

Received: 14 April 2025 / Revised: 19 June 2025 / Accepted: 28 June 2025 / Published: 30 June 2025

Abstract – Wireless Sensor Networks (WSNs) are commonly utilized in remote environments. Intelligent sensors are embedded to monitor numerous parameters, including pressure and temperature. The sensors, known as nodes, gather data and transmit it to the base station. Nodes in WSNs must acquire information from their environment due to their dynamic nature. In WSN, energy economy and network longevity are essential goals that routing can facilitate. The changing nature of WSNs complicates adaptation of conventional routing approaches and computational intelligence systems. Reinforcement Learning (RL), a subset of computational intelligence approaches, facilitates the selection of future actions based on insights acquired from previous encounters with the environment. Consequently, RL is often employed to develop routing algorithms in WSNs. Recent research indicates that employing a reinforcement learning-based Q-learning technique can yield the most energy-efficient pathways in WSNs. The document presents a comprehensive examination of routing methodologies utilizing RL in WSNs. The Q-learning strategy is extensively discussed as it underpins other RL strategies. The paper elaborates extensively on the selection of state-action pairs and rewards in Q-learning across various literary works. The experimental study presents optimization policies and challenges for enhancing Quality of Service (QoS).

Index Terms – Reinforcement Learning, WSNs, State-Action Pair, Reward Function, Value Function, Policy Function, Exploration, Exploitation.

1. INTRODUCTION

Remote monitoring systems utilize Wireless sensor networks (WSN). A multitude of nodes exists within WSN. The nodes are formed using several sensors, including those for motion, acoustics, temperature, pressure, chemical detection, weather, and optical. They are utilized gather information for military, healthcare, defense, agricultural, and several other applications. These sensors monitor many parameters, including humidity and temperature. Figure 1 represents a WSN.

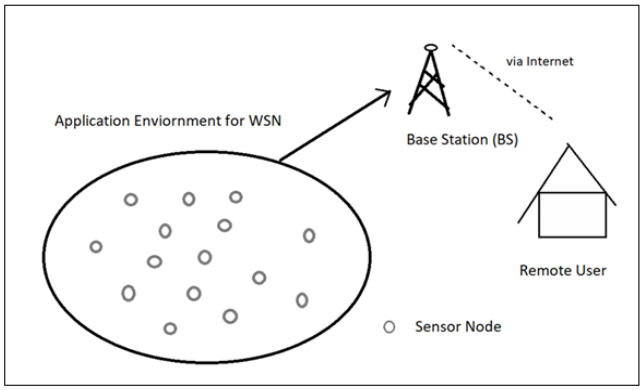

Figure 1 Illustration of WSN

WSNs are predominantly utilized for various purposes in severe and fluctuating environmental conditions [1]. WSNs encounter numerous challenges, such as memory constraints, limited energy resources, restricted bandwidth, and computational complexity, due to their application in varied contexts [2]. Thus, it is seen that the energy of wireless sensor networks deteriorates [3]. Consequently, the energy of WSNs degrades. Enhanced WSN performance necessitates the

REVIEW ARTICLE

resolution of difficulties related to routing, quality of service (QoS) [4], fault detection, energy harvesting [4], anomaly detection [4], clustering [5], security [5], location [4], dependability [5], and data aggregation [4][5].

To enhance WSN performance, it is essential to comprehend the changing network conditions and adapt accordingly, due to changing aspects of network and mobile nodes [1]. Sensors collect sensed data, extract pertinent information, process and transmit it to the base station (BS). Should the energy of any node in any of these phases be depleted, the network ceases to function and the data is irretrievably lost. Consequently, it is essential to manage the provided data in an energy-efficient manner to prolong the network's longevity.

In WSNs, optimizing network durability and energy efficiency remains a critical objective, and the routing technique provides an effective means to achieve both. The primary objective of routing algorithms is to transmit data with maximal efficiency while considering the dynamic topology of the nodes [5]. Moreover, due to the extensive deployment of sensors, a substantial volume of data necessitates transmission, processing, and reception from each sensor to the base node. A novel domain within computer science, known as machine learning (ML), has arisen in recent years. It uses data for learning, and its algorithms and methodologies demonstrate the capacity to predict outcomes [6, 7, 8, 9]. They are classed according to the manner in which information is extracted from the environment. Machine Learning is employed in WSNs to enhance performance in areas such as energy management and optimization, network lifetime management, congestion control, routing, Quality of Service monitoring, and security management [10, 11, 12]. Machine learning methodologies facilitate comprehension of the dynamic characteristics of sensor networks and enhance connection quality, network performance, energy efficiency, and service quality while optimizing network resources [5]. Machine learning algorithms can facilitate the analysis of this huge data volume with significantly reduced time and precision [2]. Hence, the demand for learning-based, computationally intelligent routing algorithms is increasing.

Applications devoid of datasets utilize a form of machine learning known as RL. A systematic model based on the collaboration amongst the environment and RL is presented. Considering dynamic characteristics of WSNs, this RL function is highly advantageous. RL can optimize routes to conserve network resources by comprehending the dynamic characteristics of sensor networks.

1.1. Purpose of the Research Study

It is outlined as:

- Examine recent developments in RL-based routing techniques for WSNs.
- To investigate the contribution of RL techniques to improving routing efficiency, energy optimization, and longevity of network in WSNs.
- To examine necessity and impetus for intelligent and decentralized routing approaches in resource-deficit WSN environments.
- To classify and evaluate reinforcement learning-based methodologies according to learning models, scalability, and adaptability to fluctuating network conditions.
- To create a comprehensive taxonomy of RL techniques utilized in WSN routing.
- To conduct a comparative examination of several RL algorithms regarding their effectiveness and applicability to WSN issues.

1.2. Comprehensive Contributions of the Research

RL thus provides promising answers by facilitating adaptive, self-learning routing algorithms for WSNs. The review paper aims to thoroughly evaluate current developments in reinforcement learning-based routing strategies for WSNs. The purpose is to investigate how RL approaches enhance routing efficiency, energy management, and network longevity. The impetus arises from the increasing demand for sophisticated, decentralized routing systems in resource-limited sensor networks. We rigorously analyse various RL methods within the setting of WSNs. The document classifies and contrasts methodologies according to learning models, scalability, and real-time adaptation. Our contributions encompass a comprehensive taxonomy, comparative analysis, and identification of research deficiencies. This establishes the groundwork for subsequent advancements in intelligent routing for WSNs with RL.

Section 1 discuss an outline objectives and contributions along with brief introduction to WSNs. Section 2 offers a comprehensive overview and a fundamental introduction to reinforcement learning. Section 3 illustrates Q-routing, a fundamental routing mechanism based on reinforcement learning. Section 4 encompasses a review of literature on RL-based routing techniques for WSN. Section 5 experimentally analyses the strengths and shortcomings of RL-based routing algorithms for WSN. Section 6 addresses the issues associated with reinforcement learning-based routing and optimization strategies. Section 7 presents the outcomes and challenges of RL-based routing for WSNs.

2. RL INTRODUCTION

RL offers a systematic approach to acquiring knowledge about the environment through previous encounters and facilitating informed decision-making in the future. Figure 2 illustrates the RL process.

International Journal of Computer Networks and Applications (IJCNA)
DOI: 10.22247/ijcna/2025/28 Volume 12, Issue 3, May – June (2025)

REVIEW ARTICLE

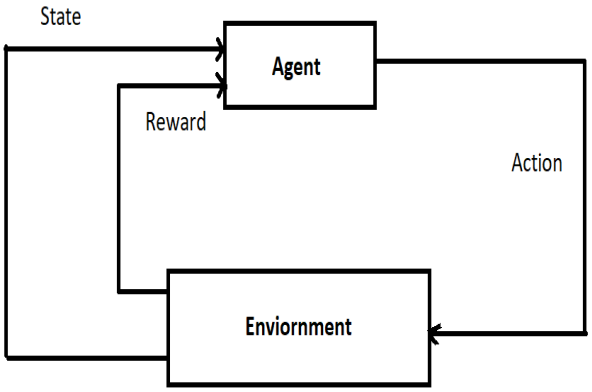

Figure 2 RL Methodology

An agent, commonly referred to as a learner, interacts with the environment in RL. The agent selects actions to implement in the environment according to its present state after acquiring knowledge of the environments. The policy specifies the state for which the agent will behave. In reinforcement learning, an agent executes an action inside the environment and then receives a response from it. This response is termed feedback or reward. The reward is a consequence of a recent action executed by the agent, dictated by its current state and knowledge, as the agent is in a continual learning process.

Rewards can be categorized into two types: short-term and long-term. The value-function denotes the long-term advantages. It illustrates potential rewards for the agent's present state. A reward, whether positive or negative, can be depicted by the value function. The worth of present condition is assessed by the agent utilizing a value function.

The literature summarizes both techniques based on the model model-free in RL. Techniques based on model computes the environmental model utilizing state transition probabilities. The agent acquires this model of the environment. This strategy accelerates learning due to the reusability of the data stored in the model.

The starting environment is a crucial determinant in this strategy [5]. The greedy approach refers to the model-based strategy wherein the agent operates to maximize rewards without concern for the potential repercussions of its actions.

The agent learns in the model-free method solely by interaction with the environment and the acquisition of experience. Q-learning and policy gradient methods are employed to enhance this experiential learning. Through the iterative execution of actions and subsequent analysis of outcomes, these algorithms enhance or adjust their policy to optimize rewards.

The model-based approach is effective in static environments, as the agent can predict the reward of an action prior to its execution. Model-free methodologies enable agents to interact with their environment and acquire knowledge from their activities. Consequently, model-free approaches are extensively employed in dynamic real-time environments, such as routing for WSNs.

In reinforcement learning, both exploration and exploitation strategies are essential. The agent explores its environment while selecting actions that utilize reward knowledge [13]. In exploitation, the agent selects actions depending on the established probability of rewards. Exploration and exploitation techniques are balanced to enhance the accumulation of rewards over time.

Heuristic search strategies are predominantly employed during exploration and exploitation phases [5]. The ε-greedy search is a widely utilized heuristic search method. In ε-greedy search, the agent seeks the optimal action by exploiting the search space with a probability of 1-ε and exploring all existing actions in the present state with a probability of $0 < \varepsilon \leq 1$.

Several routing strategies for WSNs are founded on RL-based routing algorithm, and will be discussed in the subsequent section.

3. A BRIEF OVERVIEW OF Q-ROUTING

Buoyan and Littman investigated the RL-based Q-learning known as Q-routing [14]. The Q-learning method is used in wired networks for Q-routing. It learns the routing policy to lessen the hops a packet will make on well-traveled paths.

Q-learning helps to improve agent policy without prior knowledge of the model. As a result, it adheres to RL's model-free methodology and is based on action-value function. The action-value function is important to assesses how well a state achieves a specific action [5].

Q-routing involves a separate phase of learning for the agent. During each phase, the agent is in state s_n executes an action, obtains a reward, and transfers on to the next stage s_{n+1}. Equation (1) is revised action value.

$$Q_n(s_n,a_n) = (1-\alpha)*Q_n(s_n,a_n)+\alpha[R_n + \gamma * \max_{a\varepsilon A}[Q_{n-1}(s_{n+1},a_n)]] \quad (1)$$

where α signifies learning-factor and γ discount-factor, which allows for adjustment for long-term benefits, falls between 0 and 1. In this method, preliminary Q-values are assumed.

Nodes denote an existent state in Q-routing. Action is the process of choosing the best node to be the next forwarder based on the Q-value. Reward in RL stands for packet transmission cost [5]. Deliveries between two nodes and link costs between two nodes are the incentive for Q-routing. The

REVIEW ARTICLE

neighbor node's local information determines the reward. During Q-learning, Q-value of respective node is calculated using equation (1). Equation (2) represents the function that defines the reward in Q-routing.

$$R_t = qt_i + TxT_{(i,j)} + \theta_j(d) \quad (2)$$

where qt_i is time, it takes for packet P in node $i's$ queue, $TxT_{(i,j)}$ denotes broadcast time in-between nodes i and j and $\theta_j(d)$ is the estimate of amount of residual time to destination d denoted as in equation (3) as,

$$\theta_j(d) = \min_{k \epsilon N_g(j)} \theta_j(d,k) \quad (3)$$

where $N_g(j)$ is the set of node $j's$ neighbors.

Shortest distance is chosen because the better the path amongst nodes the better the transmission of packet P, and the lower the delay.

Therefore, using equations (2) and (3), equation (1) is rewritten for Q-routing, including the reward, as equation (4).

$$Q_i(d,j) = (1-\alpha) * Q_i(d,j) + \alpha * (qt_i + TxT_{(i,j)} + \theta_j(d)) \quad (4)$$

where Q_i is Q-value matrix of node i which is randomly initialized at the beginning.

In Q-routing Q_i denotes the end-to-end delay, Q-value is linked to every node neighbor i. Therefore, in Q-routing for node i to transmit packet P to neighbor, it selects a node amongst neighbors say node j because it will have lowest Q-value calculated based on equation (4). The node selected to be next forwarder is node having lowest Q-value for optimal routing path. This is because if we observe the Q-learning equation, the immediate reward of Q-routing represents the link cost and delay. The lower the cost of the link and the less the delay of the next forwarder node the better the path.

Q-routing converges to the best action-value. Prior to reaching the best answer, it exhibits a gradual convergence. The network topology and traffic pattern information do not need to be known beforehand because Q-routing is a model-free technique that effectively finds the best routes in constantly changing situations. The flow chart for the Q-routing method is shown in Figure 3.

In the literature, several methods for WSNs are based on Q-routing. The RL-based protocols are actually improved iterations of the fundamental Q-routing for WSNs. The work's main goal is to talk about RL techniques for WSNs. A review of RL techniques is presented in a number of publications [1][5][15][16].

The following are some ways in which the suggested work differs from the methods mentioned in the literature:

1. Methods of routing based on RL
2. Selection of state actions and rewards of various RL-based routing strategies in the literature on WSN performance
4. The impact of various RL-based routing techniques' state action and incentives on QoS parameters, as well as how well they function and improve
5. Analyses of the performance of different RL-based routing strategies and obstacles to improving QoS.

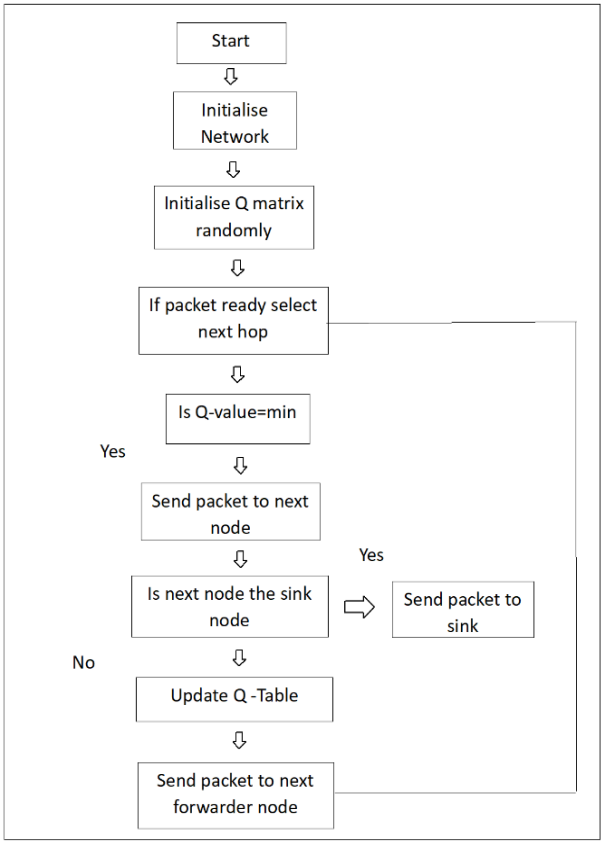

Figure 3 Flowchart of Q-Routing

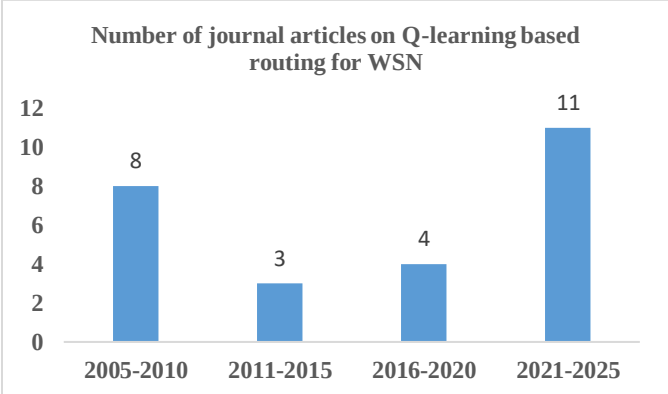

Figure 4 Number of SCI, Scopus, Web of Science Journal Articles Published from 2005 to 2025 Presented in the Study

REVIEW ARTICLE

An overview of RL-based routing methods for WSNs is discussed in the works. Selected works published in SCI, Scopus, and Web of Science journals between 2005 and 2025 comprise the publications cited in the survey. The number of journal articles from SCI, Scopus, and Web of Science compared to number of years is illustrated in Figure 4.

4. SUMMARY OF WSNS USING ROUTING PROTOCOLS BUILT ON RL

This section discusses the routing strategies employed in RL literature.

4.1. AdaR: Adaptive Routing with Reinforcement Learning for Sensor Networks

AdaR utilizes RL approach which is not dependent on the model and exhibits reactivity and adaptability in its routing strategy. It establishes an on-demand routing path to accommodate frequent topological changes and variable communication conditions [17]. For each sensor neighbor s' in AdaR, there exists an action a such that $P(s \mid s, a) = 1$, with the sensor itself representing a state s. Each action executed results in a packet being transmitted to the corresponding neighbor. The state-action pair encompasses the hop count amongst neighbor nodes and base station, remaining energy, and several routing paths that intersect neighboring nodes. Each packet is initiated by a data node. The subsequent node is chosen from the neighboring nodes according to the existing Q-values. Exploration is employed to execute actions with the maximal Q-value. The technique proposes three distinct types of rewards: a maximum reward upon reaching the base station, a neutral reward when the hop count exceeds a certain threshold, and a negative reward when a loop is detected in the path. The superiority of routing path dictates the instantaneous reward. The calculation is derived from the cumulative path reward, factoring in both the path length and an additional feature combination.

The major elements of AdaR include the utilization of LSPI and linearly weighted value-function of each state-action pair. Consequently, the method functions adaptively, modifying the Q-values and acquiring the weights from environmental data. The proposed adaptive routing utilizing RLoptimally leverages data. To enhance network longevity, the proposed method's LSPI attains a balance among several optimization objectives. AdaR [18] does not convert the requirements for QoS mapping into concrete Q-values.

4.2. UWB Wireless Sensor Network Geographic Routing Protocol Based on Reinforcement Learning

The authors proposed a RL approach built on spatial routing for ultra-wideband WSNs [19]. A variety of geographic routing protocols are discussed. Their scalability has resulted in their engagement. The literature indicates that GPSR [20] is the superior geographic routing protocol. The inadequate energy efficiency of GPRS constrains its efficacy in the WSN area. The authors examined a geographic routing technique built on RL in ultra-wideband WSNs. One advantage of UWB for global routing is its capacity to precisely predict the location of UWB nodes [21]. This model-free approach focused on the reward function, aiming to reduce the packet delivery while enhancing longevity of sensor network. The state is the entity now receiving or generating data packets. Based on the acquired Q-values, neighbor node is chosen. The Q-value in routing signifies expected overall reward. The Q-value is adjusted based on the acquired reward. The normalized progression to sink via the subsequent hop and the normalized energy of that hop are the basis for the reward function of RL algorithm. The subsequent forwarder node exhibiting the greatest relative advancement and the most remaining energy is selected to get the maximum reward. A fixed reward is also considered when a node reaches the sink in one hop. The reward function also indicates a warning signal for low energy. Should the node possess sufficient energy, it transmits the packet to neighboring nodes once its energy diminishes below a designated threshold, concurrently imposing a negative reward on its neighbors. Geographic routing is beset by "void" concerns. The packets are discarded by the node in the void node situation. The proposed method selects a negative reinforcement to tackle the void issue.

4.3. Wireless Sensor Networks Using Q-Probabilistic Routing

Arroya Valles et al. presented a strategy utilizing a Bayesian decision model combined with RL to overcome limitations of sensor energy constraints and frequent, random topological alterations. The Q-Probabilistic Routing (QPR) method utilizes local information among neighboring nodes and decisions derived from the delayed rewards of prior actions to facilitate intelligent routing. QPR is an algorithm designed for opportunistic routing strategies. It adapts to the network's dynamics to incorporate routing decisions. The assessment of a global cost metric enables it to adapt to changing network dynamics. RL rules are employed to locally acquire global cost measurements through interaction with neighboring nodes. Bayesian decision models are employed in the proposed technique to facilitate decision-making following local learning. Global cost metrics are computed greedily using the cost measure Q/P, where Q represents the expected retransmissions (ETX) to a specified node and P denotes the enhancement in that hop. It enhances the retransmissions ETX by considering energy consumption, network quality, and latency, instead of minimizing the hops QPR. Bayesian decision approaches provide predictive assessments based on candidate profiles, message significance (I), energy per transmission (ET), and energy per reception (or idle mode) (EIR), derived from historical data. The Q-values are revised based on environmental data. The QPR algorithm consists of: forwarding and information updating steps. Alternative candidate nodes transmit messages to the node on the

REVIEW ARTICLE

forwarding pages. The message is either destroyed if the forwarding criterion is unmet or transmitted using the Bayesian model to assess the likelihood criteria if the candidate nodes with high relevance transmit it. The neighbour's profile is updated upon the successful completion of the forward. RL is employed to ascertain the cost metric disseminated by each neighboring node to revise the Q-values of all nodes throughout the information update phase. Consequently, the deferred expense of prior routing decisions and the data exchanged among neighboring nodes are utilized to train the QPR algorithm [22].

4.4. FROMS: Optimization of Multiple Sinks in WSN via Feedback Routing

The RL-based FROMS protocol was proposed for multicast routing in WSNs [23]. Q-learning is employed toward illustrating the routing of several sinks. In a multiple sink situation, each node operates as a self-regulating agent, assimilating information from its surroundings. An agent state is defined as a tuple comprising the sinks that packets must reach and the requisite hops to each sink across all proximate nodes. The proposed methodology delineates the potential actions the agent may do to achieve the goal. The objective is to provide a singular route for a data packet. Q-learning facilitates the determination of the reward for the executed activity. The hop count for transmission from the agent to all sinks is estimated by the Q-value for the proposed multiple sink model. The ideal strategy is indicated by low Q-values, which diminish over the Q-learning process. The overall hops to sink for each node is computed initially. These are subsequently subtracted from the total number of hops, as broadcast communication is employed to relay information to neighbors and enable neighbors to access following hops. FROMS optimizes numerous sinks in the WSN via feedback routing.

4.5. The QoS-Aware Routing Protocol RL-QRP

Quality of Service (QoS) route paths are obtained by a distributed Q-learning mechanism within the reinforcement learning-based QoS-aware routing protocol [17]. Sensor nodes in distributed Q-learning independently establish optimal routing strategies based on historical experiences and incentives. Each node in the RL-based QoS-aware routing system signifies a state, and each adjacent node possesses a corresponding action. The operation signifies that a packet has been dispatched to the relevant neighboring node. The eminence of action at the respective state dictates the Q-value at each node, regarded as its routing table [17]. Actions transmitting the data packet are utilized to compute the rewards. A positive incentive will be awarded to the node upon the successful forwarding of data packets. Besides the anticipated future payout, node's Q-value is influenced via reward. ACK configuration is employed to facilitate the reinforcement of an action. The corresponding node Q-value dictates the ensuing decisions. Consequently, nodes utilize prior interactions and rewards to ascertain optimal paths, facilitating adaptability in dynamic situations [24]. The primary objective of RL-based QoS-aware routing is the providing Quality of Service (QoS). Proposed study delineates the QoS criterion for delivery of packets and latency. QoS criteria is validated and documented by node in Q-table prior to transmitting packet to its adjacent node. Node having maximum Q-value receives data. Each forwarding operation alters the packet's Quality of Service criteria. To establish a connection status table at individual nodes, the diverse types of QoS requirements must be standardized by a weighted function. By examining the Q-value table, nodes can evaluate the duration required to communicate data packet hence packet delivery ratio (PDR).

4.6. MRL-QPR

This RL technique extends QoS-aware routing built on RL and accommodates multiple agents [25]. The sensor node in RL-based routing does not consider data exchange with other nodes, rendering global optimization unattainable. Local network information exchange with nearby neighbors is incorporated into MRL-QPR, which employs many agents to achieve globally optimal performance. In the proposed methodology, sensor nodes function as agents, while wireless channel and quantity of packet placements constitute the environment. The information is transmitted to the optimal node. Each node use Q-learning to ascertain the optimal action for state-action pairings.

The MRL-QPR operates by granting the node acting as an agent a positive reward for each data packet forwarding action executed. The compensation fluctuates based on the packet loss rate or latency. Q-values are modified at each node by incorporating the anticipated long-term reward and the immediate reward. Every node in the MRL-QRP network computes a weighted aggregate of its own and other nodes' anticipated future rewards. The fundamental difference between the RL-QRP and MRL-QRP algorithms is as follows. The MRL-QPR node relies on local network data, engages with its neighboring node, and assimilates the weighted sum of prospective benefits with the rewards from other network nodes.

Accordingly, during packet transmission, node considers the following factors: i) its anticipated QoS performance, ii) the expected QoS performance of the subsequent nodes in the network, and iii) the effect of transmission on its neighboring nodes. Consequently, the nodes cooperate to determine behaviours that optimize global benefits by assigning suitable weights and incentives. The modified Q-values influence subsequent routing decisions. This compilation of Q-values at each node functions as a routing table. The node possessing the greatest Q-value receives the packet upon convergence.

REVIEW ARTICLE

4.7. QoS-RSCC

QoS-RSCC [26] is an extension of the techniques initially introduced by Liang et al. Al. QoS-RSCC proposes cooperative communication for resource-limited WSNs with adaptive relay selection. The proposed method is executed by multi-agent reinforcement learning. The broadcast component of cooperative communication is employed to exploit the diversity of time and space [27][28]. Users collaborate by transmitting data packets to one another. One relay is chosen to collaborate with the communication from other relaying candidates. The optimal relay is chosen for cooperative communication. The channel state information (CSI) is a valuable parameter for optimal relay selection. The source or destination of the probable relays available for cooperative communication is presumed to possess complete or partial Channel State Information (CSI). The authors in the proposed study pick the operating condition of relaying candidates, including duty cycle, processing and queuing delays, and mobility patterns, with CSI in the dynamic WSN environment. To select the optimal relay from numerous candidates that satisfy QoS requirements, an ideal relay can be determined based on packet outage possibility and channel efficacy in QoS-RSCC for multi-hop routes between each pair of routers. Cooperative transmission uses direct transmission relaying transmission. Relaying transmission is employed just if direct transmission fails. To encode and retransmit the packet to destination when relaying transmission is initiated best relay is chosen. Decision-makers are promptly compensated upon the selection of the optimal relay. The decision-maker employs anticipated long-term benefits and immediate rewards to modify the selection policy. Consequently, a sequence of trial-and-error contacts within a changing environment does not facilitate sub-optimal selections; rather, it endorses the ideal relay assignment option [26]. Relay candidate that significantly impacts outage possibility and channel efficacy is likely to be chosen as the optimum relay. The proposed method combines the route-discovery strategy with the selection of relay candidates. AODV serves as the basis for the routing algorithm that facilitates Quality of Service (QoS).

4.8. EQR-RL: QoS Routing Based on RL that is Energy Mindful

The authors employ RL [29] to propose an energy-efficient QoS routing algorithm whereby nodes represent states and forwarding data packets represents action. Q-table is revised according to feedback from each neighboring node concerning the time required to forward a packet. Data packets can be routed to sinks by utilizing the anticipated minimal future cost of each node. The feedback data is integrated into the data packet header to minimize protocol overhead. EQR-RL manages mobility and failure recovery while communicating the path cost, notwithstanding the adverse feedback from neighboring nodes. The packet forwarding node in EQR-RL calculates the round-trip time for transmitting packets to subsequent node. Three actions are performed by node receiving the packet: compresses least end-to-end delay estimate from route table into packet header, selects next hop for the data packet, and modifies the packets by altering the receiver address. Ultimately, ascertain its remaining energy, which is maximum total of remaining energy in route table, and the minimum total of energy captured in the packet. The selection of the subsequent data hop is incorporated into the exploration policy. This strategy considers irregularities in network traffic dynamics. The strategy mitigates congestion and prolongs the network's lifespan by distributing traffic instead of directing packet transfers to the optimal neighbor, while considering the current QoS parameters of each link. The QoS parameter in proposed study is end-to-end delay. Through reinforcement learning, it selects a neighboring node for the subsequent hop based on the current network state, rather than opting for neighbor with optimal QoS.

4.9. QSGrd

The proposed work integrates gradient-based routing with Q-routing [30]. The approach comprises of Q-learning and transmission gradient. Its architecture exhibits a protocol stack with the application, network, data link, and physical layer. Exchange of data takes place at physical level.. Payloads in network traffic are denoted by data frames, whereas metadata essential for network traffic management is indicated by ACK and status frames. The physical layer emulates transmission errors.

The data link layer executes the communication sequence, manages outgoing data and ACK frames, prevents packet collisions, and retransmits data frames if an ACK frame is not received.

The network layer assists a node in making decisions regarding the forwarding of a data frame it receives. To maintain a current record of adjacent nodes, forwarding decisions are incorporated utilizing the information provided in the Status Frame. Consequently, the physical layer assimilates environmental data, Status Frames revise Q-values, and the network layer implements routing decisions. Data frames are employed at the application layer.

The foundation for the proposed methodology, with the conceptual framework, is a transmission gradient. This approach estimates the chance of successful transmission between two nodes using a probability function.

This model-free RL strategy relies on positive incentives for Q-value updates, with nodes representing the states. In the broadcast status frame packets, each node retains its distinct Q-value. The Q-value, representing the node's utility degree according to the reward function, is crucial. The Q-value

REVIEW ARTICLE

rewards are contingent upon the performance of the sink, energy efficiency, and the minimization of residual transmissions required for delivery of packets to base. Movement towards sink is determined using transmission gradient, which determines the chance of a successful transmission to the subsequent node. The transmission gradient in QSGrd directs data via the most efficient tracks to sink, whereas Q-learning enhances routes for balancing packet delay and energy expenditure through its incentive function [30]. The Q-Learning factor, in conjunction with the residual energy levels at those nodes, also consider the average transmissions to base station for each node. Consequently, QSGrd acquires routes that optimize the remaining energy of nodes for these routes while decreasing their length.

4.10. Wireless Sensor Networks with QoS and Energy-Aware Cooperative Routing Protocol for Wildfire Monitoring (RSSI/energy-CC)

This study [31] integrates cooperative communication routing algorithms based on both QoS and energy consumption. QoS can be quantified through absolute Received Signal Strength Indicator (RSSI). The suggested study employs a multi-hop cooperative network. Cooperative nodes, referred to as CN, are established between sink and source nodes. The CN groups represent the modelled states of the environment.

The agents can either transmit the packets or observe their transmission. Each node in the proposed framework maintains a Q-value. This Q-value indicates the payout that each node must get by selecting either the forward packet action (af) or the monitor packet action (am). Data packets are communicated to node with highest reward. Two reward mechanisms are proposed, contingent upon the success or failure of a transmission. Node election and data forwarding protocols are predicated on optimal network quality and minimal energy consumption, or a compromise between the two. The ecosystem promptly rewards all nodes for each successful transfer, contingent upon link quality and energy usage.

4.11. An Adaptive Routing Tree Based on Learning for Wireless Sensor Networks

The study employs a real-time RL method to provide an adaptive spanning tree routing approach [32]. It utilizes meta-routing for constraint-driven approach. The strategy's operation has three phases: initiation, forwarding, and confirmation. The Q-value is defined at initialization. The Q-value indicates minimal cost from current to base node. When packets are transmitted from neighboring nodes and stored, NQ-values (neighbor Q-values) are communicated. These values are first estimated based on the attributes of neighboring entities. Consequently, at initialization, a initial spanning tree attached at sink is generated. All nodes, excluding for the sink, direct to the neighbor possessing the lowest NQ-value.

Packet forwarding from that node incorporates the node's present Q-value for the specific type of data. A neighboring node consistently recalibrates its Q-value and conveys the requisite NQ-value upon receiving data from a preceding node with a Q-value, irrespective of being the designated recipient. During the forwarding phase, a real-time search algorithm utilizing the Q-value is employed to transmit packets to the optimal neighbor.

The received data is consequently communicated to parent. If adjacent node with low NQ-value changes, the parent of the node is likely to be altered. During the confirmation phase, if forwarded node does not receive packet within designated timeframe, the parent indicator returns to neighbor possessing lowest NQ-value. The node's NQ-value is subsequently updated. This is referred to as implicit confirmation. Should the implicit confirmation fail at any node, a designated number of retransmissions will be initiated. Consequently, each of the protocol's three steps entails learning.

4.12. MRL-CC

It consists of a mesh structure that is built on multi hops using cooperative nodes. Data is transmitted through these cooperative nodes (CN) [33].

The node functioning as the agent in MRL-CC is selected according to the multiagent RL method employing Q-learning. Relative distance between two cooperative partner nodes determines Q-value. Q-value for itself and collaborating node is retained by every node

The preserved Q-value indicates the PDR and latency of existing routes to sink. The nodes involved in the cooperative communication group have two alternatives upon receiving a packet. Data is transmitted to node having high Q-value. Packet transmission is monitored by remaining cooperative nodes in CN group. Reward is based on successful transmission. Quality of packets transmitted signify positive reward whereas delay signifies negative reward. The positive reward specifies data packet's progression towards sink. The node having highest Q-value at convergence receives packet, whereas nodes with inferior Q-values oversee the transmission and provide assistance in the event of a failure.

4.13. In WSN, Distributed Adaptive Cooperative Routing that is Cognizant of QoS

The proposed work investigates QoS delivery for critical applications in latency and reliability domain using cooperative communication [34]. To optimize long-term benefits, the model-free RL technique advocates for the application of parametric optimization. The RL's states represent the routers' destinations along the path from the source. Transmission is being conducted either through

REVIEW ARTICLE

relayed or direct methods. When the neighboring node fails to satisfy the criteria for the current packet's reliability or latency, the agent remains inactive. The prize is dispensed solely upon the execution of the action.

4.14. An Effective Algorithm for Intelligent Energy Routing in Wireless Sensor Networks

The method [35] incorporates routing strategy and path discovery through the Q-learning approach as essential elements. The proposed approach consists of two operational phases. Initially, clustering is employed, followed by the utilization of a linked graph to construct the network. In the second step, the learning data is transmitted through the Q-value. Using cluster building in the initial phase the network is partitioned in numerous clusters of maximum size. Size of cluster is determined by position of sink. The dimensions of the adjacent cluster are considered last. In the initial phase nodes are not required to recognize CH node. Using cluster ID nodes of its one-hop neighbor transmit data to sink. As ideal CH nodes are taught, each node can either choose most efficient data transmission path or take the role of CH node. The present node status and its adjacent nodes dictate the selection of the appropriate CH node. Q-learning is utilized in the proposed FFIEE. This method addresses node failure problems and identifies the optimal CH node. Second stage selects optimal CH node, that incurs the minimal cost to sink. Each sensor node functions as a state and an independent learning agent within the Q-learning framework. The alternatives selected include either designating itself as CH for the subsequent hop or transmitting data to a neighboring node. It transmits a packet to the base assuming that the selected next hop is its own node after retaining it for a specified duration.

Two criteria are employed to assess the agent's actions in order to ascertain the reward. The distance between node to sink influences rewards, hence reducing overhead. The remaining energy of node constitutes secondary parameter for reward. This facilitates reduction of latency across the network.

The delays and the distance metric constitute the basis for FTIEE's cost functions. The Q-value is determined by the node's and action's Q-values. The distance between nodes and sinks is utilized for its calculation. The Q-value of action incorporates battery level of each node. Consequently, the approach acquires the capability to discern the optimal CH nodes and routes throughout network cycles. The proposed FTIEE protocol facilitates simultaneous data transfer and route discovery on demand. During data transfer with priority, all nodes direct the data to the node possessing the superior Q-value. Information is transmitted arbitrarily to any node, including those situated within another cluster, provided that Q-values of the nodes are comparable. The ε-greedy method is employed in proposed strategy to select the maximum Q-value. The protocol employs a data-driven approach that utilizes learning and data aggregation to conserve energy.

4.15. MRL-SCSO

The technique [36] proposes a unique system for unattended WSNs. The sensor of the UWSN function under demanding environmental conditions. The proposed method attains self-optimization and self-configuration in WSNs by employing an energy-aware convex hull algorithm alongside multi-agent RL. MRL-SCSO comprises both convex and non-convex nodes that operate as agents. Non-convex nodes intermittently alter states, whereas convex nodes function as the foundation. They may exist in an active, dormant, inactive, or exploratory state.

During the discovery phase, the convex nodes are operational and capable of processing, forwarding, receiving, and discarding packets or control messages. Energy conservation is generally executed during the sleep state. Each node, as a state value, monitors the energy, traffic, and link quality of all available paths to the sink for itself and its neighbors. Depending on environmental changes, the node in the discovery state may opt to route packets to proximate nodes or modify its status. The objective is to optimize rewards by exploration and exploitation. The payout is influenced by link quality, node energy, and node buffer occupancy. Three distinct categories of links are monitored for the assessment of link quality. Active link monitoring is conducted by transmitting probe packets to neighboring nodes at a preset rate. Passive link monitoring occurs when a node intercepts a packet and acknowledges a message not directed to it. Hyperlink monitoring integrates both active and passive link surveillance. Link quality is assessed using many parameters, including the quantity of packets transmitted or received, acknowledgments, timestamps, RSSI, and sequence numbers. The quality of the link is utilized to assess a positive or negative reward.

The predetermined energy threshold of a node is selected to equilibrate the network's energy consumption. To maintain coverage, nodes above the specified threshold are retained in a low-power condition until either all neighboring nodes adopt the same state or no nodes are left. The buffer occupancy of each node is crucial for determining whether to activate more neighboring nodes or to advance through convex nodes. The policy specifies the maximum state value to be conveyed to the sink. The three steps of MRL-SCSO implementation are boundary establishment, topology management and control, and data distribution. The convex and non-convex nodes are differentiated during the initial operational phase. All nodes stay active inside the network, facilitating packet forwarding, processing, receiving, and dropping; hence, the convex nodes assist in the transfer of emergency data.

REVIEW ARTICLE

The route table retains data about the convex nodes to ensure connectivity with sink during second operational stage. The convex nodes in the state remain operational until the energy falls below a specified threshold. Each non-convex node in the network utilizes residual energy and connection quality, which are contingent upon the reward, to identify two active neighbors. The neighbor list is contingent upon the prize. Active nodes are marked as higher-reward. In the MRL-based topology, the nodes are instructed to operate in either an active or sleep mode. The value function ascertains the quality of other nodes. Periodically, sleep state nodes shift to the inactive state. The active node activates the passive state nodes when the count of neighboring nodes falls below the threshold.

During the data dissemination phase, the neighbors receiving the highest rewards exchange information. The rewards of the nodes and their neighboring nodes are updated based on the actions performed. The active node selects the neighbor node that offers the highest incentive for data transfer. The network data transmission backbone consists of active convex nodes. To achieve self-configuration within the network, the multi-agent RL policy, informed by network's conditions, transforms the states of the nodes. The selection of nodes is crucial for local optimization. Global optimization is achieved through the utilization of a multi-agent RL approach that facilitates interchange of observed data across neighboring nodes throughout each contact. Consequently, the multiagent RL mechanism in the proposed method provides self-configuration and self-optimization to enhance PDR, average throughput, end-to-end delay and energy efficiency.

4.16. An Intelligent Reinforcement Learning-Based Routing Strategy for Wireless Sensor Networks

The suggested approach employs reinforcement learning-based routing to improve lifetime of network and energy efficiency [37]. The nodes signify the state in reinforcement learning, and the neighboring node is selected built on the assessment of path quality from source to sink node utilizing neighboring nodes along the way. Data packets and control signals are transmitted over the network in the proposed RLBR.

During startup, the sink acquires control packets. The node ID, geographic locations, residual energy, and hop count of control packets provide data regarding preceding forwarder node. The data packets encompass the following information: package ID, node ID, position, coordinates, remaining energy, hop count, Q-value, next forwarder, and payload data. The Q-value signifies present estimation of path quality from sink to preceding forwarder.

Each node disseminates control packets to neighboring nodes to establish the network in RLBR. To ascertain hop count to sink, sender node appends data prior to transmitting the control packet to its neighbor. A node ascertains the sender's Q-value by extracting the sender's data from the control packet upon receipt of the control packet. Q-value located in neighbor database of receiving node. The Q-value is determined by quality of path from neighbor to sink and reward received when present node transmits a packet to its neighbor. The compensation in RLBR is contingent upon the hop count of the neighboring node and its residual energy. The current node receives the highest reward for forwarding packets when the distance to its neighboring node is minimized. This enhances packet delivery and assists the network in sustaining an energy equilibrium. Following the exchange of information through control packets, data packets are communicated over network. Upon receipt of a packet and extraction of the despatcher node information, neighbor table is updated. If existing node is not designated node, packet is discarded. Upon receipt of packet, a sensor node apprises its neighbor table and retrieves sender's information. If node specified in the data packet's information table is not the active one, the packet is discarded. Upon examining the neighbor database for a record of the sink node to facilitate ongoing broadcasting, the present node transmits the packet directly to the sink. If no record of sink is available, data packet is forwarded to subsequent node using neighbor table. Subsequent node is chosen as neighbor node with highest Q-value. The current node concurrently changes its packet header information, hop counts, and Q-value. The input from this data transfer cycle informs the next one. The feedback mechanism enables the network to utilize less energy.

4.17. Routing with Energy Efficiency Using RL

The publication [38] proposes a cluster-based, reinforcement learning-driven energy-efficient routing system named EER-RL. To improve subsequent hop choice and energy depletion, the suggested method shares local data with neighboring nodes, including device ID, hop count, and remaining energy position coordinates. The three steps of EER-RL's operation include cluster formation, data communication, and network arrangement, together with cluster head election. During the selection of CH and establishment of network, base station disseminates its positional coordinates and transmits a heartbeat message. Upon receipt of the heartbeat packet, each node records the base station's position, while the initial hop count and energy are utilized to compute preceding Q-value. The heartbeat message initiates the CH election procedure. Upon the completion of the CH election, a notification is dispatched to all nodes. It includes CH ID, its primary Q-value, and its position. In the subsequent phase of data transfer, RL is employed for acquisition. During this phase, each node or device functions as a learning agent. In EER-RL, the neighbor is chosen as the next hop for routing the packet from each state. The Q-values are revised continuously during the learning process based on the immediate reward. The computation employs the following

REVIEW ARTICLE

parameters: hop count, reward, hop count for computing hop count, distance among sender and each neighbor, and nodes remaining energy. The incentive, condensed in packet header, diminishes network payloads. Monitoring the packets enables each adjacent node to refresh its routing database. The sender node in EER-RL selects the neighbor node with the highest Q-value to optimize the reward.

4.18. A Novel Approach to Utilizing RL and Fuzzy Logic to Identify a Highly Dependable Path in IOT

The suggested work introduces an innovative energy-efficient method to enhance network longevity [39]. Using Fuzzy logic in collaboration with RL enhance network longevity by considering the proximity to sink, the available bandwidth, and remaining energy of nodes to determine best pathway. The nodes symbolize the states, and the action is selecting the most sustainable information transfer pathway. The evaluation criterion ascertains the reward for each action by evaluating the node's available bandwidth, residual energy, and proximity to the sink. Suggested method for optimizing network longevity use fuzzy rules to evaluate reward for identifying the ideal path from source node to sink node. The reward consists of three input factors: the node's closeness to the sink, its available bandwidth, and its residual energy. Two fuzzy sets are established, yielding eight fuzzy bases for each input factor. Consequently, the proposed methodology acquires optimal path by utilizing state-action rewards via RL, while fuzzy rules assess the path's quality based on the calculation of the reward function.

4.19. RL-Based Cooperative Model (CRLM)

The authors provide a cooperative model for energy balance utilizing meta-heuristics and RL [40]. The study leverages the merit-seeking capability of the metaheuristic algorithm built upon RL. It addresses load balancing issue in network clusters by implementation of pruning and a multi-hop approach. It demonstrates an extended network lifespan.

4.20. Energy Balanced Routing Based on RL with Greedy Action Chains (RLR-TET)

Greedy action chain is employed by RLR-TET approaches to achieve node energy equilibrium and determine the optimal routing path. Greedy chains are employed in the technique to construct a dynamic tree structure throughout the learning process [41]. The optimal next hop is identified at the base of the tree, whereas error updates are acquired in proximity to the tree's nodes. The benefit of the tree structure is in its capacity to adapt to dynamic changes while maintaining network efficiency and energy equilibrium.

4.21. Intelligent Routing System with Distributed Neural Network Guidance

Liu et al. [42] suggested a methodology for large-scale WSNs. Network's failure to adapt to evolving conditions stems from its expanding topology. The optimization of timely routing decisions is impeded by the proliferation of mistakes. The proposed DNN-HR technique facilitates the optimization of the RL process through three sequential steps: A neural network governs the learning nodes by adjusting V values, while a hierarchical learning framework facilitates the nodes' acquisition of knowledge from their neighbors. Identifying cluster heads by weight incremental propagation trees and establishing a hierarchical subspace that divides the network based on geographic regions. The cluster heads obtain the weight increments acquired by the nodes. Subsequently, all sensor nodes obtain a broadcast containing the updated neural network weights. The three-step technique demonstrates enhanced packet delivery and node survival rates.

4.22. Enhanced Tree Routing Using the RL Algorithm

The performance characteristics of WSNs, including dependability, latency, and energy economy, must be balanced according to the specific application. Consequently, identifying the best parent node in tree-structured routing methodologies is difficult. Kim et al. utilize this constraint in the proposed investigation [43]. The researchers utilize Q-learning to gather empirical data from the network to identify the ideal parent node. Q-learning employs cognitive indicators such as hop count, short-term power consumption ratio, buffer occupancy, and the weighted average of received signal strength to determine rewards.

4.23. Q-Learning-Based Routing for Information Transmission that Uses Less Energy

Su et al. proposed a method to enhance energy efficiency through the adaptive selection of neighboring nodes in a routing system [44]. The incentives are revised based on data from the neighboring node, including transmission direction and distance, energy consumption, and residual energy. The suggested method employs a combination of rewards to compute the updated Q-values. The five incentives are information transmission distance, data transmission action, direct difference in data transmission reward, data transmission energy consumption, and residual energy. Subsequent hop will be node having maximum Q-value.

4.24. Energy-Efficient, Reinforcement-Learning-Based, Optimal Routing Protocol for Wireless Sensor Networks (RLER)

To address the problem of energy depletion in hotspots or mobile sinks, the authors propose solution in [45]. The authors employ a clustering method that utilizes a grid tree. The RLFIS algorithm is employed alongside Neighbourhood Overlapping (NOVER), Bipartivity Index (BI), and Algebraic Connectivity (AC) to ascertain root node. The nodes' energy and transmission range are utilized to construct a tree structure within the grid, originating at the root node. The

REVIEW ARTICLE

Hybrid Bat-Crow approach is employed to shift sink nodes. The approach has an extended network longevity.

4.25. Reinforcement Learning with Multiple Agents for Energy-Efficient Routing in Wireless Sensor Networks

The authors employed the reward function to identify three specific reward functions aimed at improving reliability and minimizing transmission latency [46]. The current condition of the network is utilized to ascertain the optimal paths.

4.26. EEERP-RL: An Improved Energy-Efficient Wireless Sensor Network Routing Protocol Based on Reinforcement Learning

This article examines the effects of both active and inactive nodes, as well as energy usage [47].

In [48], the authors propose energy-efficient Q-learning routing by imposing penalties on rewards based on the energy levels of adjacent nodes.

Table 1 consolidates the state-action pairs, rewards, and advantages of the aforementioned RL strategies. Furthermore, Table 1 illustrates the QoS provisioning considered by several routing algorithms built on RL for WSNs.

Table 1 Summary of Routing Methods Built on RL in WSNs

Method	Type	QoS Provisioning	State Action Pair	Reward	Analysis	Advantage
AdaR [17]	Model free	Increase throughput and reduce delay	hop count between two neighbors nodes to sink, the residual energy, the number of routing paths crossing neighbor node	Reward is computed depending on the quality of route path and includes a positive reward if sink is reached, zero when path length exceeds and sink is not reached and negative when a cycle appears in a path	Built upon Least Squares Policy Iteration (LSPI) which learns the best routing approach with a lesser number of trials and is not sensitive to primary constraint	As LSPI learns the weights by sampling from the environment, it becomes helpful to apprise the Q-values created using updated information from the state action pair Limitation: QoS mapping with Q-values not mentioned
[19]	Model free	Delivery ratio, average delay	State- Nodes obtains a data packet Action- select neighbor node as successive hop built on learned Q	Q-value updated based on reward. Reward consists of several scenarios. Reward function indicates low energy warning signal	focused on the reward function in RL to achieve a lifetime of network lifetime and improve PDR.	Reward takes care of void problem in GPRS
[22]	Bayesian decision model	Energy	local interaction among neighbor nodes	Delayed reward of previous actions	opportunistic-based routing policy algorithm, incorporates	able to make intelligent route choices from the reward of earlier actions for

REVIEW ARTICLE

					routing by adapting to network dynamics	successful packet delivery
FROMS [23]	Model multiple sinks as RL problem solved using Q-learning	Energy and packet delivery	where the packets need to reach and the hops to the sink using all the neighboring nodes to the individual sinks			Used for multicast routing,
RL based QoS aware routing [18]	RL based QoS aware	Average packet delivery, average delay	State: node in the network Action: Forward packet to the next node	Reward is built on the actions that are forward data	Optimal routes are obtained using distributed Q-learning to provide QoS aware routes	While forwarding data packet the QoS requirement is checked Limitation: no data exchanges is considered in between itself and other sensor nodes hence global optimization cannot be achieved in large scale networks
MRL-QPR [25]	Multiagent RL with QoS support	Average end to end delay	State: Every node in network Action: Execution of the forwarding data packets by the agent	When data is forwarded, positive reward is obtained	Local information and exchange with direct neighbors is incorporated	Each node evaluates its own projected QoS performance and expected QoS performance of the next node and impact of transmitting. This cooperation among nodes help select actions to exploit the global rewards
QoS-RSCC	Multi agent RL	Throughput		Immediate reward is given	Use cooperative communication	

REVIEW ARTICLE

[26]				for optimal relay selection	with adaptive relay selection	
EQR-RL [29]		Energy	State-node Action-forwarding of data packets to next neighbor node		Each neighbor in the route gives feedback of the time to send the packet forward. It updates Q table, to obtain expected minimum future cost of each node, through which packet can be sent to sink	Handle mobility and failure issues even though the feedback received from the neighbor nodes is negative. Select neighbor node with best QoS
QSGrd [30]	Combination of gradient and Q-learning, model free based RL	Energy	State: nodes	based on the progress made towards sink, energy improvement, and decrease residual transmissions for packet to reach sink	encrypts the info of shortest path using transmission gradient and Q-Learning enhance routes to balance energy consumption and reduce packet delay	
QoS and Energy Aware Cooperative Routing [31]	Cooperative communication routing	Delay packets, average Delay and energy	Cooperative nodes between sink and source are modelled as environment states Actions: forwarding the packets and monitoring the forwarding packets	Reward function depends on the success or the failure of transmission		
A learning-based adaptive routing tree [32]	Learning tree based adaptive routing	PDR			It works in 3 phases; initialization, forwarding and confirmation, Q-value is defined in	

REVIEW ARTICLE

					initialization, forwarding phase packets are forwarded once, in confirmation phase whether packet is received is confirmed	
[33]	uses multi hop mesh cooperative approach using group cooperative nodes	Average delay, PDR	State: cooperative nodes Action: Packet forwarding and packet transmission monitoring	Packet quality reflects positive reward forwarded; negative reward is a function of delay caused by unsuccessful transmissions.	Cooperative communication-based RL where nodes with high Q-values forward packets; nodes with low Q-values help packet transmission failures	
[34]	Model free learning	reliability	States: State of routers along the route Action: Execution of direct or relayed transmission	Reward is obtained for every action executed	--	QoS aware routing using cooperative communication
Efficient intelligent energy routing protocol [35]		Packet Delay, Packet delivery, energy	State: Each sensor node Actions selected is data transmission to a neighbor	Reward is obtained based on two parameters for the action taken by the agent and depends on distance among node and the BS/sink and remaining energy of the node	Implemented in two steps, the initial step involves clustering to establish a network using connected graph. In second step information is broadcasted using Q-value	on-demand detection of route and data communication
MRL-SCSO [36]	multi-agent RL	PDR, energy	States: node can have these states: discovery, active, idle	Reward depends on link quality, node energy and buffer occupancy of	Convex nodes act as the backbone of the network. The multi-agent RL policy of	Achieves global optimization

REVIEW ARTICLE

			and sleep Action: forwarding packets to neighboring nodes and state change depends on environmental conditions	the node and is based on exploration and exploitation	converting the states of the node based on the condition of the network helps it to achieve the self-configuration in the network.	
Intelligent routing algorithm [37]		Energy efficiency, packet delivery	State: node Action: to select the neighbor node based on assessment of path quality from source to sink and neighbor nodes	depend on reward to send data to neighbor and quality of the path to sink node	Two types of packets are flooded in to the network; Control and data packets	reward in RLBR is proportionate to the remaining energy and hop count to sink. This approach aids in preserving an energy balance and improves packet delivery
[38]		Energy	State: nodes Actions: selection of the neighbor	Rewards depend on hop count, link distance to calculate the hop count, distance among sender and each neighbor and the remaining node energy	local information is shared with the neighbor nodes to enhance energy	.
[39]	Combination of fuzzy logic and RL	End to end delay	State: nodes Action: to select high sustainable route for sending data information	Reward is the combination of distance amongst node, remaining energy and available bandwidth of the node.	Optimize network lifetime, fuzzy rules are used to rate the reward to search the ideal path	
CRLM [40]	Implements combination of RL and meta-heuristic algorithm		Selection of cluster heads	RL is used to enhance the features of the meta-heuristic algorithm	Performs load balancing and energy balance to improve network lifetime	Optime network lifetime and load balance in clusters

REVIEW ARTICLE

RLR-TET [41]	Greedy chains with tree structure	--	Root of tree represents the next hop	and leaf node represent the errors for updating in the learning process	Build upon greed action chains	Optimize network lifetime
DNN-HR [42]	Three step in which neural network are used to update weights to sensor networks	--	Neural networks are used optimize the learning in RL	Sensor nodes learn by themselves, form cluster heads based on the geography and weight updating is provided using neural networks as well for adapting in changing environments of large scale WSN	Forms two level learning structure along with geographical subspace formation for weight increment propagation and hierarchical weight updating	Improves node survival rate, Enhanced packet delivery
Enhanced Tree routing based on RL algorithm [43]	Tree structure based on RL, Optimize the search for optimal parent node	--	Reward is based on hop count, buffer occupancy, received signal strength and short term power consumption ratio	The network dynamics are obtained using empirical models through Q-learning	Exploits the challenge of finding ideal node	Enhance end-to end-delay, PDR and energy
Q-learning based energy efficient information transmission [44]	Based on Q-learning	--	State: current node and Action: Next optimal hop	Reward constitutes of summation of 5 types of rewards	Achieves energy efficiency in minimum number of hops	Balance energy Consumption and increase network lifetime
RLER [45]	Hybrid approach uses RL to form tree grid structure	--	Form root node using RLFIS algorithm	Hybrid BAT-Crow search algorithm	Helps to avoid mobile sinks and hotspots to overcome the problem of energy consumption	Enhance network lifetime
Multi-agent-based RL			Current status of node used	Three trustworthy reward		Improve reliability and decrease

REVIEW ARTICLE

[46]				functions defined		transmission delay

5. SIMULATION AND EXPERIMENTAL ANALYSIS FOR ROUTING UISNG RL IN WSN

An analysis of the simulation and experimental results for several routing algorithms built on RL for WSN are summarised. This section also addresses how the efficacy of RL-based routing and QoS provisioning is profoundly influenced by the choice of the state action and reward function. Table 2 presents an overview of the experimental investigation on RL based routing strategies for WSNs.

5.1. Q-Routing

Experiments indicate that Q-routing provides better efficacy than the shortest path routing approach when compared to average delivery time [14]. The fundamental Q-routing technique can thus enhance two essential QoS attributes: throughput and latency. The Q-value is contingent upon the network cost and latency. The connection cost and the latency of the subsequent forwarder node diminish with the link quality.

5.2. Geographic Routing Protocol Based on Reinforcement Learning for UWB Wireless Sensor Network

The algorithm was executed in NS2 across multiple scenarios to assess its efficacy. Average latency, network longevity, and delivery ratio exemplify performance indicators. The RLGR exhibits a delivery ratio comparable to that of GPSR [19]. The variance in node energy is responsible for the observed enhancement in RLGR's average latency. The reward function evaluates energy of node, prompting the agent to select next hop with better remaining energy to establish an extended route to base station. It has been noted that despite a significant quantity of packets being transmitted to the sink before to the network partition, many are discarded due to a limited number of nodes having insufficient energy at the conclusion. The suggested RLGR algorithm identifies surrounding nodes and adjusts to energy variations to prolong network longevity. A thorough reward function was devised that accounts for node energy, delay, routing failure, and network longevity.

5.3. Q-Probabilistic Routing

It is evaluated against GPSR (Greedy Perimeter Stateless Routing) [49]. a thoroughly examined position-aware routing system [50] and Expected progress-Face-Expected progress [51]. The absence of neighboring nodes that are nearer to the target than the node itself is a significant challenge for routing algorithms employing greedy forwarding [22]. To overcome this limitation, QPR employs candidates ranked by their probability of successful delivery. The efficacy of Q-PR is assessed using the Received Importance Ratio (RIR) and the anticipated transmissions (ETX). RIR is ratio of overall significance produced by network traffic to the overall significance that was received. Experimental testing unequivocally demonstrates that QPR surpasses GPRS regarding successful delivery rate. Nevertheless, QPR's delivery rate is equivalent to that of EFE. The QPR algorithm enhances network longevity, indicating that nodes can conserve energy for an extended period compared to GPRS and EFE. Q-PR utilizes the Bayesian probability model for forwarding, so conserving energy for the transmission of more critical information.

5.4. RL-QRP RL Based QoS Aware Routing Protocol

During the initial stages of the simulation, as the sensor nodes acquire knowledge of all potential routes and assess the quality of their connections, it underperforms compared to QoS-AODV. Upon completion of learning, its performance surpasses that of QoS AODV. In dynamic network conditions, RL-QRP demonstrates superior performance compared to QoS AODV. Under conditions of minimal network traffic, RL-QRP and QoS-AODV exhibit equivalent performance. The suggested work's limitation is that each sensor node disregards data exchanges with other sensor nodes. This approach cannot attain global optimization, particularly in extensive networks.

5.5. MRL-QPR

The OMNeT++ simulation platform underpins Castalia [52][53], utilized for simulation purposes. The initial performance is superior with QoS-AODV. Subsequent to the preliminary learning phase, MRL-QPR operates more effectively as the algorithm considers all potential possibilities, resulting in optimal node behavior based on the state information. MRL-QRP surpasses QoS-AODV regarding average end-to-end delay under high traffic conditions. QoS-AODV and MRL-QRP exhibit equivalent performance under conditions of minimal traffic movement. MRL-QRP surpasses QoS-AODV in high mobility scenarios because to the adaptability provided by machine learning in response to environmental changes.

5.6. QoS-RSCC

This method employs cooperative communication when direct transmission is unsuccessful, with CRP selecting two relays for each packet transfer. QoE-RSCC exhibits superior performance in CPR owing to diminished channel utilization, resulting from an elevated likelihood of packet collisions and channel access conflicts. The second advantage is that QoS-RSCC utilizes network resources efficiently, in contrast to CPR, which requires two relays. Relay selection is more efficient when comparing QoS-RSCC to CRP, as only the relay that enhances performance regarding outage probability

REVIEW ARTICLE

and channel efficiency is chosen. This also facilitates superior network throughput compared to CPR as data transmits.

5.7. EQR-RL

Reinforcement Learning-based energy-aware Quality of Service routing: EQR-RL functions in conjunction with RL-QRP and QoS-AODV. QoS-AODV outperforms EQR-RL and RL-QRP. Following a specified duration of simulation, EQR-RL outperforms RLQRP as the algorithm acquires insights into network dynamics via load balancing and congestion mitigation measures.

The EQR-RL protocol allocates packets among multiple routes while taking into account the network's specifications, hence enhancing performance in high-traffic situations. The EQR-RL method has a longer lifespan than QoS-AODV and RL-QRP due to the implementation of load balancing in the subsequent hop selection.

5.8. QSGrd

The simulation framework is implemented using OMNeT++. The approach is evaluated against the Energy Aware Gradient (EAGrd), Energy Smart Gradient (ESGrd), and Greedy Gradient (GGrd) algorithms. QSGrd outperforms EAGrd and enhances the network's longevity, as evidenced by simulations conducted to validate node energy. To minimize packet delay, QSGrd evaluates routes characterized by lower Packet Error Rates (PER) and superior connection quality by assessing the likelihood of successful transmissions.

5.9. QoS and Energy Aware Cooperative Routing Protocol for Wildfire Monitoring Wireless Sensor Networks (RSSI/energy-CC)

It utilizes the TOSSIM simulation environment. The performance of is compared to MRL-CC. In comparison to noncooperative algorithms, RSSI/energy CC has superior performance regarding packet loss. The RSSI/energy CC technique demonstrates superior performance when comparing the average latency to sink and the percentage of delayed packets.

The optimal RSSI for packet transmission is utilized to determine the forwarding node. It surpasses MRL-CC algorithm in energy usage and network endurance.

5.10. A Learning-Based Adaptive Routing Tree for Wireless Sensor Networks

The probabilistic sensor network simulator PROWLER, version 5.11 is employed for simulation [54]. Performance is evaluated using node failures and mobile sinks. It is contrasted with backbone tree and grid routing schemes. When it encounters failures of nodes, the adaptive tree protocol demonstrates an enhanced delivery ratio and superior performance. The adaptive tree methodology demonstrates elevated delivery rates in mobile sink settings.

5.11. MRL-CC

The Castalia WSN simulator is employed to assess performance through MRL-CC simulations [52][53][54]. It demonstrates superior performance in average end-to-end delay when simulations are executed in error-prone wireless networks. The MMCC node selection prioritizes distance, while MRL-CC algorithms analyze the environment and relay data to the subsequent node based on insights derived from incentives, experiences, and nodes exhibiting optimal connection quality. MRL-CC surpasses MMCC in PDR because to its superior flexibility and adaptability for data packet forwarding under varying packet arrival rates. MRL-CC excels in heavy network traffic due to its utilization of knowledge and rewards to select the optimal cooperative policy in highly variable conditions. MRL-CC is also enhancing QoS metrics for extensively scaled networks and dynamic scenarios.

5.12. QoS Aware Distributed Adaptive Cooperative Routing

NS2 simulator is employed for simulation. It is compared with Minimum Power Cooperative Routing (MPCR) [55], Reliable and Energy Efficient Routing (REER) [56] and MRL-CC systems. The proposed DACR architecture exhibits worse energy efficiency in comparison to REER and MRL-CC regarding the overall energy consumption of network nodes. However, due to DACR's elevated PDR, it consumes less energy per packet. DACR's implementation of energy-aware route identification and relay selection markedly enhances its performance regarding reliability, energy consumption, and latency compared to the analyzed protocols. DACR enhances latency and reliability at each hop through a proactive adaptive decision-making process for transmission mode selection.

5.13. Efficient Intelligent Energy Routing Protocol in Wireless Sensor Networks

It uses NS2 in conjunction with a C# tool specifically developed for WSN. The subsequent protocols employed for the comparison of the proposed work are: Low Energy Adaptive Clustering Hierarchy (LEACH) [57], Hybrid Energy Efficient Distributed Clustering utilizing Fuzzy Logic and Non-Probabilistic Strategy (HEED-NPF) [58], and Energy Efficient Clustering Scheme in WSNs (EECS) [59].

Network's longevity, packet delivery, and latency are utilized to evaluate FTIEE's performance. FTIEE demonstrates that data collection and an innovative, intelligent use of Q-learning have enhanced the network's longevity. The proposed approach exhibits comparable performance in packet delivery optimization. The proposed method demonstrates enhanced energy efficiency relative to other methods. The algorithm demonstrates robust fault tolerance and reliability. Rather than resolving or circumventing problems, FTIEE retransmits packets during a network disruption. Despite the FTIEE

REVIEW ARTICLE

retransmission policy being perceived as a burden on the system, it surpasses current protocols in performance.

5.14. MRL-SCSO: Multi-Agent Reinforcement Learning-Based Self-Configuration and Self-Optimization Protocol for Unattended Wireless Sensor Networks

It is simulated using the Cooja Simulator and compared with the collection tree protocol (CTP) to assess its performance regarding throughput, energy, average end-to-end delay, and PDR. MRL-SCSO exhibits superior performance under elevated PDR load as it utilizes only the requisite number of active nodes and backbone paths.

MRL-SCSO has superior average end-to-end delay performance for substantial loads, as packets are relayed through convex nodes that are consistently updated with reliable, active neighbors. MRL-SCSO attains superior throughput compared to CTP under high traffic scenarios due to its utilization of active nodes. Message transmissions during network configuration result in increased energy consumption by MRL-SCSO. Energy consumption diminishes as the quantity of nodes increases, attributable to adaptive state scheduling.

5.15. An Intelligent Routing Algorithm in Wireless Sensor Networks Based on Reinforcement Learning

It employs NS2 platform simulations to evaluate lifetime of network and energy. The subsequent algorithms are contrasted with it: BEER, EAR, MRL-SCSO, and the foundational Q-routing protocol. RLBR surpasses alternative techniques in energy efficacy, PDR, and lifetime of network by employing a feedback mechanism and continual learning to identify the most optimum broadcasting path for packets.

5.16. Routing Energy Efficiently Based on RL

MATLAB is employed to assess the EER-RL methodology. EER-RL exhibits a prolonged network lifespan relative to LEACH, PEGASIS, and Flat EER-EL. The cluster-based approach is more efficacious in extensive networks for enhancing energy efficiency and prolonging network lifespan.

5.17. A Novel Approach to Utilizing RL and Fuzzy Logic to Identify a Highly Dependable Route in the Internet of Things

The OPNET 11.5 emulator is utilized for testing, and a comparison is made with the IEEE 802.15.4 protocol. This demonstrates that the proposed RL and fuzzy methodology surpasses IEEE 802.15.4 regarding end-to-end delay. The improved outcomes are attributable to the transmission of all data by high-bandwidth nodes, guided by fuzzy logic assessments for route path discovery. A path with the highest energy and shortest distance to sink is chosen for high bandwidth node. The IEEE 802.15.4 protocol will exhibit increased end-to-end latency due to the selected node potentially possessing substantial bandwidth yet insufficient energy to transmit packets to the sink node.

The proposed RL and fuzzy technique surpass the IEEE 802.15.4 protocol regarding packet delivery and minimizes data packet error rates. The nodes chosen for the next hop towards the sink achieve superior packet delivery because to their elevated energy levels, optimal bandwidth, and proximity to the sink.

An elevated signal-to-noise ratio is an additional characteristic of the proposed data that will assist in delivering superior quality data to the consumer.

Table 2 Experimental Analysis of RL Based Routing Methods for WSNs

Method	Simulation platform	Performance Assessment/ Metrics	Algorithms Compared	Strengths and Weakness of performance comparison
Q-routing [14]	--	Throughput and Delay	Shortest path routing algorithm	--
RLGR [19]	NS-2	Network Lifetime	Greedy Perimeter Stateless Routing (GPSR)	Exhibits better performance than GPRS
Q-PR [20]	--	Delivery rate	GPSR and Expected progress-Face-Expected progress (EPFE)	Delivery rate of QPR is similar to EFE and better than GPRS

REVIEW ARTICLE

RL-QRP [18]		Average end to end delay	QoS-AODV	Demonstrates weak performance in initial phase and exhibits enhances performance for dynamic network traffics
MRL-QPR [25]	Castalia [44] built on the OMNeT++ [45]	Average end to end delay	QoS-AODV	Demonstrates weak performance in initial phase and exhibits enhances performance for dynamic network traffics
QoS-RSCC [26]	--	Network Throughput	Cooperative Routing protocol (CRP)	--
EQR-RL [29]	OMNeT++	Energy and network lifetime	GGrd, Energy Aware EAGrd), and Energy Smart Gradient (ESGrd)	QSGrd protocols significantly outperform in extending lifetime of network
[31]	TOSSIM simulation platform, discrete simulator for TinyOS	Delayed-packets and average-delay k	MRL-CC	--
Learning-based adaptive routing tree [32]	PROWLER [36],	Delivery ratio, Latency	Backbone tree and Grid routing	--
MRL-CC [33]	Castalia [17] built on OMNeT++ [18]	Average end-to-end delay	MMCC	MRL-CC demonstrate enhanced performance for many QoS metrics for large scale and vastly changing network environments
[34]	NS2	Average end-to-end delay, reliability, energy expenditure, protocol operation overhead, integrated performance, network lifetime	Minimum Power cooperative routing (MPCR), Reliable, Energy Efficient Routing (REER), MRL-CC	
Efficient intelligent	NS2 using C# tool	PDR and network lifetime	LEACH, HEED-NPF, EECS	FTIEE exhibit betters for network lifetime, acceptable

REVIEW ARTICLE

energy routing protocol [35]				performance in packet delivery, better energy efficiency, good reliability and fault tolerance
MRL-SCSO [36]	Cooja Simulator	PDR, network lifetime, delay and energy	collect tree protocol (CTP)	MRL-SCSO demonstrate enhanced performance for PDR, end to end delay, and throughput
Intelligent routing algorithm [37]	NS2	PDR, network lifetime, delay and energy	MRL-SCSO, EAR, BEER, and basic Q-routing protocol	--
[38]	MATLAB	Network lifetime	LEACH, PEGASIS and Flat EER-EL	--
[39]	OPNET 11.5	End-to-end delay, packet delivery	IEEE 802.15.4 protocol	--
CRLM [40]	MATLAB 21	Network Lifetime	E-ALWO, ChOA-HGS, GATERP, GWO, IPSO-GWO, and LEACH	--
RLR-TET [41]	--	Energy balance, network lifetime		--
DNN-HR [42]	MATLAB	PDR node survival and energy standard	--	--
[43]	OPNET	Average PDR, end-to-end delay, power consumption	Linear weighted sum-depending on parent selection approach	--
Q-learning based energy efficient information transmission [44]	MATLAB	Average and remaining deviation of energy, Energy consumption, Average number of hops, energy depletion in nodes	1. Energy aware self organizing routing 2. Decentralized RL based routing 3. Centralized RLBR approach	--

REVIEW ARTICLE

RLER [45]	OMNeT + +	Improved network lifetime	RLFIS algorithm, Hybrid Bat-Crow	Reduce energy consumption and hotspot by use of mobile sink
Energy Efficient Q-learning [48]	MATLAB	Enhanced network lifetime	Q-learning and RLBR	-

6. ROLE OF OPTIMIZATION AND ITS EFFECTS AND CHALLENGES IN RL BASED ROUTING

This section examines the implementation of several optimization strategies in RL algorithms to ascertain the optimal path. AdaR use Least Squares Policy Iteration (LSPI) to optimize routing. In LSPI, a parametric function approximator is employed to estimate the Q-values, $Q\pi$, for a specified policy π, instead of assessing the optimal value function [60][61].

Value-function is linearly weighted representation of properties associated with each state-action pair. The advantages of LSPI include accelerated convergence and a singular evaluation of the policy across the sample set. The initial parameters can be readily modified as it evaluates a history of samples. The policy is determined using linear grouping of features of state-action pair, facilitating equilibrium of optimization objectives for data transmission to neighboring nodes.

The parameters of the reward function are defined in [19] to ascertain the optimal network longevity. The reward function, utilizing exploration probability, discounted factor, and rate of learning, identifies the optimal path under diverse routing conditions.

Due to the uncertainty regarding the subsequent forwarder node, Q-Probabilistic Routing in WSNs use an average cost as the cost parameter. The RL algorithm acquires this average cost locally through data interchange among neighboring nodes. In Q-PR, the transmission energy and message significance are equilibrated by parameter γ. Regulating γ poses the primary challenge in enhancing Q-PR.

FROMS employs both stochastic and greedy exploration to provide feedback routing for optimizing multiple-sinks. Q-values are perpetually revised throughout the learning process based on the feedback incentive. The algorithm converges in a finite duration, and the Q-values remain unchanged. Subsequent to this phase, the greedy search is implemented.

The research delineates two methods for assessing convergence: a reward-based strategy that aggregates the total of neutral rewards obtained. Upon the receipt of N rewards, the exploration concludes. The second strategy is reward-based and route-based. Unlike the first, this necessitates the identification of at least M paths prior to the consideration of neutral incentives.

The RL-based QoS-aware Routing Protocol designates the current sensor node in RL-QRP as the entity capable of altering the environment's state. We do not consider currents or exchanges with the other sensor nodes. As a result, global optimization is not achieved, especially in large networks.

MRL-QPR employs a multi-agent cooperative RL approach to improve QoS in WSNs. QoS routes are determined by collecting local network data and sharing restricted information with adjacent neighbors. This facilitates the identification of optimal paths via global optimization.

The study used the ε-greedy exploration policy, characterized by a defined probability ε, to select the optimal relay in QoS-RSCC. This facilitates adaptation to the dynamic environment for optimal relay selection via rewards and experiences.

Exploration functions as the educational step in EQR-RL. Exploration use the ε-greedy strategy to reconcile low-cost approaches with alternative probable choices. The ε-greedy approach employs probability ε to ascertain the most probable course of action.

Numerous agents in MRL-CC explore the optimal cooperative policy through rewards and experiences. With restricted information dissemination and localized network data, many agents collaborate to exploit the optimal policy while considering their interactions with one other. Consequently, optimal outcomes can be realized without the need of storing exact network data and central control.

Optimal relay selection is accomplished by a multi-objective Lexicographic Optimization (LO) method in QoS aware distributed adaptive cooperative-routing [62]. The goal functions are organized in lexicographic order. The primary aim of the two objective functions is to optimize reliability, while the secondary aim is to reduce latency. The LO is augmented with a constraint that guarantees the principal objective function retains its optimal value.

In an efficient intelligent energy routing system, a node's behaviour in identifying the ideal path to the sink is dictated by the reward value. The actions are selected by the ε-greedy policy.

Two types of activities are taking place in MRL-SCSO: exploration and exploitation. Exploitation enhances network performance by selecting action with maximum V-value. For each state-action pair, a random action is chosen to improve

REVIEW ARTICLE

the approximation of V-value. The basis of data forwarding is inquiry. A node delivers packet to next nodes utilizing maximum reward and the probability of ε.

To enable the algorithm to compute optimal routes utilizing reinforcement learning, most proposed methodologies either adjust the Q function or develop a reward function.

7. CONCLUSIONS AND CHALLAENGES

This study offers a comprehensive summary of routing methods built on RL for WSNs. The study highlights the reward functions and the selection of state-action pairs for routing. In several RL-based routing systems, the objective is to identify the next optimal forwarder with a low Q-value along route to sink node, where nodes represent states. The reward function is essential when evaluating the optimal course of action. The literature has few methodologies proposing rewards, whether positive or negative, in relation to WSN applications. The proposed methods utilize exploration or exploitation for both short- and long-term routing goals. To ascertain optimal paths, several algorithms in the literature employed the greedy methodology of Q-learning. Investigations into the rate and verification of convergence remain nascent. QoS-enabled computationally intelligent routing approaches sometimes encounter communication overhead and computational burden [18]. Ensuring accurate node information in dynamic environments poses a challenge for the design of routing algorithms that enhance or support QoS. Several proposed strategies outline the creation of QoS support for routing through a RL approach. The issue of adaptive QoS support and provisioning for RL-based routing protocols remains unaddressed.

REFERENCES

[1] Meena Pundir and Jasminder Sandhu, A systematic Review of Quality of Services in WSNs using Machine Learning: Recent Trend and Future Vision, Journal of Network and Computer Applications, 188, (2021). https://doi.org/10.1016/j.jnca.2021.103084.

[2] Padmalaya Naya, G. K. Swetha, Surbhi Gupta, K. Madhavi, Routing in WSNs using machine learning techniques: Challenges and opportunities, Measurements, 178, (2021). https://doi.org/10.1016/j.measurement.2021.108974.

[3] Jingiing Yan, Mengchu Zhou, Zhijun Ding, Recent advances in energy-efficient routing protocols for WSN: A review-IEEE Access, pp-5673-5686, (2016). doi: 10.1109/ACCESS.2016.2598719.

[4] Praveen Kumar D, Tarachand Amgoth, Chandra Sekhara Rao Annavarapu, Machine learning algorithms for WSNs: A survey, Information Fusion (2018). https://doi.org/10.1016/j.inffus.2018.09.013

[5] Raja Basha, A. A Review on WSNs: Routing. Wireless Personal Communication 125, pp. 897–937 (2022). https://doi.org/10.1007/s11277-022-09583-4

[6] Zoubir Mammeri, RLBased Routing in Networks: Review and Classification of Approaches, IEEE Access, vol-7, pp. 55916-55950, (2019). doi: 10.1109/ACCESS.2019.2913776.

[7] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 2017.

[8] K. V. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[9] D. Witten, R. Tibshirani, and T. Hastie, An Introduction to Statistical Learning: With Applications in R. New York, NY, USA: Springer, 2013.

[10] S. Ayoubi et al., Machine learning for cognitive network management, IEEE Communication. Mag., vol. 56, no. 1, pp. 158–165, (2018). 10.1109/MCOM.2018.1700560.

[11] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, Computational intelligence in WSNs: A survey, IEEE Communication. Surveys Tuts., vol. 13, no. 1, pp. 68–96, 1st Quart., (2011), https://doi.org/10.1007/978-3-319-47715-2.

[12] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, Machine learning for networking: Workflow, advances and opportunities, IEEE Network., vol. 32, no. 2, pp. 92–99, Mar./Apr. (2018), 10.1109/MNET.2017.1700200.

[13] Gustavo Könzel, Leandro Soares Indrusiak, Carlos Eduardo Pereira, Latency and Lifetime Enhancement in IWSN: a Q-Learning Approach for Graph Routing, IEEE Transaction Vol. 16, pp. 5617- 5625, (2019), doi: 10.1109/TII.2019.2941771.

[14] J. A. Boyan and M. L. Littman, Packet routing in dynamically changing networks: A RLapproach, in Proc. Neural Inf. Process. Syst. Conf. (NIPS), pp. 671–678, (1994), 10.5555/2987189.2987274.

[15] Yau, Kok-Lim & Goh, Hock Guan &Chieng, David &Kwong, Kae. Application of RLto WSNs: models and algorithms. Computing. Vol. 97, (2015), 10.1007/s00607-014-0438-1.

[16] Mohamed Said Frikha, Sonia MettaliGammar, Abdelkader Lahmadi, Laurent Andrey, Reinforcement and deep RLfor wireless Internet of Things: A survey, Computer Communications, vol. 178, pp. 98-113, (2021), https://doi.org/10.1016/j.comcom.2021.07.014.

[17] P. Wang and T. Wang, Adaptive Routing for Sensor Networks using Reinforcement Learning,The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), pp. 219-219, (2006),10.1109/CIT.2006.34.

[18] Xuedong Liang, I. Balasingham and Sang-Seon Byun, A RLbased routing protocol with QoS support for biomedical sensor networks,First International Symposium on Applied Sciences on Biomedical and Communication Technologies, pp. 1-5, (2008), 10.1109/ISABEL.2008.4712578.

[19] S. Dong, P. Agrawal and K. Sivalingam, RLBased Geographic Routing Protocol for UWB WSN, IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference, pp. 652-656, (2007), 10.1109/GLOCOM.2007.127.

[20] B. Karp and H.T. Kung. GPSR: Greedy perimeter stateless routing for wireless networks, In Proc. ACM/IEEE Mobi. Com., pp. 243- 254, (2000), https://doi.org/10.1145/345910.345953.

[21] I. Oppermann, L. Stoica, A. Rabbachin, Z. Shelby, J. Haapola, "UWB WSNs: UWEN - A Practical Example," IEEE Communications Magazine, Vol. 42, Issue 12, 2004 pp. S27 - S32.

[22] R. Arroyo-Valles, R. Alaiz-Rodriguez, A. Guerrero-Curieses and J. Cid-Sueiro, "Q-Probabilistic Routing in WSNs,3rd International Conference on Intelligent Sensors, Sensor Networks and Information, pp. 1-6, (2007), 10.1109/ISSNIP.2007.4496810.

[23] A. Forster and A. L. Murphy, FROMS: Feedback Routing for Optimizing Multiple Sinks in WSN with Reinforcement Learning, 3rd International Conference on Intelligent Sensors, Sensor Networks and Information,pp. 371-376, (2007),10.1109/ISSNIP.2007.4496872.

[24] N. Ouferhat and A. Mellouk, Optimal QoS and adaptative routing in WSNs, in Proc. IEEE International Conference on Information and Communication Technologies (ICTTA'06), pp. 2736–2741, (2006), 10.1109/ICTTA.2006.1684844.

[25] Xuedong Liang, I. Balasingham and Sang-Seon Byun, A multi-agent RLbased routing protocol for WSNs,IEEE International Symposium on Wireless Communication Systems,pp. 552-557, (2008), 10.1109/ISWCS.2008.4726117.

[26] X. Liang, I. Balasingham and V. C. M. Leung, Cooperative Communications with Relay Selection for QoS Provisioning in WSNs,GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference,pp. 1-8, (2009), 10.1109/GLOCOM.2009.5425437.

REVIEW ARTICLE

[27] A. Nosratinia, T. Hunter, and A. Hedayat, Cooperative communication in wireless networks, IEEE Communications Magazine, vol. 42, no. 10, pp. 74–80, (2004), 10.1109/MCOM.2004.1341264.

[28] Y.-W. Hong, W.-J. Huang, F.-H. Chiu, and C.-C. J. Kuo, Cooperative communications in resource-constrained wireless networks,IEEE Signal Processing Magazine, vol. 42, pp. 47–57, (2007), 10.1109/MSP.2007.361601.

[29] S. Z. Jafarzadeh and M. H. Y. Moghaddam, "Design of energy-aware QoS routing protocol in WSNs using reinforcement learning," 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-5, (2014), 10.1109/CCECE.2014.6900988.

[30] B. Debowski, P. Spachos and S. Areibi, Q-learning enhanced gradient-based routing for balancing energy consumption in WSNs,2016 IEEE 21st International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), pp. 18-23, (2016),10.1109/CAMAD.2016.7790324.

[31] Mohamed Maalej, Sofiane Cherif, and Hichem Besbes, QoS and Energy Aware Cooperative Routing Protocol for Wildfire Monitoring WSNs, Hindawi Publishing Corporation, The Scientific World Journal Volume 2013, https://doi.org/10.1155/2013/437926.

[32] Ying Zhang and Qingfeng Huang, A Learning-based Adaptive Routing Tree for WSNs, Journal of Communications, vol. 1, issue-2 (2006), 10.1109/ICAEE.2015.7506827.

[33] X. Liang, M. Chen, Y. Xiao, I. Balasingham, and V. C.M. Leung, MRL-CC: a novel cooperative communication protocol for QoS provisioning in WSNs, International Journal of Sensor Networks, vol. 8, no. 2, pp. 98–108, (2010), https://doi.org/10.1504/IJSNet.2010.034619.

[34] Md. Abdur Razzaque, Mohammad Helal Uddin Ahmed, Choong Seon Hong, Sungwon Lee, QoS-aware distributed adaptive cooperative routing in WSNs, Ad Hoc Networks, vol.19, pp. 28-42, (2014), https://doi.org/10.1016/j.adhoc.2014.02.002.

[35] Kiani F, Amiri E, Zamani M, et al. Efficient intelligent energy routing protocol in WSNs. Int J Distrib Sens N vol.5, issue 27, (2015), https://doi.org/10.1155/2015/618072.

[36] Renold AP and Chandrakala S. MRL-SCSO: multi-agent reinforcement learning-based self-configuration and self-optimization protocol for unattended WSNs. Wireless Personal Communication, vol. 96, pp.5061–5079, (2017), https://doi.org/10.1007/s11277-016-3729-3.

[37] Guo WJ, Yan CR, Gan YL, et al. An intelligent routing algorithm in WSNs based on reinforcement learning. Appl Mech Mater, vol. 678, pp. 487–493, (2014), https://doi.org/10.4028/www.scientific.net/AMM.678.487.

[38] Vially Kazadi Mutombo, Seungyeon Lee, Jusuk Lee, and Jiman Hong, EER-RL: Energy-Efficient Routing Based on Reinforcement Learning, Mobile Information SystemsVolume 2021, https://doi.org/10.1155/2021/5589145.

[39] Akbari, Y., Tabatabaei, S. A New Method to Find a High Reliable Route in IoT by Using RLand Fuzzy Logic. Wireless PersCommun vol. 112, pp. 967–983 (2020), https://doi.org/10.1007/s11277-020-07086-8.

[40] Zhendong Wang, Liwei Shao, Shuxin Yang, Junling Wang, Dahai Li, CRLM: A cooperative model based on RLand metaheuristic algorithms of routing protocols in WSNs, Computer Networks, Volume 236 (2023)

[41] Liu, Z., Wang, X. Energy-balanced routing in WSNs with RLusing greedy action chains. Soft Comput (2023). https://doi.org/10.1007/s00500-023-08734-4

[42] Zhibin Liu, Yuhan Liu, Xinshui Wang, Intelligent routing algorithm for WSNs dynamically guided by distributed neural networks, Computer Communications, Volume 207, pp. 100-112, (2023), https://doi.org/10.1016/j.comcom.2023.05.018.

[43] Kim BS, Suh B, Seo IJ, Lee HB, Gong JS, Kim KI. An Enhanced Tree Routing Based on RLin WSNs. Sensors (Basel). 2022 Dec 26;23(1):223. doi: 10.3390/s23010223.

[44] Su, X., Y. Ren, Z. Cai, Y. Liang, and L. Guo. 2023. "A Q-Learning-Based Routing Approach for Energy Efficient Information Transmission in WSN." IEEE Transactions on Network and Service Management 20 (2): 1949–1961.

[45] Keerthika, A., and V. Berlin Hency. 2022. "Reinforcement-Learning Based Energy Efficient Optimized Routing Protocol for WSN." Peer-to-Peer Networking and Applications 15: 1685–1704. https://doi.org/10.1007/s12083-022-01315-6.

[46] Prabhu, D., R. Alageswaran, and S. Miruna Joe Amali. 2023. "Multiple Agent Based RLfor Energy Efficient Routing in WSN." Wireless Networks 29: 1787–1797. https://doi.org/10.1007/s11276-022-03198-0.

[47] Jaber, M. J., and Z. J. Jaber. 2024. "EEERP-RL: Enhanced Energy-Efficient Routing Protocol Based on RLfor WSN." Internet Technology Letters. https://doi.org/10.1002/itl2.385.

[48] Archana Chaudhari, Vivek Deshpande and Divya Midhunchakkaravarthy. "Energy-efficient Q-learning-based routing in WSNs". International Journal on Smart Sensing and Intelligent Systems Sciendo, 18, no. 1 (2025): https://doi.org/10.2478/ijssis-2025-0008

[49] P Bose, P. Morin, I. Stojmenovic, J. Urrutia, Routing with guaranteed delivery in ad hoc wireless networks, In Proc. of 3rd ACM Int. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications DIAL M99, pp. 48-55, (2001), https://doi.org/10.1023/A:1012319418150.

[50] B. Karp, H.T. Kung, Gpsr: greedy perimeter stateless routing for wireless networks, In MobiCom : Procs. of the 6th Annual Int. Conf. on Mobile Computing and Networking, ACM Press, pp. 243–254, (2000), https://doi.org/10.1145/345910.345953.

[51] M. Stojmenovic, A. Nayak, Localized routing with guaranteed delivery and a realistic physical layer in WSNs, In Computer Communications, vol. 29, pp. 2550-2555, (2006).

[52] H. N. Pham, D. Pediaditakis, and A. Boulis, From simulation to real deployments in WSN and back, in Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'07), pp. 1–6, (2007).

[53] A. Varga, The OMNeT++ discrete event simulation system, in Proc. The 15th European Simulation Multiconference (ESM'01), pp. 112–118, (2001).

[54] G. Simon, Probabilistic wireless network simulator. http://www.isis.vanderbilt.edu/projects/nest/prowler/.

[55] A.H. Ibrahim, Z. Han, K.J. Ray Liu, Distributed energy-efficient cooperative routing in wireless networks, IEEE Trans. Wireless Commun., Vol. 7, issue-10, pp. 3930–3941, (2008), 10.1109/T-WC.2008.070502.

[56] M. Chen, T. Kwon, S. Mao, Y. Yuan, V. Leung, Reliable and energy efficient routing protocol in dense WSNs, Int. J. Sensor Networks, vol. 4, issue- 12, pp. 104–117, (2008).

[57] Heinzelman, W., A. Chandrakasan, and H. Balakrishnan. 2000. "Energy-Efficient Communication Protocol for Wireless Micro Sensor Networks." Proceedings of the 33rd Annual Hawaii International Conference on System Sciences 8.

[58] Taheri, H., P. Neamatollahi, M. Naghibzadeh, and M.-H. Yaghmaee. 2010. "Improving on HEED Protocol of WSNs Using Non-Probabilistic Approach and Fuzzy Logic (HEEDNPF)." Proceedings of the 5th International Symposium on Telecommunications (IST '10), 193–198. https://doi.org/10.1109/ISTEL.2010.5734023.

[59] Ye, M., C. Li, G. Chen, and J. Wu. 2005. "EECS: An Energy Efficient Clustering Scheme in WSNs." Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference (IPCCC '05), 535–540.

[60] Boyan, J. A. 1999. "Least-Squares Temporal Difference Learning." In Proceedings of the 16th International Conference on Machine Learning (ICML).

[61] Lagoudakis, M. G., and R. Parr. 2001. "Model-Free Least-Squares Policy Iteration.", Advances in Neural Information Processing Systems, 14.

REVIEW ARTICLE

[62] Zykina, V. 2004. "A Lexicographic Optimization Algorithm." Automation and Remote Control 65: 363–368. https://doi.org/10.1023/B:AURC.0000019366.84601.8e.

Authors

Archana Chaudhari: Post Doctoral Fellow at Lincoln University College, Petaling Jaya, Malaysia. Doctor of Philosophy (Engineering), Assistant Professor, of the Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune, India. Areas of scientific interests: WSNs, internet of Things, Machine Learning, Deep Learning, Control System.

Vivek Deshpande: Post Doctor Fellow (Engineering), Director and Professor, Vishwakarma Institute of Information Technology, Pune, India. Areas of scientific interests: WSNs, Distributed Systems, High Performance Computer Networks Real Time Embedded Systems, Internet of things.

Divya Midhunchakkaravarthy: Director, Centre of Postgraduate Studies, Lincoln University College, Malaysia. Areas of scientific interests: Cybersecurity, network and systems, Data Science.

How to cite this article:

Archana Chaudhari, Vivek Deshpande, Divya Midhunchakkaravarthy, “A Comprehensive Review of Reinforcement Learning Based Routing Algorithms for Wireless Sensor Networks”, International Journal of Computer Networks and Applications (IJCNA), 12(3), PP: 449-474, 2025, DOI: 10.22247/ijcna/2025/28.