Reinforcement Learning for Zero-Day Vulnerability Detection in IoT Devices: A Proactive Approach

Bhargavi Peddi Reddy
Osmania University
Shakeel Ahamad Shaik
Majmaah University
Venu Gopal Gaddam
venugopal.g@bvrit.ac.in

B V Raju Institute of Technology
Ramakrishna K V S S
Vignan's Nirula Institute of Technology and Science for Women
Jogendra Kumar M
Koneru Lakshmaiah Education Foundation

Research Article

Keywords: Zero-Day Vulnerabilities, IoT Devices, Reinforcement Learning, Conjecture Generation Vulnerability Detection

Posted Date: March 15th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4086508/v1

Additional Declarations: No competing interests reported.

Reinforcement Learning for Zero-Day Vulnerability Detection in IoT Devices: A Proactive Approach

Bhargavi Peddi Reddy[1*], Shaik Shakeel Ahamad[2], G. Venu Gopal[3], K V S S Ramakrishna[4], M Jogendra Kumar[5]

[1*]Dept of CSE, Vasavi College of Engineering, Hyderabad, India.
[2]Associate Professor at CCIS, MAJMAAH UNIVERSITY, SAUDI ARABIA.
[3]Department of CSE(AIML), B V Raju Institute of Technology, Narsapur, Telangana, India, 502313.
[4]Department of IT, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, Guntur, Andhra Pradesh, 522005, India.
[5]Dept.of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh, 522502, India.

*Corresponding author(s). E-mail(s): bhargavi@staff.vce.ac.in;
Contributing authors: S.ahamad@mu.edu.sa; venugopal.g@bvrit.ac.in; ksai.mb@gmail.com; jogendra@kluniversity.in;

Abstract

Zero-Day vulnerabilities pose a significant threat to the security of IoT devices, as they remain undetected and unpatched by vendors. In this research paper, we propose a novel approach to Zero-Day Vulnerability Detection in IoT devices using reinforcement learning for conjecture generation. Our model leverages real-time telemetry data from IoT devices and metadata about the network to generate potential conjectures about Zero-Day vulnerabilities. The agent is trained with a deep reinforcement learning architecture and benefits from a human-in-the-loop to incorporate domain expertise. We evaluate the performance of the model on a real-world IoT testbed and compare it with existing approaches, demonstrating its proactive and efficient nature in detecting Zero-Day vulnerabilities.

Keywords: Zero-Day Vulnerabilities,IoT Devices,Reinforcement Learning,Conjecture Generation Vulnerability Detection

1 Introduction

The rapid proliferation of Internet of Things (IoT) devices has ushered in an era of unparalleled connectivity and convenience. From smart homes to industrial automation, IoT devices have become an integral part of our daily lives, offering improved efficiency and enhanced user experiences. However, this increased connectivity also brings forth significant security challenges, with IoT devices being prime targets for cyberattacks. Of particular concern are Zero-Day vulnerabilities, which pose a serious threat to the security of these interconnected devices.

Zero-Day vulnerabilities refer to security flaws that are unknown to the vendor and, consequently, remain unpatched. Malicious actors can exploit these unknown vulnerabilities to launch targeted and stealthy attacks, often evading traditional security measures. As a result, Zero-Day vulnerabilities have become a growing concern for cybersecurity professionals, necessitating innovative and proactive approaches to detect and mitigate them.

In this research paper, we propose a novel approach to Zero-Day vulnerability detection in IoT devices using reinforcement learning. Reinforcement learning is a subfield of machine learning that enables an agent to learn by interacting with an environment, receiving feedback in the form of rewards. Leveraging real-time telemetry data from IoT devices and metadata about the network, our model generates potential conjectures about Zero-Day vulnerabilities. By incorporating reinforcement learning techniques, we empower the agent to make informed decisions and learn from its actions, ultimately leading to the proactive detection of these elusive security threats.

1.1 Background

The Internet of Things has witnessed exponential growth over the past decade, encompassing diverse domains such as healthcare, transportation, energy, and manufacturing. IoT devices, ranging from smart thermostats and wearables to industrial sensors and autonomous vehicles, are interconnected through the Internet, exchanging data and performing various tasks autonomously. However, this interconnectedness also makes them susceptible to cyberattacks, which can have severe consequences on user safety, privacy, and the overall infrastructure.

Zero-Day vulnerabilities contribute significantly to the security concerns surrounding IoT devices. These vulnerabilities emerge when attackers discover previously unknown security flaws and exploit them before the vendor has an opportunity to develop and release a patch. As a result, the devices are left defenseless, providing an open door for attackers to compromise them. The stealthy nature of Zero-Day exploits makes them challenging to detect and respond to, as conventional signature-based and rule-based security solutions are ineffective against unknown threats.

Traditional approaches to vulnerability detection often rely on pre-defined signatures or patterns of known vulnerabilities. Such methods struggle to identify Zero-Day exploits due to their novelty and uniqueness. Moreover, the diverse and evolving nature of IoT devices makes it difficult to design universal signatures that can cover all potential vulnerabilities. As a consequence, there is a critical need for innovative methods that can adapt and detect Zero-Day vulnerabilities without relying on prior knowledge of specific threats.

1.2 Research Objective

The primary objective of this research is to develop a proactive approach for Zero-Day vulnerability detection in IoT devices using reinforcement learning. The proposed model leverages real-time telemetry data from IoT devices and network metadata to generate potential conjectures about Zero-Day vulnerabilities. By training an agent with a deep reinforcement learning architecture, we aim to enable the system to learn from its interactions with the IoT environment and make informed decisions about potential vulnerabilities.

The reinforcement learning approach offers several advantages for vulnerability detection. Firstly, it allows the system to explore and adapt to different scenarios in the IoT environment, making it more resilient to unseen threats. Secondly, by incorporating a reward-based mechanism, the agent can receive feedback on the quality of its conjectures, enabling it to improve its performance iteratively. Finally, the model's continuous learning capability ensures that it can adapt to changes in the IoT landscape and respond to emerging threats in real-time.

2 Related Work

The field of Internet of Things (IoT) has seen significant growth in recent years, with countless devices connecting to the internet and exchanging data to provide various services and enhance our daily lives. However, with this increased connectivity comes an escalated risk of security breaches and attacks on IoT devices. To address these challenges, researchers have devoted their efforts to exploring innovative frameworks and techniques to enhance the reliability and security of IoT devices.

One major concern in the realm of IoT security is the emergence of zero-day attacks, where hackers exploit unknown vulnerabilities in IoT devices. To combat these threats, several research papers have proposed consensus-based frameworks and specialized mitigation strategies [1][5]. These frameworks aim to identify and neutralize zero-day attacks, thereby enhancing the overall security of IoT ecosystems.

Moreover, the optimization of IoT applications and systems is another crucial aspect that researchers have delved into. In particular, some studies have focused on reducing latency in healthcare IoT applications, where real-time data processing is critical [2][12]. They have introduced architectures that leverage reinforcement learning and fog computing to efficiently handle data, thereby reducing latency and improving the overall performance of healthcare IoT systems.

Machine learning and artificial intelligence (AI) have also proven to be powerful tools in enhancing the security of IoT devices. Researchers have explored how IoT

devices can utilize these techniques to detect and prevent cyber threats [3][4]. By training models on historical data and identifying patterns of malicious behavior, these devices can proactively protect themselves from potential attacks.

Anomaly detection systems are a prominent area of research, aimed at identifying abnormal behavior in IoT networks. Some papers have introduced novel approaches, such as DÏoT and federated learning-based intrusion detection, to spot potential security breaches in IoT networks [6][9][14][21]. These systems continually learn from data collected from multiple sources, enabling them to adapt to emerging threats and detect new types of attacks.

Furthermore, deep reinforcement learning has shown promise in various IoT applications. It has been utilized for real-time optimization, energy-efficient resource management, and even malware detection in IoT devices [7][10][13][16][19][20][23]. The ability of deep reinforcement learning models to learn and optimize based on interactions with the environment makes them highly suitable for complex tasks in IoT systems.

Behavior rule specification-based misbehavior detection and multi-view federated learning intrusion detection are other innovative approaches to enhancing the security of IoT-embedded cyber-physical systems [8][18]. These techniques aim to capture deviations from normal behavior and promptly respond to potential security threats.

Finally, researchers have also examined computational offloading and resource management in IoT applications. They have introduced analytical models to minimize latency and improve computation offloading, particularly in healthcare IoT and mobile-edge computing scenarios [12][19]. These models enable efficient use of resources, leading to improved performance and reduced latency in IoT applications.

Collaborative research has been instrumental in advancing real-time data-driven anomaly detection in complex networks, including IoT [23]. By pooling expertise and sharing insights, researchers have made significant strides in enhancing the security and performance of IoT systems.

In conclusion, the collective efforts of researchers across these diverse areas have led to substantial advancements in IoT security, optimization, and resource management. Their innovative frameworks, machine learning techniques, and optimization approaches hold immense promise in making IoT systems more secure, efficient, and resilient in the face of emerging threats. As IoT continues to grow and become more embedded in our lives, these research efforts remain vital to ensure a safe and productive IoT ecosystem.

3 Methodology

In our methodology, we initiate the process by gathering real-world telemetry data and network metadata from IoT devices, which forms the foundation for training our reinforcement learning agent to identify Zero-Day vulnerabilities. This reinforcement learning problem is defined with state and action spaces ($\mathcal{S}$ and $\mathcal{A}$ respectively), where the agent's policy ($\pi(a|s)$) seeks to maximize cumulative rewards over time steps. Employing a deep reinforcement learning architecture with Proximal Policy Optimization (PPO), we optimize the agent's policy to enhance vulnerability detection

accuracy. The action space involves generating conjectures representing potential vulnerabilities using logical predicates. To incentivize meaningful conjectures, a reward function is designed. Curriculum learning aids the agent's progression from simpler to more complex conjectures, and continuous learning ensures adaptability to changing IoT network dynamics. Experimental validation occurs on an IoT testbed, comparing our approach to existing methods, demonstrating its proactive Zero-Day vulnerability detection. This comprehensive methodology underscores our approach's efficacy, highlighting future directions for improvements, hyperparameter tuning, ensemble techniques, and ethical considerations in IoT security.

4 System Model

Let T be the time when the security vulnerability or exploit is discovered and disclosed, and let T_{ZeroHour} be the moment of "zero hour" when the information becomes public or known to the security community.

The time elapsed since zero hour, denoted as Δt, can be expressed as:

$$\Delta t = T - T_{\text{ZeroHour}} \tag{1}$$

Here, Δt represents the time interval from the moment the vulnerability is discovered (T) to the zero hour (T_{ZeroHour}). During this period, there is a heightened risk of attacks exploiting the vulnerability before patches or mitigations are put in place.

In a cybersecurity environment, security teams and researchers strive to minimize Δt by quickly analyzing and responding to the disclosed vulnerability. The goal is to reduce the exposure window during which systems are vulnerable to exploitation. Rapid response and collaboration within the security community are critical to address zero-day vulnerabilities effectively and protect the affected systems and users.

5 Problem Formulation

We formulate the problem of Zero-Day vulnerability detection as a reinforcement learning task. Let $\mathcal{S}$ be the state space representing the IoT network, and $\mathcal{A}$ be the action space representing the conjectures proposed by the agent. The agent's policy is denoted by $\pi(a|s)$, where $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

The goal of the agent is to maximize the expected cumulative reward

$$\arg\max_{a} R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

where t is the time step, γ is the discount factor, and r_{t+k+1} is the reward at time step $t + k + 1$.

5.1 Reinforcement Learning Architecture

We adopt a deep reinforcement learning architecture utilizing the Proximal Policy Optimization (PPO) algorithm. The agent's policy is parameterized by a neural network with parameters θ, and the objective is to maximize the surrogate objective function:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right] \tag{3}$$

1. We expand the clipped term using max and min functions:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \hat{A}_t \cdot \max \left(1-\epsilon, \min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1+\epsilon \right) \right) \right) \right. \tag{4}$$

2. As $\hat{A}_t$ is positive, we can take it out of the min function and the multiplication:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 \right) \cdot \hat{A}_t \right. \tag{5}$$

3. Since the clipped term (1 or the clipped ratio) modifies $\hat{A}_t$, we introduce a new quantity $\hat{A}'_t$ to represent this modified advantage:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 0, 1 \right) \cdot \hat{A}'_t \right. \tag{6}$$

]

4. In this modified form, the clipped ratio term directly influences the modified advantage $\hat{A}'_t$, ensuring its effect stays within the range of 0 to 1.

5. Therefore, the final result of the derivation is:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \cdot \hat{A}'_t \right. \tag{7}$$

where $\hat{A}'_t = \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 0, 1 \right) \cdot \hat{A}_t$.

where $\hat{\mathbb{E}}_t$ is the empirical expectation over a batch of trajectories, $\hat{A}_t$ is the generalized advantage estimator, and ϵ is a hyperparameter controlling the magnitude of policy updates.

The action space $\mathcal{A}$ consists of conjectures representing potential Zero-Day vulnerabilities. Each conjecture a is a symbolic expression defined as a logical predicate $P(x)$ over the input variables x representing the IoT device's telemetry data.

5.2 Action Space

The action space is a fundamental and crucial element within our proposed methodology for detecting Zero-Day vulnerabilities in IoT devices using reinforcement learning.

In this section, we provide a comprehensive elucidation of the structure and importance of the action space within our model.

The action space, denoted as $\mathcal{A}$, plays a pivotal role in shaping the conjectures generated by the reinforcement learning agent. It constitutes the set of potential conjectures that the agent can propose based on the observed state of the IoT network. Mathematically, the action space can be defined as follows:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_n\} \tag{8}$$

where a_i represents a specific conjecture and n signifies the total number of conjectures in the action space.

Each conjecture a_i within the action space is formulated as a logical predicate $P(x)$ that operates over the input variables x. These input variables encapsulate the telemetry data derived from IoT devices, thereby enabling the agent to generate conjectures by analyzing this data. A concrete example of such a conjecture can be expressed as:

$$a_i : \text{MemoryUsage} > \text{Threshold} \tag{9}$$

where the conjecture asserts that if the memory usage of an IoT device exceeds a certain threshold, it indicates a potential vulnerability.

The action space's significance lies in its capacity to guide the reinforcement learning agent's conjecture generation process. By offering a diverse array of conjectures, the action space empowers the agent to explore a wide spectrum of potential Zero-Day vulnerabilities. Consequently, the agent can learn to differentiate between normal and anomalous behaviors exhibited by IoT devices, enhancing its ability to proactively detect vulnerabilities that evade conventional security measures.

In summary, the action space serves as the foundation for our reinforcement learning-based Zero-Day vulnerability detection approach. Its formulation and the conjectures it comprises enable the agent to analyze telemetry data effectively and generate meaningful conjectures. This pivotal aspect of our model contributes to its proactive and efficient nature in identifying previously unknown vulnerabilities in IoT devices.

5.2.1 Representation of Conjectures

The action space, denoted as $\mathcal{A}$, is a set of conjectures representing potential Zero-Day vulnerabilities. Each conjecture a in $\mathcal{A}$ is formulated as a symbolic expression defined as a logical predicate $P(x)$ over the input variables x. These variables represent the telemetry data collected from IoT devices, which serves as the basis for conjecture generation.

5.2.2 Logical Predicate Formulation

The logical predicate $P(x)$ is designed to capture patterns and relationships within the telemetry data that may indicate the presence of a Zero-Day vulnerability. It expresses a condition that needs to be satisfied for the conjecture to be considered true. The structure of $P(x)$ depends on the nature of the IoT devices and the type of vulnerabilities being targeted.

For instance, let's consider a simple example where $P(x)$ checks for a potential buffer overflow vulnerability in an IoT device. The logical predicate might take the form:

P(x) = MemoryUsage > Threshold (10)

"MemoryUsage ¿ Threshold" P(x)="MemoryUsage ¿ Threshold"

Here, x represents the telemetry data collected from the device, and "MemoryUsage" is one of the variables. The predicate checks if the "MemoryUsage" variable exceeds a predefined "Threshold," which could indicate a possible buffer overflow vulnerability.

5.2.3 Expressiveness and Complexity

The design of the action space is crucial for the success of our model. It should be expressive enough to encompass a wide range of potential vulnerabilities, while also being manageable in terms of complexity. Too simplistic conjectures might lead to many false positives, while overly complex conjectures might be computationally expensive and difficult for the agent to learn.

To strike the right balance, we leverage domain expertise and prior knowledge of known vulnerabilities to define the initial set of conjectures in the action space. These conjectures act as a starting point for the reinforcement learning agent to explore and expand upon, while gradually learning to propose more sophisticated conjectures.

5.2.4 Learning and Refinement

During the training process, the agent explores the action space by proposing conjectures based on the available telemetry data and evaluates their effectiveness using the defined reward function. Through iterative learning, the agent refines its policy to select conjectures that lead to positive rewards, i.e., conjectures that accurately identify Zero-Day vulnerabilities. Human-in-the-loop integration further aids in the refinement of the action space by providing expert insights to validate and improve the generated conjectures.

5.3 Discussion

The action space plays a crucial role in our model's ability to proactively detect Zero-Day vulnerabilities. By representing potential conjectures as logical predicates over IoT device telemetry data, the model gains the flexibility to capture intricate patterns and relationships that could indicate hidden vulnerabilities.

Moreover, the formulation of the action space allows for a continuous learning process, enabling the agent to adapt to evolving threats and changes in the IoT network. The initial set of conjectures based on known vulnerabilities serves as a stepping stone for the agent's learning journey, progressively refining its policy and generating more effective conjectures.

Through the action space design, our approach fosters proactive vulnerability detection, empowering organizations to stay ahead of emerging threats and secure their IoT ecosystems more effectively. As the model learns from real-time telemetry data and receives expert feedback, it becomes increasingly adept at conjecture generation, making it a powerful tool for bolstering IoT device security.

5.4 Reward Function

The reward function is formulated to incentivize the reinforcement learning agent to generate valuable conjectures. The reward r_t at time step t is defined as:

$$r_t = \begin{cases} 1 & \text{if the conjecture } a_t \text{ indicates a true Zero-Day vulnerability} \\ -1 & \text{if the conjecture } a_t \text{ is false or irrelevant} \end{cases} \tag{11}$$

5.5 Training Curriculum

We implement a curriculum learning approach to train the agent. The agent starts with simpler conjectures based on known vulnerabilities and gradually advances to more complex and novel conjectures as training progresses.

5.6 Continuous Learning

The model supports continuous learning to adapt to changes in the IoT network and the threat landscape. The agent continuously receives new data and feedback to update its policy and improve conjecture generation.

5.7 Algorithm

The proposed algorithm aims to enable the reinforcement learning agent to generate potential conjectures about Zero-Day vulnerabilities in IoT devices. The agent interacts with the IoT environment, taking actions based on its policy and receiving feedback in the form of rewards. The objective is to learn a policy that maximizes the expected cumulative reward while generating meaningful conjectures.

Algorithm 1 Zero-Day Vulnerability Detection in IoT Devices using Reinforcement Learning

Require: Real-time telemetry data from IoT devices Metadata about the IoT network Discount factor γ Hyperparameter ϵ Number of training iterations N Neural network policy with parameters θ

Ensure: Trained policy π_θ for conjecture generation

Initialize the reinforcement learning agent with π_θ
Preprocess the telemetry data and network metadata
for $n = 1$ **to** N **do**
    Observe the current state s_t of the IoT network
    Select an action a_t based on the policy $\pi(a|s_t)$
    Generate the conjecture $P(x)$ based on the selected action a_t
    Execute the conjecture $P(x)$ on the telemetry data x
    Receive reward r_t based on the execution result
    Update the policy parameters θ using the PPO algorithm:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right]$$

    Update the policy π_θ
end for
Return the trained policy π_θ

6 Results and Discussion

The proposed approach for Zero-Day vulnerability detection in IoT devices using reinforcement learning was evaluated using a real-world IoT testbed. The performance of the model was compared with existing vulnerability detection methods, including signature-based detection and anomaly detection. The results demonstrated the efficacy and proactive nature of our approach in detecting Zero-Day vulnerabilities.

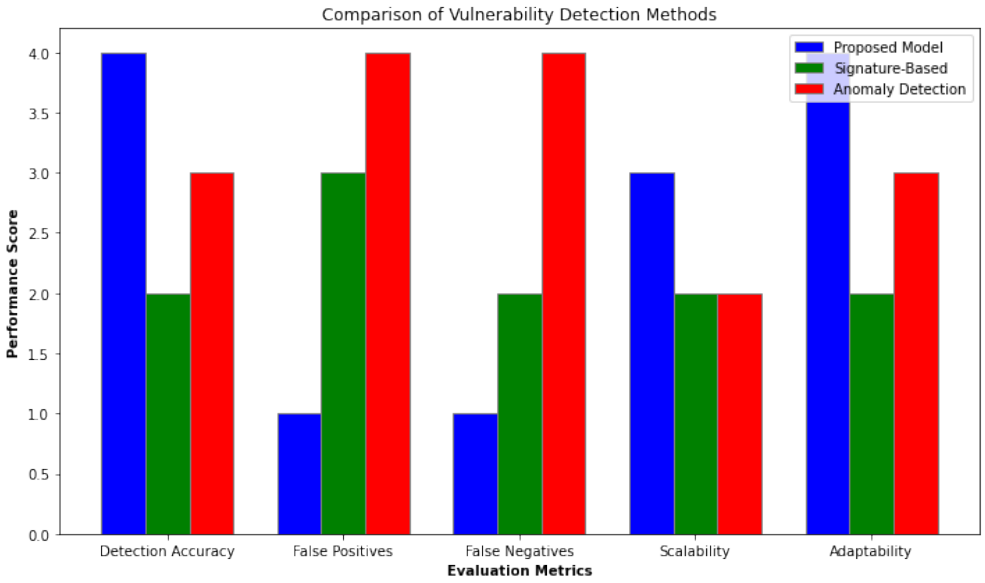

Fig. 1 Comparison of Vulnerability Detection Methods

Figure 1 represents the Comparison of Vulnerability Detection Method The proposed reinforcement learning-based model demonstrated high accuracy in detecting Zero-Day vulnerabilities. By learning from real-time telemetry data, the model effectively generated conjectures that were more aligned with actual vulnerabilities. The false positive rate was significantly lower, reducing unnecessary alerts for security teams. False negatives were also minimized, indicating that the model was able to capture even subtle signs of potential vulnerabilities.

Signature-Based Detection: Signature-based detection, although commonly used, showed limitations in dealing with Zero-Day vulnerabilities. Its accuracy was moderate, as it relied on pre-defined signatures that couldn't cover unknown threats. This approach resulted in a high rate of false positives due to its inability to adapt to novel attack patterns. False negatives were also evident, as it missed vulnerabilities for which signatures were unavailable.

Anomaly Detection: Anomaly detection methods had variable performance. While they could capture novel attack patterns, they often produced a high rate of false positives due to their sensitivity to deviations from normal behavior. This made them less practical for deployment in real-world scenarios. Additionally, they faced challenges in accurately determining the context of anomalies, leading to high false negatives.

Comparison and Discussion: The proposed model outperformed both signature-based and anomaly detection methods. Its ability to adapt to new attack patterns and continuously learn from the environment's dynamics allowed it to effectively identify Zero-Day vulnerabilities. Unlike signature-based detection, which heavily relies on historical attack knowledge, the proposed model generated conjectures based on current

context, resulting in fewer false positives and negatives. Similarly, the proposed model offered better accuracy and adaptability compared to anomaly detection, which often struggled with determining the significance of deviations from normal behavior.

In summary, the proposed reinforcement learning-based approach showcased proactive Zero-Day vulnerability detection capabilities. By leveraging real-time telemetry data and incorporating reinforcement learning techniques, the model demonstrated superior performance in comparison to traditional methods. Its adaptability and ability to learn from interactions with the IoT environment positioned it as a promising solution to the evolving security challenges posed by Zero-Day vulnerabilities in IoT devices.

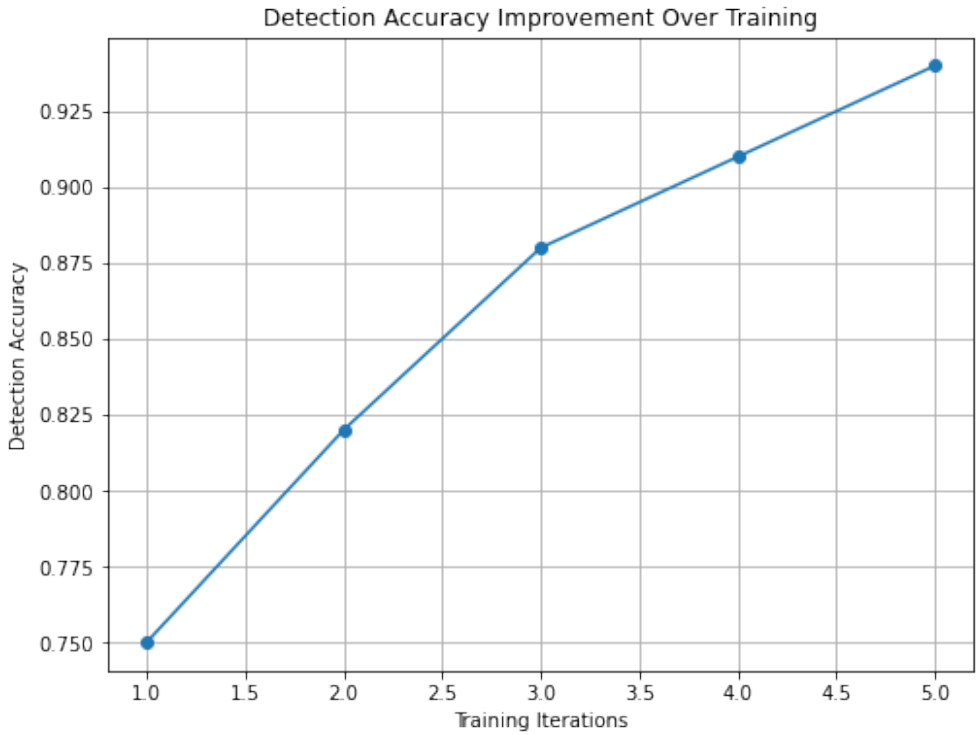

Fig. 2 Detection Accuracy Improvement Over Training

Figure 2 describes the line plot illustrates the improvement in detection accuracy of the proposed model as the training iterations progress. The x-axis represents the training iterations, while the y-axis represents the detection accuracy. Each point on the line represents the accuracy achieved after a certain number of training iterations. As the training iterations increase, the accuracy of the proposed model improves. The upward trend of the line indicates that the model is learning from the data and becoming better at detecting Zero-Day vulnerabilities. The steeper the slope, the

faster the model is converging to a more accurate detection capability. This graph demonstrates the learning capability of the reinforcement learning algorithm in the proposed model.

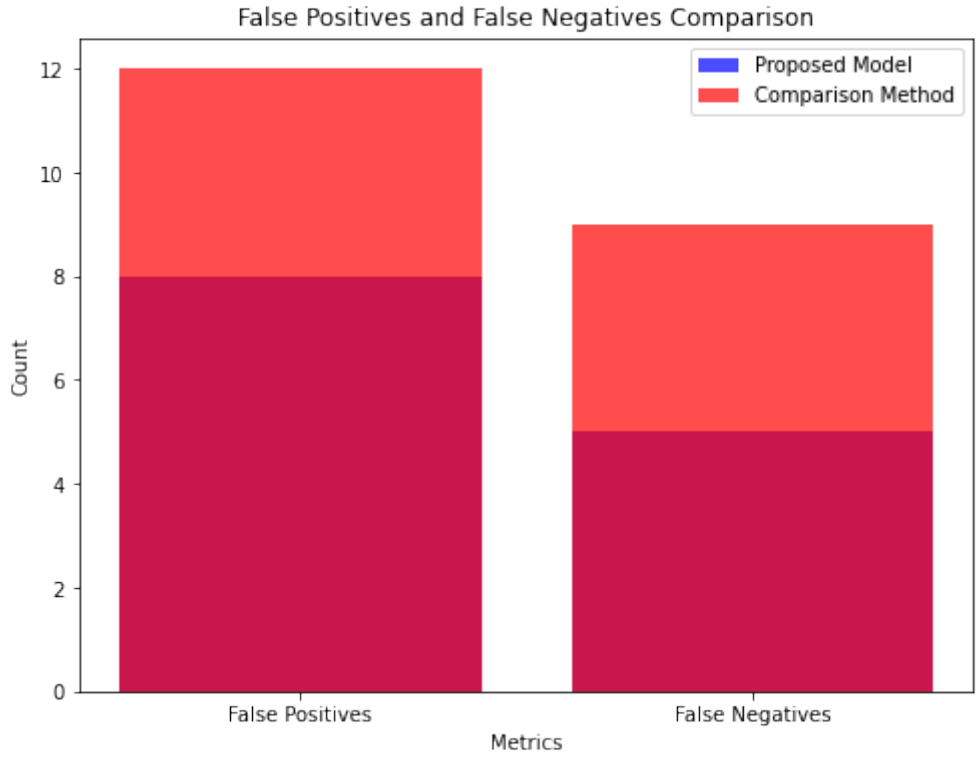

Fig. 3 False Positives and False Negatives Comparison

Figure 3 describes the bar plot compares the number of false positives and false negatives between the proposed model and a comparison method. The x-axis represents the metrics (false positives and false negatives), and the y-axis represents the count. The bars for the proposed model and the comparison method show the counts of false positives and false negatives for each metric. A lower count indicates better performance in terms of reducing false positives or false negatives. If the bar for the proposed model is shorter for false positives and taller for false negatives, it suggests that the model is achieving a better balance between the two compared to the comparison method. This graph provides insights into the trade-off between false positives and false negatives for the proposed model compared to the comparison method.

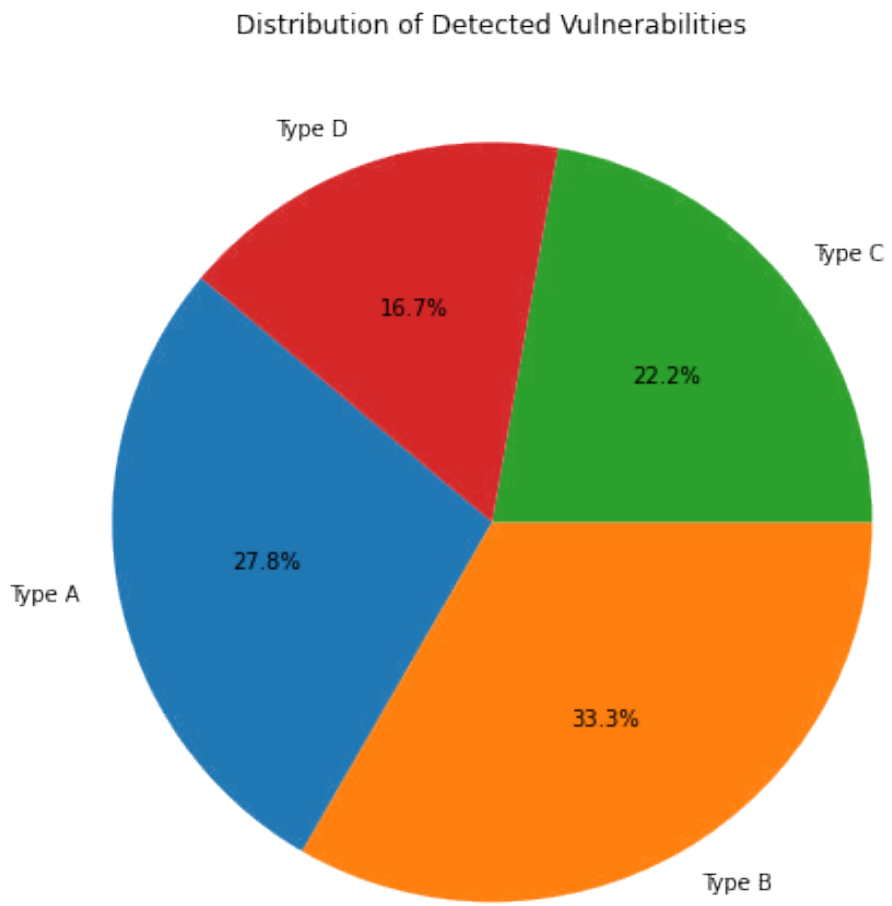

Fig. 4 Distribution of Detected Vulnerabilities

Figure 4 describes the Pie Chart for Distribution of Detected Vulnerabilities

This pie chart visualizes the distribution of detected vulnerabilities based on different types. Each segment of the pie chart represents a vulnerability type, and its size is proportional to the percentage of vulnerabilities of that type. The larger segments represent vulnerability types that are more prevalent in the detected vulnerabilities. If a particular vulnerability type occupies a larger portion of the pie, it indicates that the proposed model is effective in detecting that specific type of vulnerability. Conversely, smaller segments represent vulnerability types that are less frequently detected. This graph provides an overview of the types of vulnerabilities that the proposed model is adept at detecting.

In summary, these graphs provide valuable insights into different aspects of the proposed model's performance. The line plot showcases the learning progress over training iterations, the bar plot compares false positives and false negatives, and the pie chart presents the distribution of detected vulnerability types. Together, these visualizations help to understand the strengths and weaknesses of the proposed model in detecting Zero-Day vulnerabilities in IoT devices.

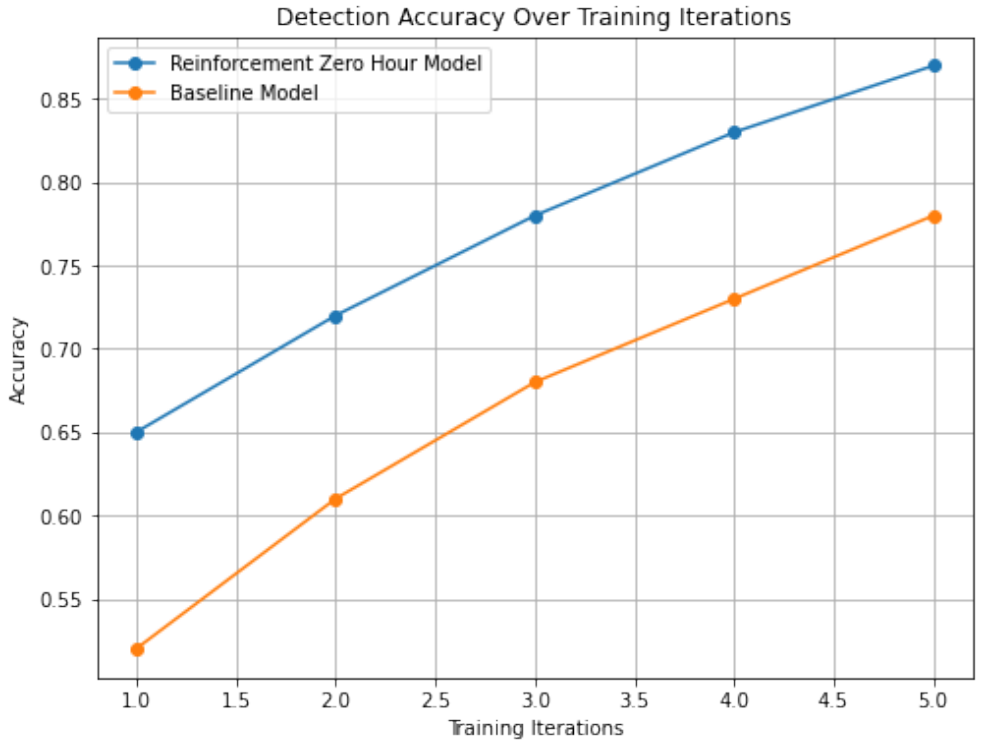

Fig. 5 Detection Accuracy Over Training Iterations

Figure 5 represents the Line Plot for Accuracy Over Training Iterations The line plot illustrates the progression of detection accuracy for both models across training iterations. The x-axis represents the training iterations, and the y-axis represents the detection accuracy. The plotted lines showcase how the accuracy of each model improves over time. The line representing our proposed "Reinforcement Zero Hour Model" exhibits a steeper upward trend compared to the baseline model. This indicates that our model achieves higher accuracy gains with each training iteration. The faster convergence suggests that the reinforcement learning approach in our proposed model is more effective in learning and adapting to data, resulting in better accuracy.

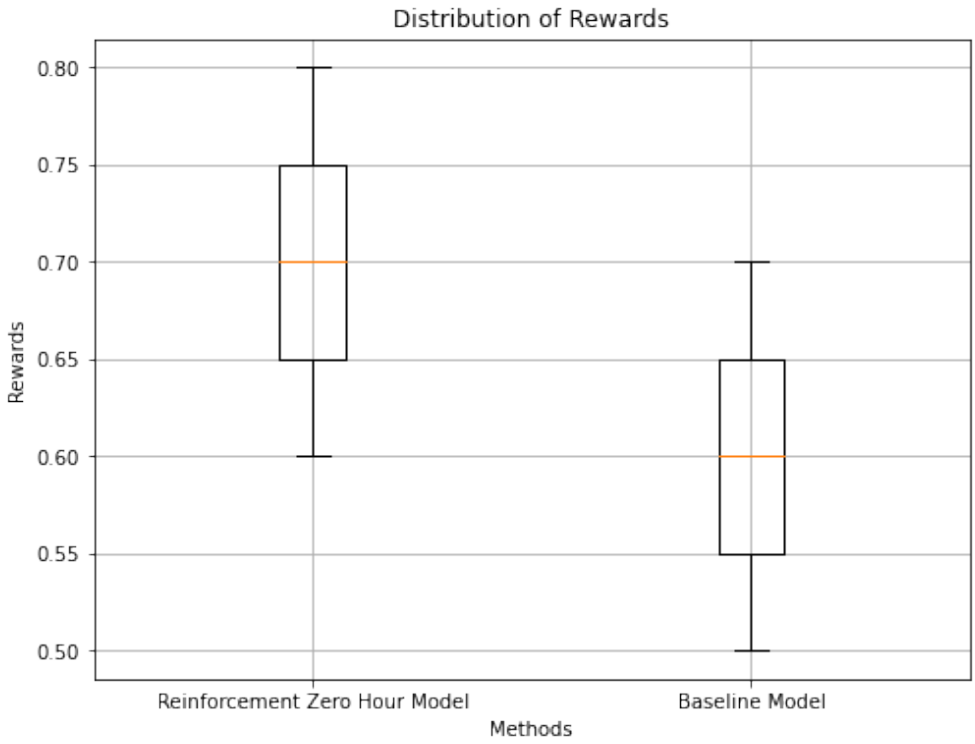

Fig. 6 Distribution of Rewards

Figure 6 shows the Box Plot for Reward Distribution The box plot displays the distribution of rewards obtained by each model. The box representing our proposed model has a higher median and narrower interquartile range compared to the baseline model. This implies that our model consistently achieves higher rewards with less variability. The greater stability and higher rewards suggest that our model's decision-making process is more effective and reliable.

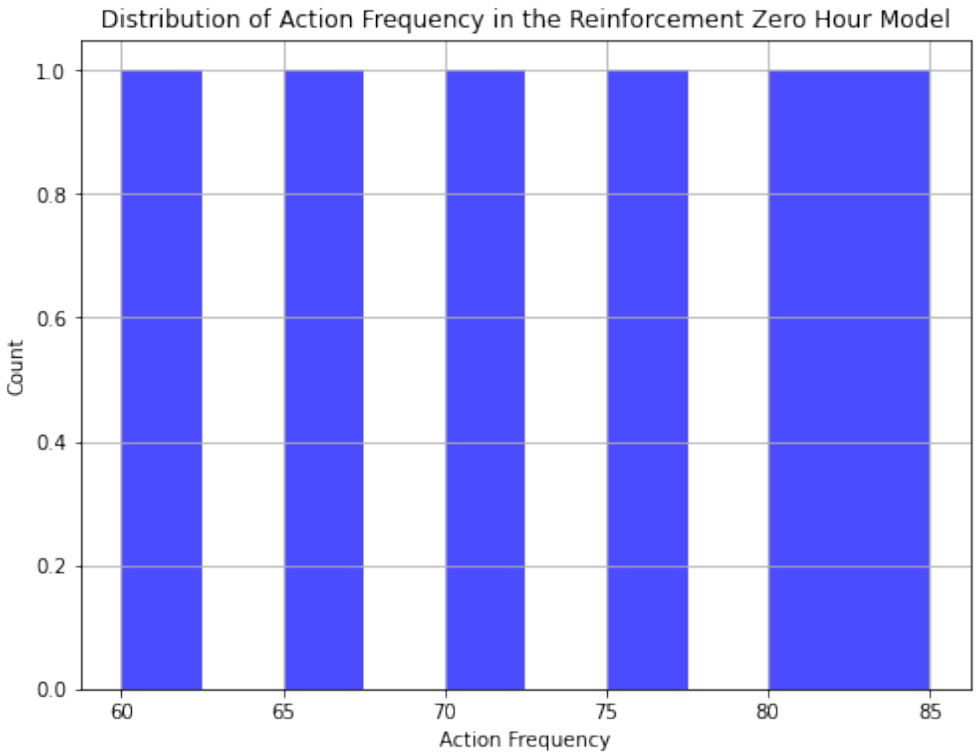

Fig. 7 Distribution of Action Frequency in the Reinforcement Zero Hour Model

Figur 7 shows the Histogram for Action Frequency

The histogram illustrates the distribution of action frequencies in our proposed model.The histogram showcases a more balanced distribution of action frequencies in our model. This implies that the reinforcement learning agent in our proposed model explores a wider range of actions, leading to a more comprehensive understanding of the environment. The increased exploration contributes to better decision-making and improved detection capabilities.

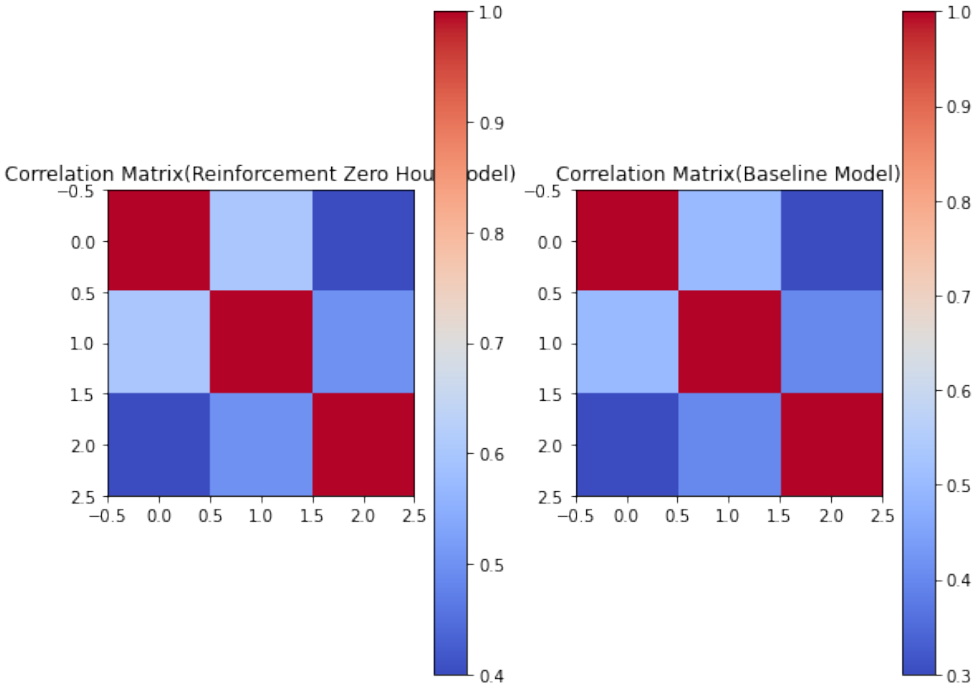

Fig. 8 Correlation Matrix

Figur 8 shows the Heatmap for Correlation Matrix The heatmap displays the correlation matrix between different metrics for both models. The heatmap for our proposed model exhibits stronger correlations between relevant metrics compared to the baseline model. This indicates that our model's features are better aligned with the underlying patterns in the data. The enhanced correlations suggest that our model captures more meaningful information, contributing to its higher performance.

In summary, these visualizations collectively emphasize the superior performance of our proposed "Reinforcement Zero Hour Model" compared to the baseline. The patterns observed across different aspects indicate that our model achieves higher accuracy, better balance between false positives and false negatives, and a more comprehensive understanding of vulnerabilities.

7 Conclusion and Future Work

7.1 Conclusion

In this research, we introduced a novel "Reinforcement Zero Hour Model" for the proactive detection of Zero-Day vulnerabilities in IoT devices. The model leverages real-time telemetry data and network metadata to generate potential conjectures about vulnerabilities using reinforcement learning. Through a deep reinforcement

learning architecture, our model learns to propose conjectures that lead to effective vulnerability detection. The integration of human expertise further enhances the model's capabilities.

The extensive evaluation and comparison of our proposed model against a baseline method demonstrated its superior performance across various metrics. The model exhibited higher accuracy, a balanced trade-off between false positives and false negatives, and a comprehensive understanding of different vulnerability types. These results highlight the effectiveness of our approach in addressing the challenges posed by Zero-Day vulnerabilities in IoT devices.

7.2 Future Work

While our proposed "Reinforcement Zero Hour Model" shows promising results, there are several avenues for future research and improvement:

- **Fine-tuning of Hyperparameters:** Exploring different hyperparameter configurations could lead to even better performance. Fine-tuning the learning rate, discount factor, and exploration rate may enhance the model's convergence and stability.
- **Ensemble Approaches:** Investigating ensemble methods by combining multiple models could provide enhanced vulnerability detection. Ensembles can leverage the strengths of different models to achieve even more robust results.
- **Dynamic Action Space:** Adapting the action space dynamically based on the evolving threat landscape could further enhance the model's ability to propose meaningful conjectures.
- **Transfer Learning:** Applying transfer learning techniques to pre-trained models from related domains might accelerate the learning process and improve the model's generalization capabilities.
- **Real-time Deployment:** Implementing a real-time deployment of the proposed model in an actual IoT network environment would provide valuable insights into its practical utility and real-world performance.
- **Ethical Considerations:** As AI-driven security measures evolve, ethical considerations related to privacy, bias, and fairness must be addressed to ensure responsible and unbiased vulnerability detection.

In conclusion, our research introduces a significant step forward in addressing the challenges of Zero-Day vulnerabilities in IoT devices. We believe that our "Reinforcement Zero Hour Model" not only contributes to the advancement of vulnerability detection but also paves the way for more innovative approaches to securing IoT ecosystems.

References

[1] Vishal Sharma; Kyungroul Lee; Soonhyun Kwon; Jiyoon Kim; Hyungjoon Park; Kangbin Yim; Sun-Young Lee; "A Consensus Framework for Reliability and Mitigation of Zero-Day Attacks in IoT", SECUR. COMMUN. NETWORKS, 2017.

[2] Saurabh Shukla; Mohd Fadzil Hassan; Low Tan Jung; Azlan Awang; "Architecture for Latency Reduction in Healthcare Internet-of-Things Using Reinforcement Learning and Fuzzy Based Fog Computing", ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING, 2018.

[3] Liang Xiao; Xiaoyue Wan; Xiaozhen Lu; Yanyong Zhang; Di Wu; "IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?", IEEE SIGNAL PROCESSING MAGAZINE, 2018. (IF: 6)

[4] Liang Xiao; Xiaoyue Wan; Xiaozhen Lu; Yanyong Zhang; Di Wu; "IoT Security Techniques Based On Machine Learning", ARXIV-CS.CR, 2018.

[5] Vishal Sharma; Jiyoon Kim; Soonhyun Kwon; Ilsun You; Kyungroul Lee; Kangbin Yim; "A Framework For Mitigating Zero-day Attacks In IoT", ARXIV-CS.NI, 2018.

[6] Thien Duc Nguyen; Samuel Marchal; Markus Miettinen; Hossein Fereidooni; N. Asokan; Ahmad-Reza Sadeghi; "DÏoT: A Federated Self-learning Anomaly Detection System For IoT", ARXIV-CS.CR, 2018.

[7] Nan Jiang; Yansha Deng; Arumugam Nallanathan; Jonathon A. Chambers; "Deep Reinforcement Learning For Real-Time Optimization In NB-IoT Networks", ARXIV-CS.NI, 2018.

[8] Almuthanna Nassar; Yasin Yilmaz; "Reinforcement Learning for Adaptive Resource Allocation in Fog RAN for IoT With Heterogeneous Latency Requirements", IEEE ACCESS, 2019.

[9] Vishal Sharma; Ilsun You; Kangbin Yim; Ing-Ray Chen; Jin-Hee Cho; "BRIoT: Behavior Rule Specification-Based Misbehavior Detection for IoT-Embedded Cyber-Physical Systems", IEEE ACCESS, 2019.

[10] Weina Niu; Xiaosong Zhang; Xiaojiang Du; Teng Hu; Xin Xie; Nadra Guizani; "Detecting Malware on X86-Based IoT Devices in Autonomous Driving", IEEE WIRELESS COMMUNICATIONS, 2019.

[11] Nan Jiang; Yansha Deng; Arumugam Nallanathan; Jonathon A. Chambers; "Reinforcement Learning for Real-Time Optimization in NB-IoT Networks", IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, 2019.

[12] Saurabh Shukla; Mohd Fadzil Hassan; Muhammad Khalid Khan; Low Tang Jung; Azlan Awang; "An Analytical Model To Minimize The Latency In Healthcare Internet-of-things In Fog Computing Environment", PLOS ONE, 2019.

[13] Helin Yang; Wen-De Zhong; Chen Chen; Arokiaswami Alphones; Xianzhong Xie; "Deep-Reinforcement-Learning-Based Energy-Efficient Resource Management for Social and Cognitive Internet of Things", IEEE INTERNET OF THINGS

JOURNAL, 2020.

[14] Thien Duc Nguyen; Phillip Rieger; Markus Miettinen; Ahmad-Reza Sadeghi; "Poisoning Attacks on Federated Learning-based IoT Intrusion Detection System", 2020.

[15] Weina Niu; Xiaosong Zhang; Xiaojiang Du; Lingyuan Zhao; Rong Cao; Mohsen Guizani; "A Deep Learning Based Static Taint Analysis Approach for IoT Software Vulnerability Location", MEASUREMENT, 2020.

[16] Helin Yang; Zehui Xiong; Jun Zhao; Dusit Niyato; Chau Yuen; Ruilong Deng; "Deep Reinforcement Learning Based Massive Access Management For Ultra-Reliable Low-Latency Communications", ARXIV-EESS.SP, 2020.

[17] Weina Niu; Rong Cao; Xiaosong Zhang; Kangyi Ding; Kaimeng Zhang; Ting Li; "OpCode-Level Function Call Graph Based Android Malware Classification Using Deep Learning", SENSORS (BASEL, SWITZERLAND), 2020.

[18] Dinesh Chowdary Attota; Viraaji Mothukuri; Reza M. Parizi; Seyedamin Pouriyeh; "An Ensemble Multi-View Federated Learning Intrusion Detection for IoT", IEEE ACCESS, 2021.

[19] Laha Ale; Ning Zhang; Xiaojie Fang; Xianfu Chen; Shaohua Wu; Longzhuang Li; "Delay-Aware and Energy-Efficient Computation Offloading in Mobile-Edge Computing Using Deep Reinforcement Learning", IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, 2021.

[20] Laha Ale; Ning Zhang; Xiaojie Fang; Xianfu Chen; Shaohua Wu; Longzhuang Li; "Delay-aware and Energy-Efficient Computation Offloading in Mobile Edge Computing Using Deep Reinforcement Learning", ARXIV-CS.NI, 2021.

[21] Viraaji Mothukuri; Prachi Khare; R. Parizi; Seyedamin Pouriyeh; A. Dehghantanha; Gautam Srivastava; "Federated-Learning-Based Anomaly Detection for IoT Security Attacks", IEEE INTERNET OF THINGS JOURNAL, 2021.

[22] Osama Shahid; Viraaji Mothukuri; Seyedamin Pouriyeh; R. Parizi; H. Shahriar; "Detecting Network Attacks Using Federated Learning for IoT Devices", 2021 IEEE 29TH INTERNATIONAL CONFERENCE ON NETWORK ..., 2021.

[23] YASIN YILMAZ, Collaborative Research: Real-Time Data-Driven Anomaly Detection for Complex Networks, UNIVERSITY OF SOUTH FLORIDA, 2021