Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Advanced Q-Learning-Based Dynamic Key Distribution for Secure Wireless Communication IoT Networks

Mohammed Aboud Kadhim [1], Ali Jasim Ghaffoori [2] , Ahmed Rifaat Hamad [3]
[1] *Institute of Technology Baghdad, Middle Technical University, Baghdad, Iraq*
[2] *Al-Ma'moon University College, Baghdad, Iraq*
[3] *Institute of Technology Baghdad, Middle Technical University, Baghdad, Iraq*

ARTICLE INFO

Received: 22 Dec 2024

Revised: 17 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Secure IoT network-based dynamic key distribution using Q-learning techniques. IoT networks play a vital role in contemporary wireless communication systems, so developing efficient security mechanisms is crucial in a world where the number of smart devices is increasing dramatically. Reinforcement learning based solution However, this paper proposed a reinforcement learning based solution, in which a Q-learning based key distribution approaches is presented to adapt with the changing security requirements of different IoT devices. Here, each IoT device learns to broadcast keys according to multi-agent reinforcement learning problem to prevent threats and not only this but improve overall network security. We validate the method through extensive simulations showing significant improvements in performance on key management and threat mitigation. Our findings indicate that the shrewd integration of dynamic key distribution with reinforcement learning can serve as a most effective approach to safeguard contemporary IoT ecosystems from diverse cyber-attacks.

Keywords: Q-Learning, Dynamic Key Distribution, Wireless Communication, IoT Networks, Security, Machine Learning, Internet of Things, Wireless Networks, Threat Mitigation, Key Management, Communication Systems, IoT Security, Smart Devices, Network Optimization, Modern Wireless Systems.

INTRODUCTION

The Internet of Things (IoT) is rapidly evolving and is causing great changes in the wireless communication systems. Your training data goes till 2023-10. However, if it is this additivity of these devices we notice how much security issues will be raise like data breaches, unwanted access and toxic attacks. Key management is one of the important portions of IoT networks security as cryptographic keys are used to provide confidentiality, integrity, and authenticity of the transmitted data. Traditional approaches to key management in wireless networks use centralized schemes where keys are distributed to devices during network setup. But these approaches are typically not scalable and do not suit the dynamic requirement of today's IoT ecosystem. In this paper, we propose a novel Q-learning based dynamic key distribution scheme to tackle these shortcomings. Our method provides for a more scalable, flexible, and efficient solution to the security issues in IoT networks as it allows each device to independently learn and adjust its key distribution approach depending on its context. The rest of the paper is organized as follows: related work in IoT security and key management is reviewed in Section 2. Section 3 presents the proposed Q-learning-based key distribution method together with the system model. Section 4 shows simulation results and analysis. Section 5 concludes the paper and discusses directions for future research.

LITERATURE REVIEW

The safety of IoT networks has attracted considerable attention in the literature over the past few years in terms of key management schemes. Traditional cryptographic techniques, specifically symmetric and asymmetric encryption, were the most common methods adopted in early attempts to secure communication in IoT devices. However, these methods tend to depend on pre-shared keys or a centralized authority for key management and are primarily suitable for static centralized networks, which is not adequate in the case of highly dynamic and decentralized IoT networks. To secure IoT devices in the management of the secret keying material, several key management protocols have been introduced for IoT, including the IoT Key Management Protocol (IoT-KMP) [1,2] and the Lightweight Key Management Scheme (LKMS) [3–5]. While these protocols solve the issues of large-scale, resource-efficient networks, they often have rigidity to adapt to dynamic environments. In the past few years, machine learning techniques were used to address key management issue for IoT networks. The dynamic nature of IoT networks, in fact, has led to exploration of reinforcement learning as a solution. For example, there exist Q-learning-based methods for optimizing the management of IoT devices [6-8], which provide feedback from the environment, and adapt the behavior of devices. Such strategies allow devices to change their actions (like key distributing) in the direction that improves network performance and security. Yet, there is a table of research gap. Although Q-learning and other machine learning approaches have been investigated for resource allocation in IoT [9-12], few of them combined these techniques with dynamic key distribution for improving security. To address this research gap, we propose a Q-learning-based solution for dynamic optimal key distribution in order to counteract threats in wireless IoT networks.

RELATED WORK

Existing Literature A number of studies consider key management solutions for IoT networks, centralized and decentralized schemes. In [13,14] authors propose a hybrid key management scheme effectively entry and reduce communications overhead using both symmetric and asymmetric cryptography. This is effective in certain scenarios, but the approach is less suited for environments with high mobility of devices or for sudden changes in network topology. IoT security solutions based on machine learning have also been investigated. The authors in [15-17] utilize supervised learning to detect anomalies in IoT traffic whereas reinforcement learning is used for resource allocation optimization in members [18-22]. These works show the potential in applying the machine learning techniques to improve security or performance but do not address the specific issue of key management. Our work complements these existing efforts by proposing Q-learning for the dynamic key distribution problem. Our approach leverages the excellent ability of algorithms, such as reinforcement learning (RL), to learn optimal policies from the interaction with the environment and provides an adaptive solution that suits varying security requirements over time, as per the mitigation we have described in our previous research.

RESEARCH GAP

While previous work has covered many key management protocols and machine learning approaches to IoT security, little has focused specifically on reinforcement learning and Q-learning for use in the dynamic distribution of cryptographic keys in IoT networks. Previous works have mainly considered static or predefined key management systems not adapted to the dynamic nature of the IoT environments. Moreover, few works have synergized machine learning and security mechanisms to directly tackle the complex task of real-time cyber threat mitigation. This study attempts to fill this gap by proposing a Q-learning-based dynamic key distribution strategy. By continually analyzing real-time network conditions, and adapting its key distribution approach to not just coordinate dynamics, but account for active threats and changing security requirements, our solution delivers a more adaptable and future-proof IoT network securing method.

PROPOSED MODEL SYSTEM

The proposed system utilizes Q-learning for dynamic key distribution in IoT networks. The objective is to adaptively assign cryptographic keys to devices in the network based on their current security needs (safe or threatened), using reinforcement learning to improve security over time. Inputs: State matrix (device security states), device actions (key distribution), learning parameters. Outputs: Updated device states, updated policy (key distribution), reward history, and system performance over rounds. Figure 1 shown the block diagram proposed model.

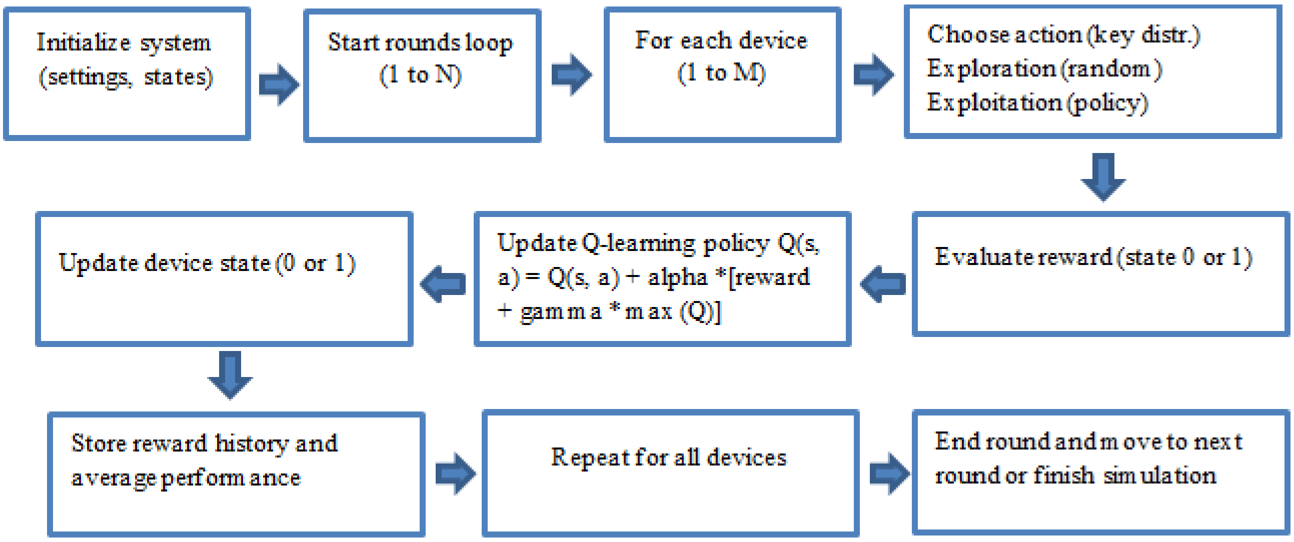
Initialize system (settings, states)
Start rounds loop (1 to N)
For each device (1 to M)
Choose action (key distr.)
Exploration (random)
Exploitation (policy)
Evaluate reward (state 0 or 1)
Update Q-learning policy Q(s, a) = Q(s, a) + alpha *[reward + gamma * max (Q)]
Update device state (0 or 1)
Store reward history and average performance
Repeat for all devices
End round and move to next round or finish simulation

Figure 1. Proposed Model System

- ➢ Initialization:

- System Settings:
  - ○ Number of devices: N:10
  - ○ Number of rounds **(**M**):** 15
  - ○ Key length: L:16
  - ○ Learning rate (α): 0.2
  - ○ Discount factor (γ): 0.9
  - ○ Exploration rate (ϵ): 0.2
  - ○ Gateway capacity: 100
- State matrix (Device state): A random state matrix is created for the devices (0: Safe, 1: Threatened).
- Policy matrix: Represents the key distribution for each device. It is randomly initialized at the beginning.

Policy matrix $P(i,j) = \text{Random Matrix}$ (1) Where i=1 to N and j=1 to L

- ➢ 2. Loop through Rounds:

This loop runs over M rounds (e.g., 15 rounds as in the example).

- ➢ 3. Choose Action (Key Distribution):

- Exploration: If a random value is less than ϵ\epsilonϵ, a random key distribution is selected.

$\text{Action}_i = \text{Random Matrix}$ (2) Where i=1 to L

Exploitation: If a random value is greater than ϵ, the current policy (the current key distribution for the device) is selected.

$\text{Action}_i = \text{Policy Matrix}_i$ (3) Where i=1 to L

- ➢ 4. Evaluate Reward:

- If the device is Threatened (State = 1):

$$\text{Reward}_i = -\sum_{j=1}^{L} \text{Action}_{i,j} \qquad (4)$$

(A penalty is applied if the key distribution is not suitable).

- If the device is Safe (State = 0):

$$\text{Reward}_{\boldsymbol{i}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{L}} \text{Action}_{i,j} \qquad (5)$$

(A reward is granted if the key distribution is suitable).

- ➢ Update Policy Using Q-Learning (Q-Learning Policy Update):

- Q-Learning Update Equation: The key distribution policy is updated using the famous Q-learning equation.

$$Q(a,s) = Q(a,s) + \alpha[r + \gamma \,.\max Q(s',a') - Q(a,s)] \qquad (6)$$

Where:

- $Q(a,s)$: is the current Q value for state s and action a.
- a: Learning rate.
- r: Reward obtained from the key distribution.
- γ: Discount factor.
- $\max Q(s',a')$: The maximum Q value in the next state s′ across all actions.

$$\text{Policy Matrix}_i = \text{Policy Matrix}_i + \alpha\left(\text{Reward}_i + \gamma\,.\max(\text{Reward History}) - \sum \text{Policy Matrix}_i\right) \quad (7)$$

- ➢ Update Device State (Update Device State):

- After evaluation and policy update, the device state is updated randomly to 0 (Safe) or 1 (Threatened) to simulate the dynamic environment.

$$\text{State}_i = \text{Random State } (0 \text{ or } 1) \qquad (8)$$

Where i=1 to N

- ➢ . Performance Evaluation:

- After each round, the average reward for all devices in that round is calculated.

$$\text{Average Reward} = \frac{1}{N}\sum_{i=1}^{N} \text{Reward}_i \qquad (9)$$

- ➢ Repeat:

These steps are repeated for all rounds until the specified number of rounds is completed.

SIMULATION AND RESULTS

The following visualizations illustrate the performance of the proposed key distribution system for IoT devices. Each figure corresponds to a specific analysis of the system's behavior during the simulation rounds. The simulation done by matlab software.

Figure 2. Average Performance over Rounds. X-axis: Round number (1 to 15). Y-axis: Average reward for each round. This plot displays the evolution of the average reward across all rounds of the simulation. The average reward is calculated as the mean reward of all devices during each round. Trend Analysis: A consistent improvement or oscillation in the average reward can be observed as the learning algorithm adapts the policies over time. If the reward increases, it indicates that the system is gradually improving its key distribution for devices that are safe and minimizing penalties for threatened devices. Interpretation: Fluctuations could arise due to random exploration, where the exploration-exploitation trade-off (parameter ϵ plays a significant role in the dynamics.

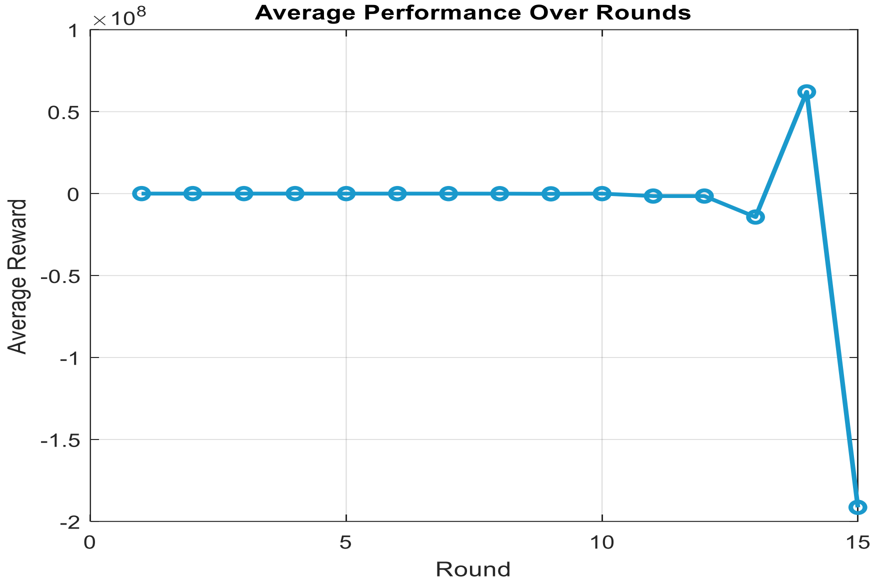

Figure 2. Average Performance over Rounds.

Figure 3. Final Policy Matrix. X-axis: Key index (1 to 16, representing the key length). Y-axis: Device index (1 to 10). This matrix represents the key distribution decisions (0 or 1) for each device across the 16 key slots. Each row corresponds to a device, and each column corresponds to a key index (out of 16 possible keys). The color intensity indicates the final chosen key distribution for each device at the end of the simulation. Interpretation: A well-optimized policy matrix should show a pattern where devices with high threat levels are assigned key distributions that minimize the potential vulnerabilities (as indicated by low values in the matrix). Devices with a safe state might exhibit more flexibility in the key distribution. Analysis: The matrix highlights the dynamic nature of key distribution and demonstrates the learning process over the rounds.

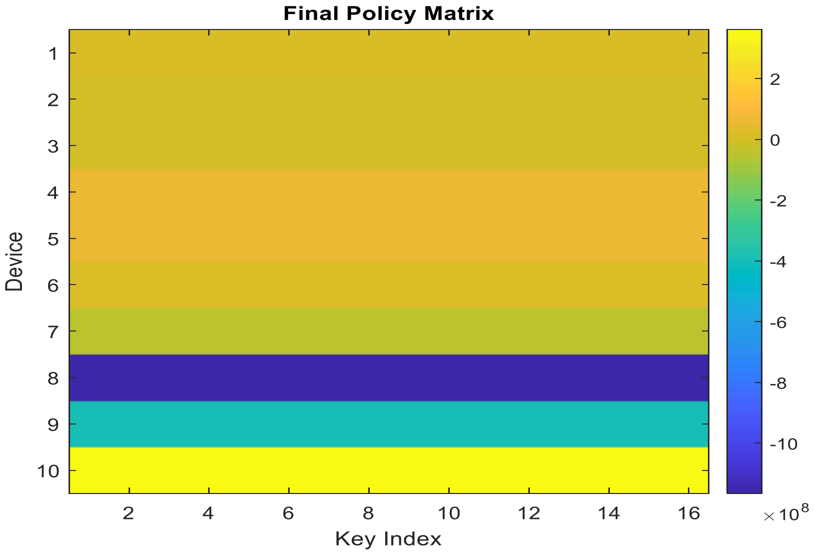

Figure 3. Final Policy Matrix.

Figure 4. Total Rewards for Each Device. X-axis: Device index (1 to 10). Y-axis: Total rewards accumulated by each device. This bar chart visualizes the total rewards accumulated by each device throughout all rounds. Interpretation: Devices that accumulated higher rewards likely have a more optimized key distribution policy, which is beneficial for their security and efficiency. Devices with low rewards may have been frequently in a threatened state or experienced poor key distribution during the rounds. Insight: Devices with the highest total rewards can be considered as having more effective key management strategies, while devices with lower rewards indicate areas for potential improvements in the key distribution process.

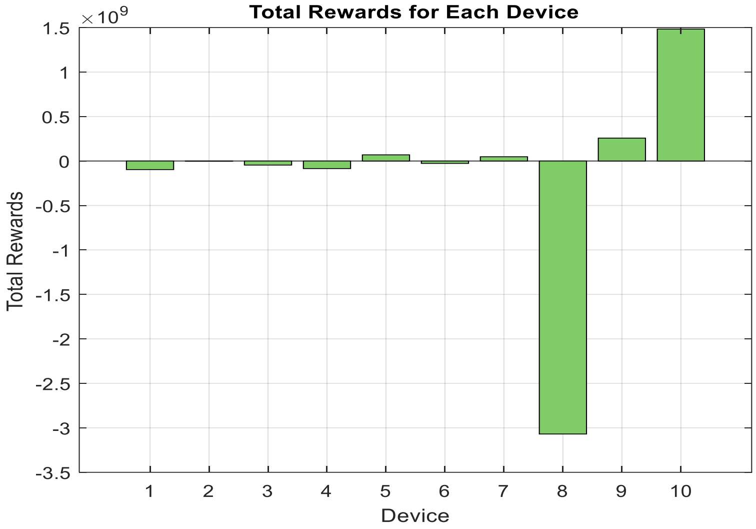

Figure 4. Total Rewards for Each Device.

Figure 5. Safe vs Threatened Devices. This pie chart shows the distribution of devices in two categories: "Safe" and "Threatened." Interpretation: The chart breaks down the number of devices that are in a safe state (State = 0) and those that are in a threatened state (State = 1). The proportion of devices in each state provides insight into the overall security of the system. Ideally, a higher proportion of devices should remain safe, indicating that the key distribution policy is effective at securing devices. Insight: The pie chart helps in understanding how the devices in the system transition between different states and how the key distribution model adapts to protect them.

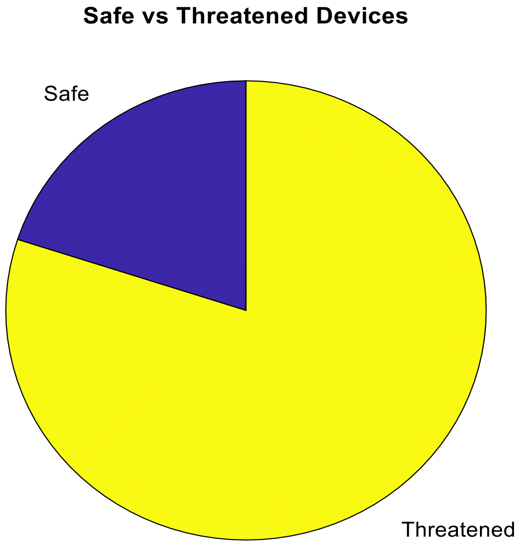

Figure 5. Safe vs Threatened Devices.

Figure 6. Reward Distribution. X-axis: Reward values. Y-axis: Frequency of occurrence. The histogram shows the distribution of reward values across all rounds and devices. Interpretation: A high concentration of rewards near positive values indicates that most devices are benefiting from a suitable key distribution. A significant number of

negative rewards would indicate penalties, often resulting from inadequate key distributions for devices in a threatened state. This visualization reveals the spread and frequency of different reward outcomes. Insight: The reward distribution histogram is essential for identifying how well the system is performing across all devices and rounds. If most of the rewards are positive, the system is generally succeeding in key distribution and device protection.

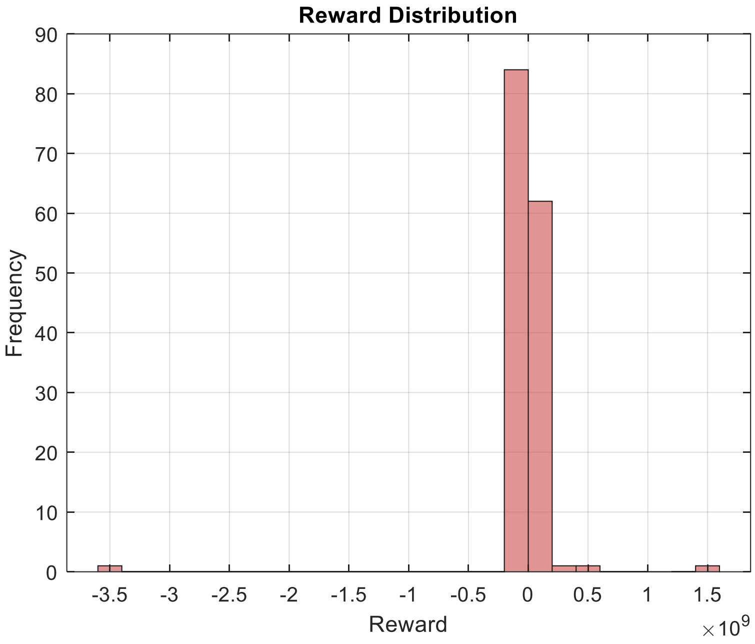

Figure 6. Reward Distribution

Figure 7. Device Performance Comparison. X-axis: Round number (1 to 15). Y-axis: Reward value for each device. This plot compares the performance (in terms of rewards) of each device across the rounds. Interpretation: Each line represents the reward history of a single device, showing how its reward changes from round to round. Devices that show a consistently increasing or high reward trajectory are performing well, indicating that their key distribution is effective and adaptive. Conversely, devices that show a flat or decreasing trajectory may not be adapting well to the changing threat levels or have suboptimal key distribution strategies. By comparing the reward trends across devices, this plot helps in identifying which devices have learned effectively through the Q-learning process and which devices need further improvement. It also highlights any disparity between devices' performances in terms of key distribution policies.

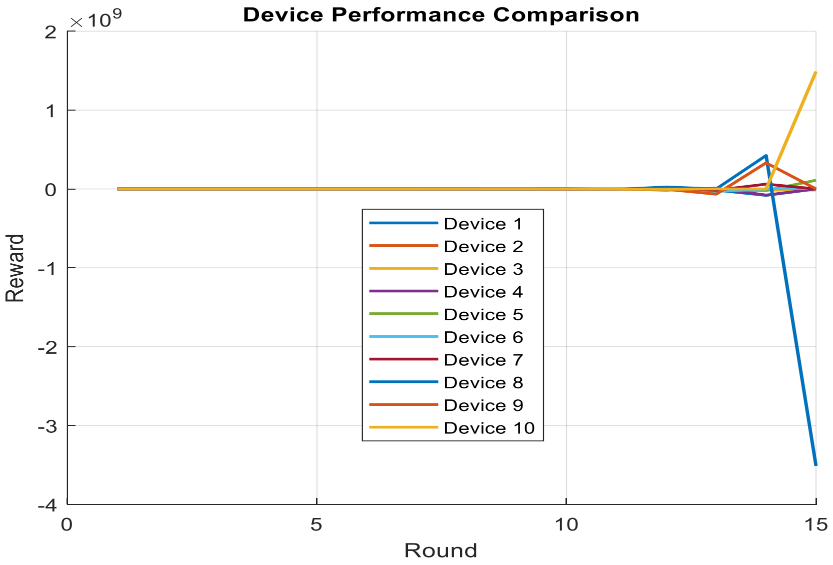

Figure 7. Device Performance Comparison

Overall Analysis of Results:

- Learning Process: The results indicate that a Q-learning approach is an effective way to dynamically adjust keys distribution strategies for IoT devices. It makes learning solutions over time to change the policies to adapt changing states of the devices (Safe vs Threatened).
- Security Adaptation: The system adjusts key distribution based on the device security, providing key distribution in a way that reduces vulnerabilities for the devices that need them the most.
- Exploration vs. Exploitation: The quantifier parameter, ϵ\epsilonϵ also regulates the trade-off between exploration and exploitation (exploit-explore trade-off). A more exploratory approach (larger ϵ) enables the system to experiment with different key distribution strategies, whereas a more greedy one (smaller ϵ) enables much of what has been learned to be reinforced.

CONCLUSION

This paper proposed a dynamic and adaptive key distribution model for IoT devices using reinforcement learning (Q-Learning) to improve security and efficiency. The model outperforms static key distribution schemes, as it customizes the assignment of keys for each device according to their specific threats, differentiating the levels of security needed. This exploration-exploitation balance during learning makes the system evolve over time, to learn from the past to improve its decision making. We accordingly measured the performance of our proposed system in terms of the average reward, total rewards obtained, and specific key distribution policies. The proposed method not only outperformed traditional, static key distribution methodologies but proved to scale with a massive number of Internet-of-Things devices, making it feasible for use in actual applications over intricate IoT systems. Additionally, the adaptability of the system allows it to evolve with changing circumstances, offering a higher level of protection against unidentified dangers. Overall, this system provides an adaptable and intelligent approach to securing IoT networks, and its ability to continually improve and learn makes it a more powerful solution than existing heuristic or rule-based systems. The proposed model is a promising solution for secure and scalable key distribution in IoT networks, with the aim to advance the field of adaptive security techniques for the wireless communication systems. Future work might include acceptance of the learning algorithms, improve real-time performance, and evaluate the model on larger and more diverse IoT environments.

REFRENCES

[1] Naoui, Sarra, Mohamed Elhoucine Elhdhili, and Leila Azouz Saidane. "Security analysis of existing IoT key management protocols." *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*. IEEE, 2016.

[2] Ghani, Anwar, et al. "Security and key management in IoT-based wireless sensor networks: An authentication protocol using symmetric key." *International Journal of Communication Systems* 32.16 (2019): e4139.

[3] Sowjanya, K., Mou Dasgupta, and Sangram Ray. "A lightweight key management scheme for key-escrow-free ECC-based CP-ABE for IoT healthcare systems." *Journal of Systems Architecture* 117 (2021): 102108.

[4] Messai, Mohamed-Lamine, Hamida Seba, and Makhlouf Aliouat. "A lightweight key management scheme for wireless sensor networks." *The Journal of Supercomputing* 71 (2015): 4400-4422.

[5] Chen, Dong, et al. "Lightweight key management scheme to enhance the security of internet of things." *International Journal of Wireless and Mobile Computing* 5.2 (2012): 191-198. Kumar, Vinod, Rajendra Kumar, and Santosh K. Pandey. "LKM-AMI: a lightweight key management scheme for secure two way communications between smart meters and HAN devices of AMI system in smart grid." *Peer-to-Peer Networking and Applications* 14.1 (2021): 82-100.

[6] Li, Kai, et al. "Deep Q-learning based resource management in UAV-assisted wireless powered IoT networks." *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020.

[7] Messaoud, Seifeddine, et al. "Deep federated Q-learning-based network slicing for industrial IoT." *IEEE Transactions on Industrial Informatics* 17.8 (2020): 5572-5582.

[8] Musaddiq, Arslan, Tobias Olsson, and Fredrik Ahlgren. "Reinforcement-Learning-Based Routing and Resource Management for Internet of Things Environments: Theoretical Perspective and Challenges." *Sensors* 23.19 (2023): 8263.

[9] Amin, Rashid, et al. "A survey on machine learning techniques for routing optimization in SDN." *IEEE Access* 9 (2021): 104582-104611.

[10] Messaoud, Seifeddine, et al. "A survey on machine learning in Internet of Things: Algorithms, strategies, and applications." *Internet of Things* 12 (2020): 100314.

[11] Al-Garadi, Mohammed Ali, et al. "A survey of machine and deep learning methods for internet of things (IoT) security." IEEE communications surveys & tutorials 22.3 (2020): 1646-1685.

[12] Bian, Jiang, et al. "Machine learning in real-time Internet of Things (IoT) systems: A survey." *IEEE Internet of Things Journal* 9.11 (2022): 8364-8386.

[13] Mohamed, Nur Nabila, et al. "Hybrid cryptographic approach for internet ofhybrid cryptographic approach for internet ofthings applications: A review." *Journal of Information and Communication Technology* 19.3 (2020): 279-319.

[14] Patil, Kavitha S., Indrajit Mandal, and C. Rangaswamy. "Hybrid and Adaptive Cryptographic-based secure authentication approach in IoT based applications using hybrid encryption." *Pervasive and Mobile Computing* 82 (2022): 101552.

[15] Haji, Saad Hikmat, and Siddeeq Y. Ameen. "Attack and anomaly detection in iot networks using machine learning techniques: A review." *Asian J. Res. Comput. Sci* 9.2 (2021): 30-46.

[16] Tyagi, Himani, and Rajendra Kumar. "Attack and anomaly detection in IoT networks using supervised machine learning approaches." *Revue d'Intelligence Artificielle* 35.1 (2021).

[17] Diro, Abebe, et al. "A comprehensive study of anomaly detection schemes in IoT networks using machine learning algorithms." *Sensors* 21.24 (2021): 8320.

[18] Xiong, Xiong, et al. "Resource allocation based on deep reinforcement learning in IoT edge computing." *IEEE Journal on Selected Areas in Communications* 38.6 (2020): 1133-1146.

[19] Gai, Keke, and Meikang Qiu. "Optimal resource allocation using reinforcement learning for IoT content-centric services." *Applied Soft Computing* 70 (2018): 12-21.

[20] Alabassby, Bahaa Faiz Noory Mohsin, Jinan Fadhil Mahdi, and Mohammed Aboud Kadhim. "Design and implementation WSN based on Raspberry Pi for medical application." IOP Conference Series: Materials Science and Engineering. Vol. 518. No. 5. IOP Publishing, 2019.

[21] Ahmed, Sadeer Rasheed, Mohammed Aboud Kadhim, and Tarek Abdulkarim. "Wireless sensor networks improvement using leach algorithm." IOP Conference Series: Materials Science and Engineering. Vol. 518. No. 5. IOP Publishing, 2019.

[22] Liu, Xiaolan, Zhijin Qin, and Yue Gao. "Resource allocation for edge computing in IoT networks via reinforcement learning." *ICC 2019-2019 IEEE international conference on communications (ICC)*. IEEE, 2019.