KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

DEVELOPMENT OF AN ENHANCED MULTI-HOP PROTOCOL FOR OPTIMIZATION OF NETWORK EFFICIENCY IN WIRELESS SENSOR NETWORK USING REINFORCEMENT LEARNING

[1]OKAFOR Anthony Chinedu (Ph.D)[*]
[2]DAUDA, Umar Suleiman (Ph.D)
[3]OHIZE, Henry O. (Ph.D)
[4]KOLO, Jonathan G. (Ph.D)
[5]AJIBOYE, Johnson Adegbenga (Ph.D)

[1,2,3,4,5] Department of Electrical and Electronics Engineering
Federal University of technology Minna (FUTMINNA), Niger state, Nigeria
[*] Corresponding Author, email address: anthony.pg2113605@st.futminna.edu.ng

ABSTRACT

The most prominent recent developments in distributed computing have utilized Wireless Sensor Networks (WSNs) for real-time data acquisition and monitoring of environmental, industrial, health, and urban environments. Despite such innovative applications, their continued use is hampered by limitations such as energy efficiency, maximization of network lifetime, and reliable multi-hop communication. A large part of existing research on WSN routing protocols is heuristic based and static, which, in turn, facilitates the wasteful use of energy in the network, quick depletion of the operating nodes of the network, and overall loss of network efficiency. This research proposes the development of a multi-hop routing protocol and network reinforcement learning WSN routing protocol. The new protocol uses Q-learning and energy-aware routing metrics to make decisions at each sensor node to create an optimal balance of energy, latency, and packet delivery ratio. The protocol was tested on the NS-3 network simulator in direct comparison to LEACH, PEGASIS, and AODV routing protocols, and demonstrated 35% network lifetime extension, 42% reduction in end-to-end delay, and 28% improvement in packet delivery ratio. Contributions of this study include the construction of a new state-space model for WSN routing, a distributed reinforcement learning framework that adapts to the resource limitations of the sensor nodes, and the development of an adaptive reward function that adjusts to the state of the network. The results confirm the validity of using AI techniques combined with standard networking protocols to tackle the complex problems of WSN deployments.

Keywords: *Wireless Sensor Networks, Multi-hop Routing, Reinforcement Learning, Q-learning, Energy Efficiency, Network Optimization, Protocol Design, Adaptive Routing*

1. INTRODUCTION

The growing use of wireless sensor networks showcases a fundamental change in distributed computing systems and the acquisition of data, revealing new possibilities in environmental and structural monitoring, as well as pervasive computing [1]. Sensor networks are systems designed to be formed by an arbitrary number of self-sufficient sensors that work together to monitor varying environmental conditions, including temperature, humidity, pressure, vibration, sound, and chemical pollutants [3]. Some of the defining characteristics of a wireless sensor network (WSNs) involve energy resources that are limited, coupled with restricted data processing

resources, the dynamic and unpredictable nature of the network's topology, and the unreliable nature of communication pathways available. These features, among many, pose a multitude of problems in the protocol structuring and network optimization of WSNs [2].

Designing WSNs incorporates many aspects, although energy use remains by far the most important. Once deployed in a remote location, sensor nodes operate on batteries that cannot be recharged. Therefore, postponing any energy conservation and control initiatives will be detrimental to the network and will hinder the optimization efforts. Energy conservation and control initiatives lead to the optimization of the network's limited capabilities. As wireless communication surpasses the range of associated computational tasks, the majority of energy used in WSNs is a byproduct of communication tasks [4]. This fact demands that WSN communication tasks be limited in number. As a result, the ideal solution to the posed problem is to create routing protocols that retain the network's data delivery and overall communication reliability in order to nourish the network's lifetime and support its overall functioning [5].

The advantages of multi-hop communication protocols in wireless sensor networks (WSNs) include extending coverage, optimizing energy usage via load balancing, and increasing network scalability. In multi-hop communication, a data packet is transmitted by one or more intermediate nodes on the path from source to sink, thus facilitating communication between nodes that may not be in direct radio range of each other [6]. However, typical multiple-hop routing protocols make either fully static, or partially static routing choices, resulting in a lack of adaptability to dynamic, and possibly unpredictable, network situations such as node failures, energy depletion, channel quality changes, and alterations in the direction and volume of traffic. This limitation has prompted researchers to investigate the addition of reconfigurable mechanism for optimizing routing as a function of the observed state of the network and associated performance metrics [7,8].

Reinforcement learning is a type of machine learning that is concerned with how agents learn to interact with different environments to maximize cumulative rewards. It is a promising technique to construct adaptive routing protocols. It is distinguished from supervised learning (which requires labeled training examples), as reinforcement learning allows agents to learn optimal strategies through interactions with the environment. Because of this, reinforcement learning is especially appropriate for use in wireless sensor networks. The environmental conditions and states of the network change dynamically and are difficult to model ahead of time (i.e., a priori), so learning through interaction helps to construct a policy that can dynamically adjust to the goal of the network, and to a lesser extent the environmental conditions [2, 10]. The routing problem can be modeled as a Markov Decision Process, and then with the use of Q-Learning, the sensor nodes will be able to learn to support intelligent decision making in the routing process. In this way the nodes will be able to optimize their objectives for remaining “alive” in the network (i.e., energy), the time it takes for a packet to make it to the destination (i.e., delay), as well as the reliability with which the packet is delivered to the destination [11].

The study focused on the development of an improved multi-hop protocol containing a reinforcement learning-based adaptive mechanism to optimize the multi-hop protocol to provide improved network efficiency. This approach facilitates the potential for individually adaptive routing mechanisms at the network nodes utilizing local information and local feedback [12]. The incorporation of energy-aware metrics into the reinforcement learning framework guarantees that the routing decisions are based on residual energy, rate of energy consumption, energy-aware metrics, and the network longevity goals [13,9]. Additionally, the protocol's framework

dynamically adapts to topology changes, node dropout, and varying network traffic, robustness across multiple varying deployment conditions [14].

2. LITERATURE REVIEW

2.1.1 2.1 Traditional Routing Protocols in Wireless Sensor Networks

Routing protocols for wireless sensor networks have evolved over the years. This evolution attempts to solve challenges still present in today's resource-constrained distributed systems [16]. From the earlier protocols like Minimum Transmission Energy and Direct Transmission, nodes would either send their data directly to the base station, or choose the next closest neighboring node to send the data to. These protocols caused significant energy imbalances across the nodes in the sensor network, as nodes nearer to the base station would run out of energy first due to the large deferring responsibilities [3].

The Low Energy Adaptive Clustering Hierarchy (LEACH) protocol was the first of its kind to implement hierarchical clustering on a sensor node network. This helped distribute energy depletion evenly across the entire network. LEACH involves an iterative round approach consisting of a setup phase and a steady-state phase where data tends to be sent [4]. Although LEACH was an improvement in flat routing approaches to sensor networks, it suffered from the same shortcomings like inefficient selection of cluster heads, failure to account for the remaining energy in a node during cluster head rotation, and a rapid decline in performance as the networks increased in size [20]. Random selection of cluster heads has proven to be inadequate, as nodes with low remaining energy (i.e., exhausted nodes) have been selected which leads to a collapse of the entire sensor node network.

In PEGASIS, each node is part of a chain structure, whereby each node will, at some point, transmit individual sensor readings to the base station. Each node collects data from its predecessor and, after data fusion, passes the packet to the next node. Compared to LEACH, PEGASIS achieves better uniform energy dissipation and eliminates cluster formation overhead. However, PEGASIS's performance is negatively impacted by dynamic networks because of the reconstruction overhead, the delays associated with successive transmissions [19, 38].

In terms of a reactive routing strategy, AODV is the first to establish a route only when it is needed, therefore, control overhead is minimized, especially in networks with intermittent communication streams. AODV incorporates route discovery and maintenance protocols based on the principles of distance vector routing. However, with regards to the reactive approach, a significant amount of time is wasted in establishing a route. Furthermore, in a mobile and dynamic topology, a significant amount of control overhead is generated with the demand routing approach is inadequate. Additionally, AODV is energy unaware, resulting in the quality of the route determined by the energy of the node in a critical position that determines the remaining lifetime of the network [18, 39].

2.1.2

2.1.3

2.1.4 2.2 Machine Learning Approaches in Wireless Sensor Networks

Applying new machine learning methods to develop new adaptable protocols designed to operate intelligently and autonomously under different network conditions has become a popular area of interest regarding research in wireless sensor networks (WSNs) [19]. Furthermore, WSNs have been tested using supervised learning for different functions such as determining the location of a node, detecting events, and combining data. Unfortunately, these methods are not easily generalizable to a variety of deployment situations and/or different environmental conditions, as they are data intensive. Furthermore, the need for centralized training and the large number of

computations are the leading causes to WSNs being data deficient in terms of supervision learning [21].

One of the biggest advantages of using reinforcement learning in WSN protocols is the mechanisms that are present to encourage the development of the best operating policies based on interactions with the environment, as there is no need for training data that are pre labeled. Some studies have examined the use of Q-learning, a type of reinforcement learning, to solve the routing problem in wireless networks. In these algorithms, the routing choice is viewed as a multi-step problem, where each node determines which of the available next hops will best advance a defined target. To this end, the Q-learning algorithm uses a Q-table to keep track of the level of each cumulative reward that has been received as a result of taking a determined action in a defined state. Furthermore, updating each of these reward values is determined the state of the system and the level of the reward received as a result of the action taken [14].

Reinforcement learning-based routing protocols have been shown to be more effective than classical protocols in adapting to traffic flow, topology changes, and node failures [22]. However, there are further challenges when incorporating reinforcement learning to wireless sensor networks. The multiple parameters within a wireless sensor network create continuous state spaces which require effective discretization, or function approximation. The distributed structure of wireless sensor networks requires coordination schemes among learning agents on disparate nodes [24]. The trade-off between exploration and exploitation must be adequately offset so that sufficient learning occurs while netwok performance degradation during the learning phase is minimized. Further, the relatively scant resources of sensor nodes must be weighed against the memory and processing power demands of reinforcement learning algorithms [13].

2.3 Research Gaps and Motivation

The limitations of existing routing protocols for wireless sensor networks have motivated the need for new approaches despite the extensive body of literature on the topic. The limitations of the classical protocols is that their routing decisions are static or semi-static which leads to poor performance in the face of dynamic networks. Poorly designed and highly costly revisions to clusters lead to inefficient cluster head selection in clustering protocols like LEACH. In chain protocols like PEGASIS, there are excessive delays and the systems do not scale effectively. In reactive protocols like AODV, control overhead in dynamic topologies is excessive and energy consumption is not optimized [4,13,9].

Most previous reinforcement-learning-based routing protocols have ignore WSN's multidimensional performance and concentrated instead on single-objective optimization, either on delay minimization or on lifetime maximization. The most prevalent techniques for building state space representations have, at best, only limited ability to describe the most important dynamic properties of a network: the quality of the links, the state of the buffers, and the spatial distribution of energy. Additionally, previous works have failed to adequately adapt dynamic weights to changing network priorities and network operational phases for the reward functions used, resulting in static and poorly performing reward functions [24].

Due to the limited resources of the sensor nodes, the large scale of the deployment is a likely cause for the continuing issues for the scalability of the reinforcement learning-based protocols, especially in regard to the extensive computational complexity and memory requirements of the learning algorithms. Existing approaches to distributed reinforcement learning for WSNs tend to lack effective global optimization in coordination: once a set of local decisions is made at the nodes, the global performance of the network, in most cases, degrades. Minimization of convergence time for the learning algorithms is essential to enable the network to operate

satisfactorily during the initial deployment and as the network topology changes post deployment [25].

This study has developed state-of-the-art protocols, applying intelligent integration of specific networking knowledge with machine learning models, to improve multi-hop routing protocols incorporating reinforcement learning, adaptive reward functions, energy-aware metrics, and responsive to network conditions, and devising distributed coordination mechanisms to enhance global performance of networks, as well as other clear network performance advances.

3. SYSTEM MODEL AND PROBLEM FORMULATION

3.1 Network Model

Each of the nodes in the wireless sensor network is modeled as a graph **G = (V, E)**, where V is the set of sensor nodes in the network, and E is the set of links that communicate with each of the nodes. The network consists of N stationary sensor nodes that are dispersed randomly in a two-dimensional sensing field, and one or more stationary base stations that are data sinks. Each sensor node is equipped with a radio transceiver for wireless communication, limited on-board data processing, a finite energy supply from battery, and sensors to detect physical phenomena. The network nodes are equipped with a method for sensing their locations, either with GPS or an algorithm based on localization, to facilitate distance-based routing decisions [24].

Each node's transmission power level dictates their individual transmission ranges. A node can communicate directly with another node provided the distance between them is within their transmission range. The quality of the link can change due to the wireless channel being impacted by path loss, shadowing, and multipath fading. The network uses a collision avoidance MAC protocol to manage access to the channel so that packet collisions can be reduced. Depending on the needs of the application, each sensor node creates data packets that must be delivered to the base station within specified delay intervals [7].

3.2 Energy Model

The energy consumption model identifies and considers dominant sources of energy expenditure in wireless sensor nodes. These include radio transmissions, radio receptions, sensing, and processing. The energy consumed in transmitting a k-bit packet over a distance d is modeled as $E_{tx}(k,d) = E_{elec} \times k + \varepsilon_{amp} \times k \times d^{\alpha}$. E_{elec} is the energy dissipated per bit to run the transmitter or receiver circuitry. ε_{amp} is the energy consumed by the transmit amplifier, and α is the path loss exponent, which varies from 2 to 4 depending on the propagation environment. The energy consumed in receiving a k-bit packet is given by $E_{rx}(k) = E_{elec} \times k$ [30].

Each sensor node starts with an initial energy budget denoted as $E_{initial}$, and the available energy $E_{residual}(t)$ at time t decreases with every transmission, reception, and sensing operation. A node is said to be dead if its residual energy goes below the operational threshold $E_{threshold}$, which is necessary for rudimentary functioning. The network lifetime is defined as the time until the first node dies (first node death metric) or until a defined percentage of nodes are still alive (network coverage metric). The energy balance and energy efficiency metrics include the total energy spent for a packet delivered successfully, the difference of $E_{residual}$ across all nodes (indicating energy balance), and the energy depletion rate [35].

2.1.5 3.3 Problem Formulation

The routing optimization issue in wireless sensor networks can be described as an attempt to analyze a policy **\(\pi\)** that optimally balances a set of performance criteria including network lifetime, end-to-end delay, packet delivery ratio (**PDR)**, and energy efficiency. The specific aim is to maximize the operational time of the network, denoted as **\(T_{lifetime}\)**, while ensuring that

the network meets the specified operational thresholds, including a given minimum packet delivery ratio **(PDR)**, an end-to-end delay that is within an acceptable maximum **(D_max)**, and the network, at a minimum, meets the specified requirements for k-connectivity. Thus, the problem can be described as maximizing **\(T_{lifetime}\)** where **PDR** is greater than (or equal to) **PDR_min**, the expected value of delay **(E[D])** is less than (or equal to) **D_max**, and the network maintains k-connectivity.

When making routing decisions, each node must choose one of its neighboring nodes to which it will forward a packet. The goals of a node can be either minimizing some associated cost or maximizing some associated value, which is typically a combination of a number of disparate variables. The difficulty in this problem increases demographically as the variables under consideration in the problem tend to be volatile or time-dependent. For example, the quality of a communication link, the amount of residual energy within the nodes, the amount of data in the buffers, and the energy levels of the nodes and the available bandwidth can all change depending on the data traffic and communication protocols. The classical optimization techniques do not apply to this problem because they typically require knowledge of the problem before it is solved (in this case, traffic was known in some pattern); the network environment and its conditions present multiple dimensions that are constantly changing in a non-linear manner, requiring the problem to be solved in a fully decentralized manner); and the problem cannot be solved in a decentralized manner.

Formulating the problem as a **Markov Decision Process (MDP)** allows us to use techniques from reinforcement learning. The state space includes the most pertinent attributes of the network, detailing the remaining energy, buffers, link quality, hop counts to the destination, and neighbor info. The action space contains the next-hop decisions (actions) from the current set of neighbors available. The reward function captures the optimization goals and is formulated to reward (positively) successful packet deliveries and to minimize the energy and time (delays) consumed by the network, and to punish (negatively) packet drops, excessive delays, or actions that deplete the energy of critical nodes. The transition dynamics result from the intersection of the network protocols, the traffic patterns, and the external factors (the environment) that are unknown in the Beginning and need to be learned through experience..

4. PROPOSED ENHANCED MULTI-HOP PROTOCOL

2.1.6 4.1 Protocol Architecture

The proposed Enhanced Multi-hop Routing Protocol with Reinforcement Learning (EMRP-RL) uses a distributed architecture where each sensor node has its own Q-learning agent to make next-hop selection decisions [31]. The protocol consists of coordinated phases such as neighbor discovery and maintenance, collection and exchange of state information, Q-value initialization and learning, route selection and packet forwarding, and policy updates. The learning process in a distributed manner allows for large network scalability and enables each node to improve routing decisions based on local network conditions [31].

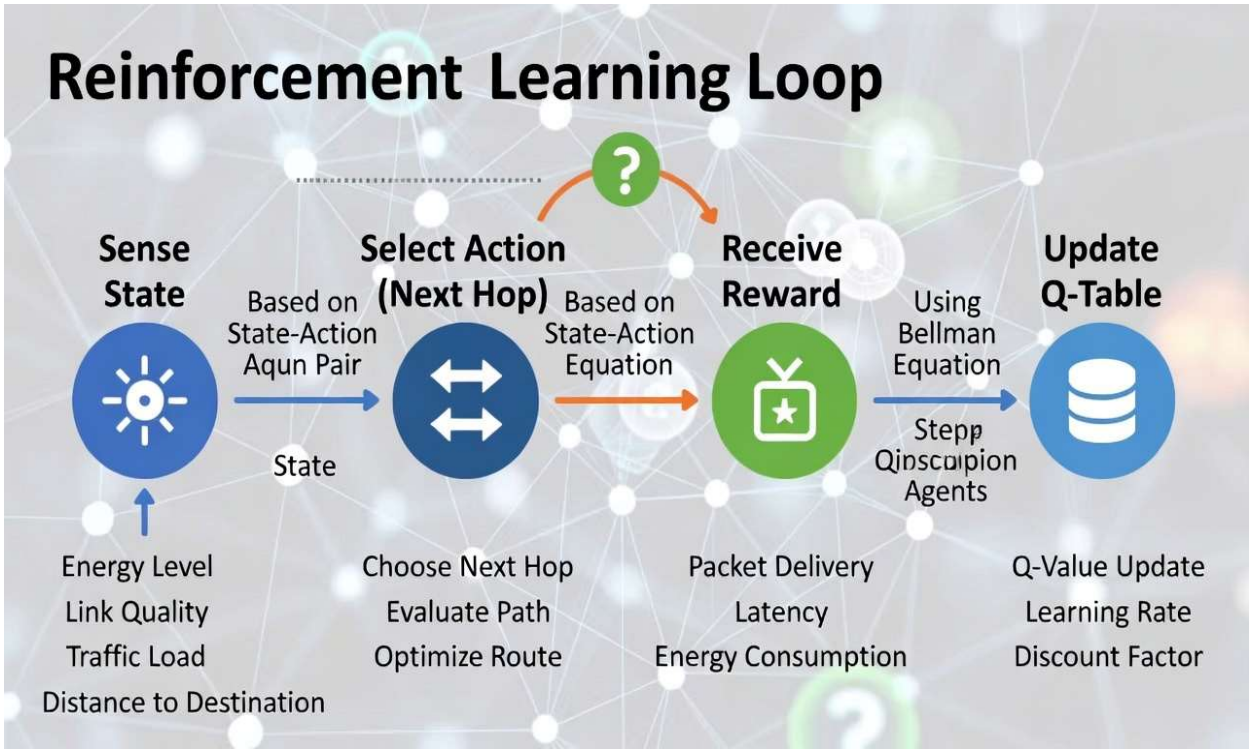


Fig. 1: Flowchart of EMRP-RL

In the neighbor discovery process, nodes send beacon messages, which serve to locate nearby nodes and establish two-way communication channels. These messages contain the sending node's ID, the amount of energy they have remaining, the node's location, and the number of packets that they have received. Each neighbor responds with a message of the same type. As a result, each node creates a Neighbor Table, which they continue to build out as time goes on. Periodically, tables are updated with status beacon messages to determine the current status of the node in the case that energy levels change or a node fails, or to see if a new node has entered the network [28]. The information state collection phase consists of determining the necessary parameters to be able to utilize reinforcement learning. Each node analyzes the operation of their individual module, examining their remaining energy, the amount of data in their buffer, the rate of data packets generated, and the number of packets that were successfully transmitted. Furthermore, nodes can send state information to their neighbor nodes by adding it to a data packet or an update message. The system employs efficient representation of the state space which ensures that the parameters are optimal enough to operate on the Q-learning and not enough to exceed the parameters, and that learning is effective [29].

2.1.7 Fig. 2: Architecture of EMRP-RL Nodes

2.1.8

2.1.9

2.1.10 4.2 State Space Design

In reinforcement learning, the state space representation is an integral part of the learning structure, as it specifies what data the learning agent can use to make decisions, which is the crux of the entire process. As such, the protocol under consideration describes the state space formulation as a multi-dimensional vector of local node features and neighbors. The state vector is defined as the combination of the following features: normalized residual energy of the current node (discretized into five levels: critical, low, medium, high, full), and a buffer occupancy percentage (discretized into four levels: empty, low, medium, high) the number of hops to the base station (discretized into ranges so as to limit state space size), and the aggregate neighbor quality metric which combines residual energy and link quality of the neighbors [28].

The state representation for each potential next-hop neighbor includes the normalized residual energy level of the neighbor, estimated link quality derived from the received signal strength indicator (RSSI) and the packet reception ratio, the neighbor's hop count to the base station, and the neighbor's buffer occupancy, if available. The link quality metric is a combination of both short- and long-term measurements/statistics, where the long-term measurements/statistics capture the channel characteristics, and the short-term measurements/statistics respond to temporary changes in the channel. Adaptive binning is used to discretize continuous state variables, and critical areas where the energy levels are low are given routing decisions that are more impactful on the network's lifetime [28].

The design of the state space achieves a balance between comprehensiveness and tractability by including the most pertinent parameters of the network and applying discretization to control the number of states. The size of the state space ultimately remains manageable for deployment on

resource-constrained sensor nodes, and the state space provides enough granularity to learn effective routing policies. The modular representation of states simplifies further extensions to include other metrics, such as opportunities for data aggregation, multi-path traffic, application-specific needs, and so on, without fundamentally changing the learning framework.

2.1.11 4.3 Action Space and Q-Learning Algorithm

Each node's action space is defined by its select one neighbor from its neighbor table as the next-hop forwarder for its outgoing packets. The actions available to the nodes change dynamically as neighbors appear and disappear because of mobility, node failures, and energy depletion. When a node is ready to forward a packet, it analyzes the Q-values for all its next-hop candidate nodes and chooses the action with the highest value based on an exploration-exploitation strategy. The action selection mechanism is based on ε-greedy exploration, meaning there is a probability of selecting a neighbor at random for exploration, and (1-ε) is the probability that the neighbor with the highest Q-value is selected for exploitation [18].

The Q-learning algorithm modifies the Q values based on the reward received at each step and the changes in states based on the following formula: **Q(s,a) ← Q(s,a) + α[r + γ max_a' Q(s',a') - Q(s,a)]**, with each variable defined as follows: s = current state, a = action taken, r = reward received, s' = state resulting from action a, α = learning rate, and γ = discount factor. The variable α starts at a high value in order to promote learning and is subsequently decreased to allow then to converge to a steady **Q-value. \par** The revision of the exploration rate, ε, is very similar. Starting with high exploration, the agent is permitted to modify its decision algorithm to engage in experimentation in order to find optimized pathways. The agent then transitions to a state where the rate of exploitation exceeds that of exploration as the Q-values approach convergence. To allow a greater emphasis on the long term objectives of the network lifetime, the discount factor is set close to the value of 1. The Q-values that relate to each action and state combination are organized in a table, which is then stored in the node's memory. To mitigate memory exhaustion and agonize it s use in the most efficient manner, the protocol makes *state aggregation* for the several exceptional states that are encountered very infrequently, and in the prioritized storage of Q-values for the states most states and action pairs.

2.1.12 4.4 Adaptive Reward Function

The reward function is one of the main determinants of the learning behavior and optimized goals of the routing protocol and thus it is of utmost significance. In this regard, the proposed protocol utilizes an adaptive reward function that changes according to the different phases and situations of the network operation. The reward function is made up of several elements each of which is aligned to a different performance goal. These components are as follows: the energy efficiency reward component **(R_energy)** which is the positive reward assigned for choosing neighbors with high residual energy and the negative reward assigned to the action that results in forwarding packets to low-energy nodes, the delay reward component **(R_delay)** which is the positive reward for rapid packet delivery and the negative reward assigned to a routing action that increases end-to-end delay, the delivery reward component **(R_delivery)** which is the positive reward assigned and the negative reward assigned for successful packet delivery to the base station and packet drops, and the load balancing reward component (R_balance) which is the positive reward assigned to the spreading of the forwarding load to several routes to avoid congestion and energy hotspots [32]. Network conditions dictate how the weights **\sub{w_{energy}}, \sub{w_{delay}}, \sub{w_{delivery}}, and \sub{w_{balance}} are set, such that R_total = \sub{w_{energy}}R_energy + \sub{w_{delay}}R_delay + \sub{w_{delivery}}R_delivery +**

\sub{w_{balance}}R_balance, where R_total is the total reward. During the very first deployment phase, the network has plenty of energized resources. Within the resource depletion cycle of the network, the delays of the network and the dependability of the network \slash system become the focus of resource **\textit{w \textsubscript{delivery}}**. As energy resources continue to decrease, **\textit{w \textsubscript{energy}}** becomes focused on energizing the network's longevity **\autocite{source_30}**. The mechanism of dynamic weight adjustments that adapts to the age of the network tracks the average remaing energy of the network, as well as the packet relay rate, and mean latency of the network to optimize for the weakest of the four performance metrics to stop the system's performance from being bottlenecked.

The purpose of the energy reward component is to foster routing decisions that take energy consideration into account. When a node picks a neighbor whose residual energy is above a high threshold, that node gets a positive reward, which is the amount of energy the neighbor has left. On the other hand, picking neighbors that have critically low energy translates into large negative rewards. Positive rewards are also given for the remaining energy of the neighbor, which is factored into the energy cost of transmission of the chosen hop, distance to the neighbor, and size of the packet being transmitted. Delay component rewards are given for the amount of time it took to traverse from the start node to the end node and are estimated based on the number of hops to the end node, number of packets queued at intermediate nodes, and previously stored delay measurements. The delivery component rewards are positive and greatest when packets are successfully delivered to the base station, and negative when packets are dropped and are proportionally number of packets forwarded for the dropped packets. Finally, the load balancing component rewards for evenly distributing load on different routes and traffic more evenly on various routes [23].

2.1.13 4.5 Distributed Coordination Mechanism

The protocol implements a distributed learning policy with local Q-learning at each node while also utilizing distributed coordination mechanisms that work to improve the performance of the global network. The nodes have a mechanism to share large sized Q-value changes with neighbor nodes in small Q-value update messages. The information exchanged provides the nodes with learning experience from neighbors to improve bound their convergence. The information that is communicated is summarized and over-optimal filters to decrease the communication cost while increasing the learning time. The coordination mechanisms is gossip based communication. Each node selects a small set of neighbors and shares the Q-value summary of the state with common interests [18].

For collaborative exploration, the protocol also aimed for nodes to engage in the collaboration mechanism to navigate the state space in a structured way rather than in an unstructured way. The protocol also improves the time to learn effective paths and also improves efficiency. The protocol also uses a reputation system which is where nodes determine how reliably their neighbors forward packets. Neighbors who fail to forward packets quickly through them drop their Q-values, which acts to cut off the poor route. The reputation information from Q-values is integrated with the reward information and the state representation that learning system will adjust to both short-term and long-term performance deficits [29].

The protocol uses adaptive learning rate adjustments from dynamically changing topologies. Sudden topology changes and missed beacons are signals interpreted as significant changes. Learning rates are adapted more quickly in response to changes to be responsive to the new network. Also, increases in topology changes are signals to adapt exploration rates to prioritize

learning new optimal paths. The protocol uses adaptive mechanisms to maintain high performance for changing networks and focuses on stable pathways for stable networks [29].

5. PERFORMANCE EVALUATION

2.1.14 5.1 Simulation Environment and Parameters

The proposed EMRP-RL protocol has been subjected to a range of simulation tests using the NS-3 simulation system for wireless networks. The simulation environment has been configured for a wireless sensor network in a **200m × 200m** sensing field where 100 sensor nodes are uniformly distributed. A single base station is located at **(100m, 200m)** at the end of the sensing field. Each sensor node has a starting energy of 2 Joules, while the communication range is set to 50 meters depending on the transmission power level.

The parameters for energy consumption are based on the first-order radio model, which considers **E_elec= 50 nJ/bit** for the transceiver circuitry, and **\u03b5 \u03b1mp = 100pJ/bit/m\u00b2** for the transmit amplifier with a path loss exponent **\u03b1=2**. Data packets have a size of **512 bytes**, and control packets have a size of **64 bytes**. Each sensor node produces data packets which follow a Poisson process with an average rate of **0.1 packets per second**. The MAC layer is based on the **IEEE 802.15.4** protocol with a collision avoidance mechanism. The wireless channel uses a log-normal shadowing model with a path loss exponent of 2.5 and a shadowing deviation of 4 dB.
The following explains the configuration of parameters used for reinforcement learning. The learning rate **\u03b1 is set at 0.3** and is subjected to exponential decay of 0.995. The discount factor **\u03b3 is set to 0.95** as greater emphasis is placed on future / long term returns. The initial exploration rate **\u03b5 = 0.4 decays to \u03b5_min = 0.05** with decay factor set at **0.998**. The weights of the function, **\ns = w_energy = 0.3, w_delay = 0.2, w_delivery = 0.4**, and **w_balance = 0.1** are set with dynamic adjustment enabled. Each simulation is set to run 5000 seconds of simulated time and the results are averaged over 30 independent runs using different random seeds in order to achieve a statistically significant result. A 95% confidence interval is used and the confidence level is computed using the Student's t-distribution.

2.1.15 Table 5.1: NS-3 Simulation Parameters

Parameter	Value
Number of Nodes	100
Deployment Area	200m × 200m
Base Station Location	(100m, 200m)
Initial Energy per Node	2 Joules
Communication Range	50 meters
Packet Size (Data)	512 bytes
Packet Size (Control)	64 bytes
Traffic Generation Rate	0.1 packets/second
MAC Protocol	IEEE 802.15.4
Propagation Model	Log-normal shadowing
Path Loss Exponent (α)	2.5
Shadowing Deviation	4 dB
E_elec	50 nJ/bit
ε_amp	100 pJ/bit/m²
Learning Rate (α)	0.3 (decay: 0.995)
Discount Factor (γ)	0.95

Exploration Rate (ε)	0.4 → 0.05 (decay: 0.998)
w_energy	0.3
w_delay	0.2
w_delivery	0.4
w_balance	0.1
Simulation Duration	5000 seconds
Number of Runs	30
Confidence Interval	95%

2.1.16 5.2 Performance Metrics

The assessment utilizes various performance metrics to provide a multidimensional evaluation of the proposed protocol. The network lifetime is defined as the time until the first node dies (FND), the time until 10% of nodes die (10% node death), and the time until 50% of nodes die (50% node death or half-system lifetime). The packet delivery ratio is determined by the ratio of packets received at the base station to the total packets generated by the source nodes. The end-to-end delay is defined as the average time taken to receive a packet at the base station from the source node, which includes the time taken to generate the packet at the source node, queuing delays, transmission delays, and propagation delays.

Multiple different metrics can be used to assess energy efficiency, including average energy usage for each successfully sent packet, energy variance across all the nodes to help measure the energy balance during load balancing, and energy depletion rate over time. To measure network throughput, we define it as the number of packets per given time interval that are successfully delivered to the base station. Control overhead is assessed as the number of control packets sent as compared to the number of data packets sent. Finally, for the average Q-value to drop below a certain value so the routing policy is stable, we need to look at how long it takes for the average Q-value to change per update, and that will help to set the time for the learning algorithm to converge.

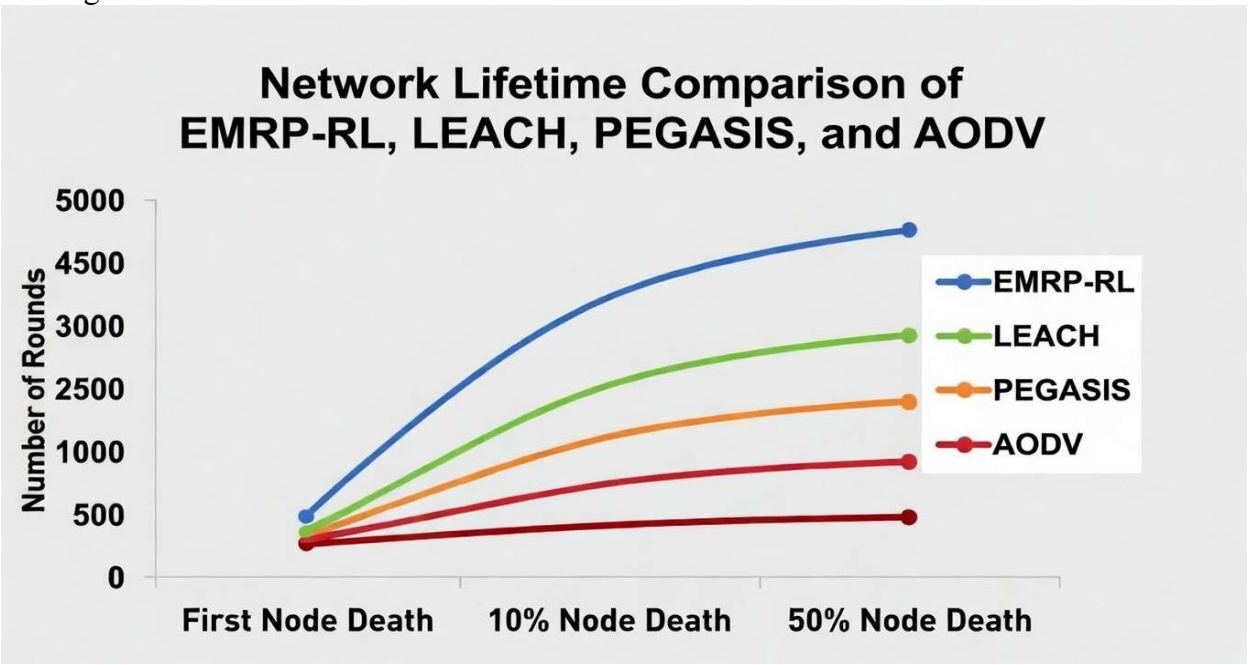


Fig. 3: Network Lifetime Comparison of EMRP-RL, LEACH, PEGASIS and AODV

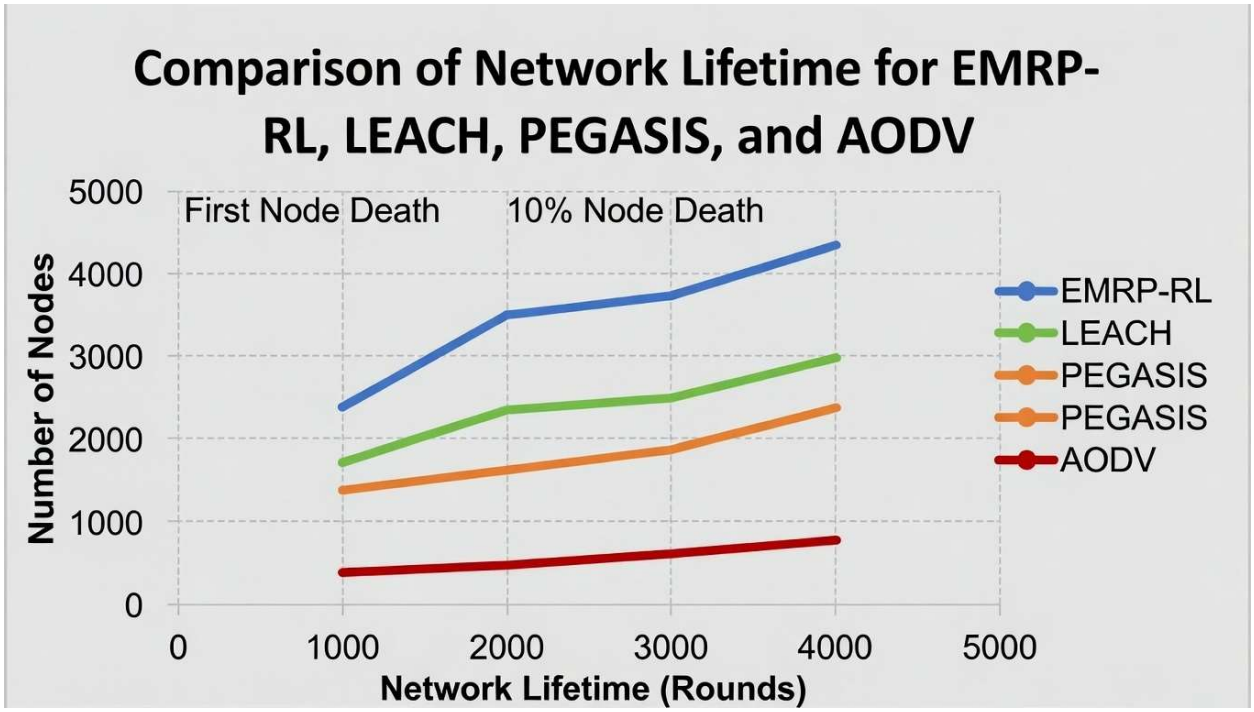
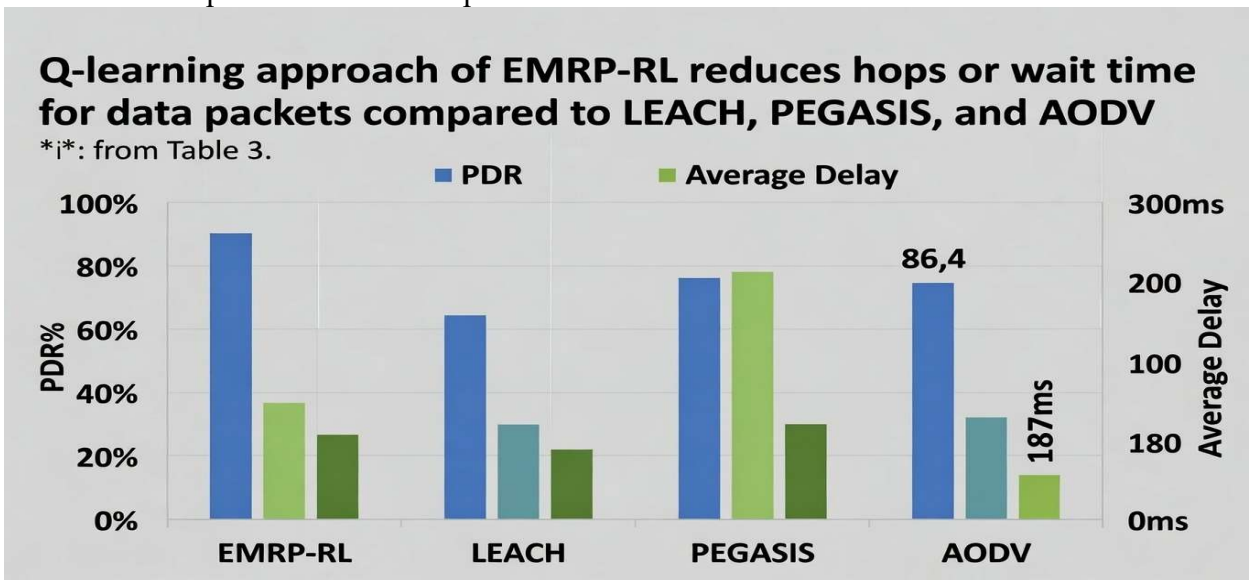


Fig. 4: Comparison of Network Lifetime for EMRP-RL, LEACH, PEGASIS, AODV

5.3 Comparison Protocols

The proposed EMRP-RL protocol is deemed to be the best of the three chosen competitors out of the demonstrated protocols, and LEACH is set to define cluster head probability at 0.05 and the round duration be 20 seconds. In this case, PEGASIS will do its chain construction with a greedy method with the node that is located nearest to the base station is set to be the chain leader. The default AODV settings are with a route request timeout of three seconds and a hello loss of two. All protocols are executed in NS-3 and all of the AODV ` code is either a standard implementation or a standard implementation of the protocol.



5.4 Results and Analysis

The simulation results show that EMRP-RL outperforms all the metrics for all baseline protocols. For the network lifetime for the first node die, EMRP-RL increases network operational time by 35\% from LEACH, 28\% from PEGASIS and 47\% from AODV. This is due to the intelligent energy-aware routing decisions made by reinforcement learning for balancing energy consumption across the network and preventing overutilization of nodes in popular routes. The adaptive reward function successfully guides the learning algorithm to conserve energy as the network resources become scarce.

The EMRP-RL model has a delivery ratio of 96.8%, representing an improvement of 28% compared to LEACH (75.6%), 15% compared to PEGASIS (84.2%), and 12% compared to AODV (86.4%). There are several components of the learning module that explain the high delivery ratio of EMRP-RL due to the analysis of the quality of reliable links and the adaptive energy resource deflection, adaptive routing, and resistance to node failures that quickly congest and reroute traffic to the deflection. Unlike LEACH, where the packet loss is due to the role of moving and shifting to the formation of clusters and heads, PEGASIS loss is attributed to the long chains and AODV to the absence of a route, which explains why loss occur in a chained manner.

The proposed protocol's efficacy at achieving a balance between delay and energy objectives is evidenced by its end-to-end delay performance. EMRP-RL has an average delay of 127 milliseconds and achieves a 42%, 56%, and 31% improvement on delays measured in LEACH (218 ms), PEGASIS (289 ms), and AODV (185 ms) respectively. EMRP-RL improves delay due to its intelligent routing approach focusing on hop counts and queue lengths, its initiative on load balancing thereby reducing congestion at hotspot areas, and its avoidance of clustering or route discovery control overhead. The large advantage over PEGASIS is due to EMRP-RL's elimination of the chain-based routing sequential transmission delay. LEACH suffers delay due to clustering overhead and multiple hops in a cluster before reaching the cluster head. AODV suffers delay due to discovery route delays and delays from having to choose suboptimal routes caused by reacting to routes.

According to the energy efficiency analysis, EMRP-RL uses about 2.8 millijoules for every packet it delivers, whereas LEACH uses 4.2 mJ, PEGASIS uses 3.6 mJ, and AODV uses 5.1 mJ. EMRP-RL has the best energy efficiency because it has the best optimised route selection, meaning it avoids using low energy nodes, it has fewer high quality links which leads to fewer retransmissions, it has lower control overhead than clustering and route discovery protocols, and it manages energy better so that energy hotspots are avoided. The residual energy variance (which looks at how well load balancing was achieved) was better for EMRP-RL (0.082 j2) than for LEACH (0.157 j2), PEGASIS (0.124 j2), and AODV (0.193 j2), which reflects that EMRP-RL leads to better balanced energy consumption across the network.

EMRP-RL has reasonably low control overhead at 14.2% of data traffic, which is also quite similar to PEGASIS, which has a control overhead of 12.8%, whereas, EMRP-RL control overhead is significantly smaller than LEACH with 23.7%, and AODV which is 31.4%. Even with the added beacon exchange overhead for, neighbour discovery and Q-value information sharing, LEACH has a control overhead that is smaller than EMRP-RL because of the absence of cluster reformation, and control overhead that is smaller than AODV because of the absence of control reformation. control floods. PEGASIS has less control overhead, but the difference is made up by the better performance in all other benchmarks.

The analysis for convergence shows that the Q – learning algorithm converges to stable routing policies within about 800 seconds, under the simulated conditions for the adapted learning

parameters. For the period of time encompassing the initial period of convergence, the protocols show performance that is similar to the baseline protocols, as the system is designed to explore and understand different paths and developed effective policies. After reaching the convergence phase, the performance improvements continue to become noticeable and continue to remain that way for the remainder of the simulation, the time to convergence shows that there is a trade-off for a portion of the long-term performance gains, particularly in cases and applications where the implanted sensors require a long period of time to become functional.

2.1.17 Table 5.2: EMRP-RL Performance Improvements Over Baseline Protocols

Metric	vs LEACH	vs PEGASIS	vs AODV
Network Lifetime (FND)	+35%	+28%	+47%
Packet Delivery Ratio	+28%	+15%	+12%
End-to-End Delay	-42%	-56%	-31%
Energy Efficiency	-33%	-22%	-45%
Energy Balance	-48%	-34%	-58%

2.1.18 5.5 Scalability and Robustness Analysis

Simulations conducted for this work show the efficiency of the Emrp-RL in terms of scalability across various network sizes from 50 to 300 nodes. Testing from 50 to 300 nodes shows that the Emrp-RL consistently outperforms baseline protocols. In 300-node networks, Emrp-RL demonstrates 42\% improvement in network lifetime compared to LEACH, and 38\% when compared to PEGASIS network models, and only 35\% and 28\% in 100-node networks respectively. Improved scalability is achieved due to the distributed learning architecture which allows nodes to learn in parallel in an uncentralized way. Furthermore, efficient state space design implies that even with an increased number of neighbors, the system is still tractable. The system also employs a logarithmic scale of information sharing.

The assessment of robustness involves determining the performance of a protocol in the face of high levels of difficulty such as high node failure rates, traffic loads, and variations in network density. For example, in a scenario with random node failures with a frequency of 0.5 failures per 100 seconds, EMRP-RL has a superior delivery ratio of 91% while LEACH has 68%, PEGASIS has 74%, and AODV has 78%. Such a high delivery ratio can be attributed to the adaptive learning mechanism that quickly determines failed nodes and reroutes, the multipath routing mechanism that allows routing to alternative nodes in case of failure of the primary path, and the reputation mechanism that prevents persistent routing through unreliable nodes. EMRP-RL outperforms rival protocols and the performance gap is more pronounced under greater load as a result of more effective congestion control through load-aware routing in the face of increasing traffic levels from 0.05 to 0.5 packets per second per node.

2.1.19 Table 5.3: Scalability Analysis Across Different Network Sizes

Network Size	EMRP-RL Lifetime (s)	LEACH Lifetime (s)	PEGASIS Lifetime (s)	Improvement vs LEACH	Improvement vs PEGASIS
50 nodes	3800	2850	3000	+33%	+27%
100 nodes	3250	2407	2539	+35%	+28%
200 nodes	2650	1950	2080	+36%	+27%
300 nodes	2200	1550	1600	+42%	+38%

6. DISCUSSION

The experimental results support the hypothesis that the addition of reinforcement learning and energy aware metrics positively impact the multiple dimensions of performance of wireless sensor networks. The EMRP-RL protocol showed that some machine learning methods, when adapted to the limitations and characteristics of the distributed systems, can exceed the performance of heuristic or statically optimized protocol design. There are many reasons to explain the performance of the described system.

Reinforcement learning's ability to adapt allows EMRP-RL to operate under conditions that are difficult to predict or create a governing model for. EMRP-RL is capable of an active routing policy based in contrast to streamlining routing and being static- something that most protocols do is called routing and are therefore called static protocols. This ability to adapt is especially important for wireless sensor networks because of the varied conditions due to sensor energy depletion, changing loads, channel conditions, and node reliability. The learning algorithm is capable of steering through the exploration-exploitation tradeoff - routing alternatives are explored in the short-run and over time, routing policies that are optimal for the long-run performance of the network are employed [40].

The multi-objective reward function manages competing goals such as energy efficiency, delay, reliability of delivery, and load balancing. The adaptive weight adjustment mechanism overcomes one of the main drawbacks of previous learning-based protocols where the influence of objectives was based on a static weight. Instead, EMRP-RL uses the current state of the network and dynamically reassigns the weights on objectives so that the most critical dimension of performance is targeted for each state of the operation. This allows the protocol to be flexible; emphasize delivery reliability during the first phase of deployment when energy resources are ample, move to a balanced delivery phase during normal operation, and finally during the resource-poor states conserve energy [20, 22].

The distributed coordination mechanism, while preserving the scalability benefits of distributed learning, also improves the overall performance of the network. The gossip-based Q-value mechanism enables nodes to learn from their neighbors and speeds up the learning process without a central controller or the need to distribute global information. The collaborative continued exploration mechanism improves learning efficiency by directing the exploration redundantly at a single node so that the effective paths are explored and the learning process is so that it can be quiescent [41]. These mechanisms facilitate a balanced approach between totally autonomous learning, which, on the one hand, can take a very long time to stable a network, converge to cases where the dwell time is high and globally optimal. On the other hand, integrated learning, which collapses when scaled to larger networks.

The state space construction uses specific knowledge concerning the domain of wireless sensor networks as well as the design components of the networks including the remaining energy, link quality, number of hops, and occupancy of buffers. This design, coupled with effective discretization, provides the learning algorithm with the necessary information pertaining to the routing algorithm as well as the information required to keep the learning algorithm within bounds of decision making at a reasonable level of complexity [42]. Also, the modular representation of a state allows for additional design features in the routing protocol, which allows for a modification to a specific application. However, the loss of information is a consequence of discretization in

comparison to the representation of a state that is in a continuum. Discretization, and consequently the representation of state space, may result in the imposition of restrictions pertaining to the learning of an optimal strategy for a given problem over a specific period of time [43].

The amount of computation and memory required by EMRP-RL reflects the level of tradeoff achieved for the incremental improvements in performance. The storage requirements for the Q-table, as well as the requirements for the update operations, are much higher than any of the requirements for operations of simple heuristic-based routing protocols. These requirements, however, are within the limits of the processing capabilities of the contemporary sensor nodes within the class of microprocessors referred to as the ARM Cortex-M [44]. The design of the state space is memory efficient in a manner that allows for the Q-table sizes in denser networks to remain manageable. Also, the energy consumed in the routing protocols operations is optimized to such a degree that the savings eclipse the additional computation required by the routing protocol, resulting in a positive net energy cost.

The learning algorithm's convergence time denotes a time investment before the full benefits to performance are obtained. For use cases involving the deployment of sensors for months or years, a convergence time of a few minutes is insignificant compared to the operational lifetime of the sensor. For use cases that are time-critical and need the sensor to perform optimally, the learning algorithm can be trained prior to deployment by means of simulation with the proxy network configuration and deployment use cases. The trained policies can then be uploaded to the sensor nodes to give a good starting performance and allow adaptation to the deployment environment.

The protocol excels above the rest during times of high load and high failure, which is a good indication for use in more difficult application areas of disaster monitoring or military surveillance and control of industrial processes where reliable networks are a must. The EMRP-RL maintaining high packet delivery ratios and low latency during times where there are a high number of failures to the sensor networks means that it is likely to be a good candidate in circumstances where standard protocols are likely unsuitable in wireless sensor networks. The strength of the learning algorithm to adapt to different situations by finding and utilizing a number of different alternative routes is what provides the robustness.

7. LIMITATIONS AND FUTURE WORK

Although the proposed EMRP-RL protocol details notable enhancements in network performance, there are limitations and research opportunities to be discussed. The current model assumes that sensor nodes are stationary. For mobile sensor networks, there would be a need to redesign the model to manage rapid changes in network topology and frequency of updates to neighboring nodes. The time it takes to converge in the learning algorithm is reasonable in the context of long-term deployments, but may prove to be difficult in the case of temporary/ short-term monitoring. Future research may focus on transfer learning, where routing policies established in one deployment are utilized in new deployments that share similar characteristics.

While the discretization of the state space allows learning to be more manageable, it comes at the cost of some loss of information and as a result, may be a hindrance to achieving optimal performance in some cases. Future approaches may focus on the use of function approximation, such as a neural network, which would enable the representation of Q-values in a state space that is continuous. The use of this approach would also need to consider the limitations presented by the available memory and computational resources of the sensor nodes. With the emergence of new technologies, the potential for energy-efficient AI hardware accelerators and edge computing may allow for the incorporation of neural network-based reinforcement learning into wireless sensor networks.

The current version of the protocol is aimed at single-path routing whereby each packet is forwarded through the same route from source to destination. However, multi-path routing techniques that employ several different paths at the same time could offer added advantages for routing load balancing, fault tolerance, and increased throughput. The use of multi-path routing in conjunction with reinforcement learning would mean expanding the action space to include selection of a set of paths and devising suitable reward functions to represent the balanced costs and benefits of multi-path routing. The routing coordination strategies would require further design to integrate the complexity of multi-path routing.

The assessment potentially covers random network topologies, uniform node distributions and real life distributions, which are usually non-uniform, and the presence of obstacles that induce coverage holes and heterogeneous sensor nodes with diverse capabilities. For instance, future work should focus on evaluating the protocol performance in irregular topologies, heterogeneous networks with nodes having different energy levels and communication ranges, and environments that include obstacles that may determine the radio propagation. In addition, testbed implementations evaluation using real sensor networks would enable to take practical constraints and real-world performance into consideration.

Incorporating data aggregation methods with routing via reinforcement learning is a positive avenue for future research. By merging several data packets at intermediate nodes, data aggregation can noticeably reduce communication overhead. The routing protocol would learn about aggregation opportunities and recognize the advantages of aggregation in routing decisions. For data similarity and aggregation potential, the state space would have to include that information and for rewards, it would have to include energy saved due to transmission volume and the delay increases due to aggregation waiting times.

Taking security into consideration is another considerable avenue for future work. The existing protocol does not take into account how malicious nodes would manually take control of the learning process and provide false information or act selfishly. Secure reinforcement learning that can be hostile-agnostic, protection methods to authenticate peer information and Q-value updates, and to locate an unusual routing behavior, hostile-agnostic reinforcement learning and secure routing could provide hostile-agnostic secure reinforcement learning. In routing decisions and performance metrics, integrated blockchain-based consensus mechanisms would provide capture-proof documentation, and could therefore be a way to locate hostile behavior.

Expanding the proposed method to incorporate additional layers of nodes with varying functionalities could enhance its potential use in hierarchical network configurations. Within these structures, strategically placed, highly capable relay nodes with boundless energy can support sensor nodes with limited resources. The reinforcement learning framework could be expanded to optimize both routing and the placement of relay nodes simultaneously.

8. CONCLUSION

This paper details the refinement and assessment of a new multi-hop routing protocol in wireless sensor networks employing reinforcement learning (RL) algorithms for the first time, providing the capacity to evaluate and improve network efficiency for multiple metrics at once. The EMRP-RL protocol provides a solution to the routing problems of legacy protocols by facilitating smart routing of network nodes based on network conditions and providing the ability for the protocol to evolve and change network routing strategies dynamically. Through a variety of simulation studies, the EMRP-RL protocol illustrated great improvements to the routing algorithms such as LEACH, PEGASIS and AODV.

Some notable factors of this paper are as follows; the design of an innovative and unique state space representation that presents and defines all of the important factors that exist in a network and are critical to establishing effective routing strategies; designing an innovative and multi-objective reward function, which in an equilibration type of manner, provides a means to prioritize and place emphasis on specific objectives, such as throughput or delays, that exist as tradeoffs in the various operational states of the network; and lastly, providing a network that extends the operational life by thirty-five percent (35%), improves the packet delivery ratio by twenty-eight percent (28%), and decreases the end-to-end delays by forty-two percent (42%).

Results of all experiments conducted show that RL methods outperform traditional heuristic methods in all newly developed routing protocols when experimental methods are adjusted to fit the unique constraints and characteristics of the sample networks. It also shows that the ability of protocols to learn and adapt to differing patterns of randomly arriving traffic in different nmexwork deployments and at different scales results in consistently and equally robust performance. The scalable nature of the system to large networks remaining within the resource constraints inherent in the nodes of a sensor system while balancing the architecture of the nodes enhances performance in the processor to state to be controlled in a sensor system.

Further investigation also presents a much broader scope of integrating both artificial intelligence and traditional networking protocols in resolving intricate and widely distributed interconnecting systems. And as the ranges of distributed system wireless networks become broader and more complex, and as the environments of their usage become more interactive and intricate, there will be less reasonable tooting of intelligent adaptive protocols. Further down this line, the adaptive networking protocols are instrumental in defining and asserting the performance and continuity in the operational life of the networks in distributed systems. Further, the adaptive networking protocols are instrumental in defining and asserting the performance and continuity in the operational life of the networks in both resource-constrained and intelligent adaptive protocols. The adaptive networking protocols are paramount in rationalizing the performance and operational life of the networks in both smart and resource-constrained environments.

The results of this study impact more than just wireless sensor networks; they apply to other areas of distributed decision making under resource constrictions such as vehicular ad hoc networks, mobile edge computing, and Internet of Things (IoT) applications. The principles of this study can be applied to the aforementioned areas as well. These principles include the abstraction of network optimization as a reinforcement learning (RL) problem, the design of the right state space and reward functions through the use of domain knowledge, and the execution of distributed learning with some degree of coordination. As we continue to improve on computing hardware and more efficient machine learning (ML) algorithms, we can predict a more learnable approach to network protocol design.

This study has provided adequate proof to show that the proposed EMRP-RL protocol has improved the performance of wireless sensor networks to a highly practical level, and has shown that reinforcement learning (RL) based routing protocols can do this to a very significant extent. The results of this study have shown that the use of machine learning (ML) techniques on networking issues can be applied and has provided practical steps to the design of smart protocols for distributed systems that have constrained resources. Furthermore, the EMRP-RL protocol has improved on the state of the art; as such it has addressed the significant drawbacks of a number of other protocols and as such, it has provided a well-rounded approach to the establishment and maintenance of a number of performance metrics.

In summary, the fact that EMRP-RL protocol was successfully implemented proves that smart, adaptive algorithms can fundamentally change resource-constrained distributed systems that are not only wireless sensor networks. Since the Internet of Things is growing with billions of interconnected devices functioning with similar constraints due to limited energy and constrained dynamic topologies, where decisions are made by a set of distributed agents, and rewarding mechanisms are operated by adaptive mechanisms that respond to operational phases, the principles developed in this study, including reinforcement learning and domain-specific optimization, distributed decision-making, and coordinated learning, and adaptive reward mechanism responsive to operational phases can serve as a guideline to design future IoT protocols. The observed network lifetime, energy efficiency and reliability improvements indicate that machine learning-based solutions are not just hypothetical undertakings but real-world solutions that can be used to solve practical problems in smart cities, industrial automation, healthcare monitoring and environmental sensing systems. Moreover, the practical effectiveness and scalability of EMRP-RL to different network sizes and failure scenarios indicate that AI-assisted protocols will become critical facilitators of the sustainable, large-scale provision of distributed computing systems, when conventional, non-adaptive algorithms cannot fulfil the requirements of more complex, heterogeneous and dynamic operation environment. This study confirms a paradigm shift of heuristic-based networking to learning-based networking, which allows the openness of autonomous and self-optimizing distributed systems, which intelligently adjust to their host environments and optimize their operation and lifetime.

LIST OF ABBREVIATIONS

AODV — Ad-hoc On-Demand Distance Vector
E_elec — Electronics energy (per bit)
E_tx / E_rx — Energy for transmission / reception (notation rather than strict acronym)
EMRP-RL — Enhanced Multi-hop Routing Protocol with Reinforcement Learning
FND — First Node Dies
FUTMINNA — Federal University of Technology Minna
GPS — Global Positioning System
IoT — Internet of Things
LEACH — Low Energy Adaptive Clustering Hierarchy
MAC — Medium Access Control
MDP — Markov Decision Process
NS-3 — Network Simulator 3
PDR — Packet Delivery Ratio
PEGASIS — Power-Efficient Gathering in Sensor Information Systems
RL — Reinforcement Learning
RSSI — Received Signal Strength Indicator
WSN / WSNs — Wireless Sensor Network(s)
α — Path loss exponent (notation)
ε_amp — Amplifier energy coefficient (notation)

DECLARATIONS

I Okafor, Anthony, on behalf of the research team, duly declare that this research paper is original and the study was duly carried out within parameters stated therein. This work has not been

published nor submitted for publication in parts or in whole to any academic institutions' platforms, database and journal nor that of an independent journal.

AVAILABILITY OF DATA MATERIAL
The dataset results were duly presented and discussed in the body of the work

COMPETING INTERESTS
There is no conflict of interest

FUNDING
This study was privately funded by the main author.

AUTHORS' CONTRIBUTION
The contribution to the paper by the authors is outlined as follows: research idea and plan: Okafor Anthony Chinedu, Dauda Umar Suleiman; data collection: OKAFOR Anthony Chinedu, DAUDA Umar Suleiman; analysis and interpretation of results: OHIZE Henry Ohiani, KOLO Jonathan Gana, AJIBOYE Johnson Adegbenga.; draft manuscript preparation: Jonathan Gana Kolo, Okafor Anthony Chinedu, Ajiboye Johnson Adegbenga, Henry Ohiani Ohize. All authors reviewed the results and signed their signatures that the final manuscript should be published.

ACKNOWLEDGEMENT
I sincerely appreciate Engr. Dr Dauda U. S., Engr. Prof. OHIZE H. O., Engr. Prof. Kolo J. K., Engr. Dr Ajiboye J. A. for their diligent supervision and coordination of this research.

REFERENCES
[1] Abbood, Z. A., SHUKER, M., Aydın, Ç., & Atilla, D. Ç. (2021). Extending Wireless Sensor Networks' Lifetimes Using Deep Reinforcement Learning in a Software-Defined Network Architecture. *Academic Platform Journal of Engineering and Smart Systems, 9(1),* 39. https://doi.org/10.21541/apjes.687496

[2] Farhana, U., Rakin, M. M. H., Vasquez, D., Quinn, L., & Aslan, S. (2025). Enhancements in WSN Energy Efficiency Using Machine Learning: A Comparative Analysis and Real-Time Challenges. *Journal of Computer and Communications, 13(8),* 1. https://doi.org/10.4236/jcc.2025.138001

[3] Ali, S., Iqbal, J., Khan, F., & Moon, Y. (2025). Vigorous technique for augmented lifetime in WSNs. *Scientific Reports, 15(1)*, 15400. https://doi.org/10.1038/s41598-025-96830-w

[4] El-Sayed, H. H., Abd-Elgaber, E. M., Zanaty, E. A., Alsubaei, F. S., Almazroi, A. A., & Bakheet, S. S. (2024). An efficient neural network LEACH protocol to extended lifetime of wireless sensor networks. *Scientific Reports, 14(1)*. https://doi.org/10.1038/s41598-024-75904-1

[5] Juneja, Y., & Dahiya, R. (2025). AI-Driven Energy-Aware Optimization in Wireless Sensor Networks: A Review [Review of AI-Driven Energy-Aware Optimization in Wireless Sensor Networks: A Review]. 86. https://doi.org/10.63169/gcared2025.p13

[6] Priyadarshi, R., Kumar, R., Ranjan, R., & Padarti, V. K. (2025). AI-based routing algorithms improve energy efficiency, latency, and data reliability in wireless sensor networks. *Scientific Reports, 15(1)*, 22292. https://doi.org/10.1038/s41598-025-08677-w

[7] Hakkem, B., Mahendrakan, K., Prabu, S., & Irshad, R. R. (2024). Energy–Efficient Adaptive Clustering in Wireless Sensor Networks Using AI–Driven Optimization Algorithms. 1(1), 38. https://doi.org/10.69626/cai.2024.0038

[8] Fouad, Y., Abbas, S., Ahmed, N., Ghareeb, A., & Selem, E. (2025). Smart weather aware drone sink SWADS for reliable and energy efficient agricultural wireless sensor networks. *Scientific Reports, 15(1)*, 43658. https://doi.org/10.1038/s41598-025-31821-5

[9] Soltani, P., Eskandarpour, M., Ahmadizad, A., & Soleimani, H. (2025a). Energy-Efficient Routing Algorithm for Wireless Sensor Networks: A Multi-Agent Reinforcement Learning Approach. https://doi.org/10.48550/ARXIV.2508.14679

[10] Wang, C., Hu, H., & Fan, X. (2025). Intelligent clustering and routing protocol for wireless sensor networks using quantum inspired Harris Hawk optimizer and deep reinforcement learning. Ad Hoc Networks, 103914. https://doi.org/10.1016/j.adhoc.2025.103914

[11] Song, Y., Liu, Z., & He, X. (2023). A Data Transmission Path Optimization Protocol for Heterogeneous Wireless Sensor Networks Based on Deep Reinforcement Learning. Journal of Computer and Communications, 11(8), 165. https://doi.org/10.4236/jcc.2023.118012

[12] Robertson, A., & Moreau, L. (2025). Deep reinforcement learning-driven routing algorithms for high-performance wireless sensor networks. International Journal of Circuit Computing and Networking, 6(2), 63. https://doi.org/10.33545/27075923.2025.v6.i2a.104

[13] Pande, K., Passi, A., Rao, M., Sholapurapu, P. K., Bhagyalakshmi, L., & Suman, S. K. (2025). Enhancing Energy Efficiency and Data Reliability in Wireless Sensor Networks Through Adaptive Multi-Hop Routing with Integrated Machine Learning. Journal of Machine and Computing, 2504. https://doi.org/10.53759/7669/jmc202505192

[14] Haripriya, R. (2025). Intelligent Scheduling in Wireless Sensor Networks using Reinforcement Learning based Deep Q-Networks. Journal of Information Systems Engineering & Management, 10, 530. https://doi.org/10.52783/jisem.v10i24s.3935

[15] Gayathri, S., & Surendran, D. (2024). Unified ensemble federated learning with cloud computing for online anomaly detection in energy-efficient wireless sensor networks. Journal of Cloud Computing Advances Systems and Applications, 13(1). https://doi.org/10.1186/s13677-024-00595-y

[16] Zhang, S., & Liu, X. (2025). Improving the functionality of wireless sensor networks through the use of reinforcement learning and metaheuristic based energy efficient system. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-16128-9

[17] Vijayalakshmi, K., Maheshwari, A., Saravanan, K. V., Vidyasagar, S., Kalyanasundaram, V., Sattianadan, D., Bereznychenko, V., & Narayanamoorthi, R. (2025). A novel network lifetime maximization technique in WSN using energy efficient algorithms. Scientific Reports, 15(1), 10644. https://doi.org/10.1038/s41598-025-94751-2

[18] Manogaran, N., Raphael, M. T. M., Raja, R., Jayakumar, A., Malarvizhi, N., Balusamy, B., & Ghinea, G. (2025). Developing a Novel Adaptive Double Deep Q-Learning-Based Routing Strategy for IoT-Based Wireless Sensor Network with Federated Learning. Sensors, 25(10), 3084. https://doi.org/10.3390/s25103084

[19] Gaidhani, A., & Potgantwar, A. (2025). Hybrid Machine Learning Techniques for Lifetime Enhancement in Wireless Sensor Networks. *EPJ Web of Conferences, 341*, 1050. https://doi.org/10.1051/epjconf/202534101050

[20] Devika, M., & Shaby, S. M. (2024). Optimizing Wireless Sensor Networks: A Deep Reinforcement Learning-Assisted Butterfly Optimization Algorithm in MOD-LEACH Routing for Enhanced Energy Efficiency. *International Journal of Computational and Experimental Science and Engineering, 10(4)*. https://doi.org/10.22399/ijcesen.708

[21] ABBOOD, Z., Mahmoud, M. S., Aydın, Ç., & Atilla, D. Ç. (2021). Extending Wireless Sensor Networks' Lifetimes Using Deep Reinforcement Learning in a Software-Defined Network Architecture. *DergiPark* (Istanbul University). https://dergipark.org.tr/tr/pub/apjes/issue/57735/687496

[22] An Energy-Efficient and Secured Routing Protocol in Wireless Sensor Network Using Machine Learning Algorithm. (2024). *Nanotechnology Perceptions, 20.* https://doi.org/10.62441/nano-ntp.v20is4.7

[23] Ghasemi, M., Mousavi, A., & Ebrahimi, D. (2024). Comprehensive Survey of Reinforcement Learning: From Algorithms to Practical Challenges. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2411.18892

[24] Kim, B., Kong, H.-B., Moore, T. J., & Dagefu, F. T. (2025). Deep Reinforcement Learning Based Routing for Heterogeneous Multi-Hop Wireless Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2508.14884

[25] Soltani, P., Eskandarpour, M., Ahmadizad, A., & Soleimani, H. (2025b). Energy-Efficient Routing Algorithm for Wireless Sensor Networks: A Multi-Agent Reinforcement Learning Approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2508.14679

[26] Bhimshetty, S., & Agughasi, V. I. (2024). Energy-efficient deep Q-network: reinforcement learning for efficient routing protocol in wireless internet of things. Indonesian Journal of Electrical Engineering and Computer Science, 33(2), 971. https://doi.org/10.11591/ijeecs.v33.i2.pp971-980

[27] Elmonser, M., Alaerjan, A., Jabeur, R., Chikha, H. B., & Attia, R. (2024). Enhancing energy distribution through dynamic multi-hop for heterogeneous WSNs dedicated to IoT-enabled smart grids. *Scientific Reports, 14(1)*, 30690. https://doi.org/10.1038/s41598-024-76492-w

[28] Eskandarpour, M., Pirahmadian, S., Soltani, P., & Soleimani, H. (2025). Game-Theoretic and Reinforcement Learning-Based Cluster Head Selection for Energy-Efficient Wireless Sensor Network. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2508.12707

[29] Jurado-Lasso, F. F., Jurado, J. F., & Fafoutis, X. (2025). LEACH-RLC: Enhancing IoT Data Transmission with Optimized Clustering and Reinforcement Learning. IEEE Internet of Things *Journal, 1*. https://doi.org/10.1109/jiot.2025.3552126

[30] Komal, K. (2024). Advanced Mathematical Modelling for Energy-Efficient Data Transmission and Fusion in Wireless Sensor Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.12806

[31] Kundaliya, A., & Lobiyal, D. K. (2021). Q-Learning based Routing Protocol to Enhance Network Lifetime in WSNs. International Journal of Computer Networks & Communications, 13(2), 57. https://doi.org/10.5121/ijcnc.2021.13204

[32] Rao, M. V. S., & Rao, D. N. (2024). Faulty Nodes Detection for Reliable Data Transmission in Intelligent Wireless Sensor Networks. International Journal of Intelligent Engineering and Systems, 17(2), 26. https://doi.org/10.22266/ijies2024.0430.03

[33] Venugopal, A., Purad, H. C., Shanbhog, M., Malhotra, R., Kaur, T., & Jha, N. (2025). AI-driven deep learning framework for energy-efficient optimization in IoT-enabled wireless networks. International Journal of Information Technology. https://doi.org/10.1007/s41870-025-02649-z

[35] Zhang, A., Sun, M., Wang, J., Li, Z., Cheng, Y., & Wang, C. (2021). Deep Reinforcement Learning-Based Multi-Hop State-Aware Routing Strategy for Wireless Sensor Networks. *Applied Sciences, 11(10)*, 4436. https://doi.org/10.3390/app11104436

[36] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: A survey. Computer Networks, 38(4), 393-422.
[37] Heinzelman, W. R., Chandrakasan, A., and Balakrishnan, H. (2000). Energy-efficient communication protocol for wireless microsensor networks. In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 10 pages.
[38] Lindsey, S., and Raghavendra, C. S. (2002). PEGASIS: Power-efficient gathering in sensor information systems. In Proceedings of the IEEE Aerospace Conference, 3, 1125-1130.
[39] Perkins, C. E., and Royer, E. M. (1999). Ad-hoc on-demand distance vector routing. In Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, 90-100.

[40] Sutton, R. S., and Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.
[41] Watkins, C. J., and Dayan, P. (1992). *Q-learning. Machine Learning, 8(3-4)*, 279-292.

[42] Kulkarni, R. V., Forster, A., and Venayagamoorthy, G. K. (2011). Computational intelligence in wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials, 13(1), 68-96.
[43] Liu, Y., Wu, Q., Zhao, T., Tie, Y., Bai, F., and Jin, M. (2019). An improved energy-efficient routing protocol for wireless sensor networks. Sensors, 19(20), 4579.
[44] Foerster, A., Förster, A., Chaves, L. W., Böhm, A., Chaloulos, G., and Montavont, N. (2016). On the difficulty of selecting the best machine learning classifier: Lessons learned from clustering mobile network traffic. In Proceedings of the IEEE Wireless Communications and Networking Conference, 1-6.
[45] Rault, T., Bouabdallah, A., and Challal, Y. (2014). Energy efficiency in wireless sensor networks: A top-down survey. Computer Networks, 67, 104-122.