The Empowerment and Countermeasures of Artificial Intelligence in Internet of Things Security

Hangfan Lu *

Faculty of Data Science, City University of Macau, Macao, China

* Corresponding Author Email: 1943193042@qq.com

Abstract. Amid the backdrop of a surging number of Internet of Things (IoT) devices, a rapidly expanding attack surface, and the extension of security threats from the virtual realm to the physical world, traditional security mechanisms are confronted with severe challenges of insufficient scalability, real-time performance, and adaptability. Concurrently, the rapid development of artificial intelligence (AI) technology, particularly its advantages in pattern recognition, anomaly detection, and automated decision-making, offers new empowerment paths for IoT security protection. This research systematically reviews the key applications of AI in IoT security, including intelligent intrusion detection based on deep learning, continuous identity authentication based on behavior and radio frequency fingerprints, and privacy-enhanced threat detection relying on federated learning. These technologies have significantly improved the accuracy of threat identification, the real-time response, and the feasibility of cross-device collaborative defense. However, the research also delves into the new threat of "aggressive AI" formed by the exploitation of AI technology by attackers, including evading detection through generative adversarial networks, automated penetration based on reinforcement learning, and direct adversarial attacks against AI models themselves. This reveals that while AI enhances defense capabilities, it also introduces new vulnerabilities and adversarial dimensions. The conclusion of this study emphasizes that AI plays a dual role in IoT security, both empowering and attacking, driving the security game into a new stage of rapid escalation.

Keywords: Intelligent defense, Adversarial challenges, Internet of Things.

1. Introduction

The security of the Internet of Things is facing an unprecedentedly severe situation. The root cause lies in the fundamental contradiction between the technical characteristics of the IoT ecosystem itself and its large-scale popularization. This has led to a sharp expansion of systemic risks. The number of globally connected IoT devices has exceeded one billion, ranging from smart cameras to industrial sensors. Each device could potentially serve as a jumping board for attackers to invade, forming an attack plane with an extremely large exposure. Malware can convert a large number of devices into controlled "botnets" to launch large-scale distributed denial-of-service attacks on network infrastructure, which have caused multiple large-scale network disconnection incidents in history. Data leakage incidents are also frequent, from private conversations in households to production data in factories, all of which may be exposed due to device breaches. The most worrying aspect is that as the IoT integrates more deeply with the physical world, security threats have extended from the virtual space to the real world. Tampered smart vehicle control systems, maliciously manipulated medical equipment or industrial robots have posed direct physical threats to personal safety, critical infrastructure and public security. Attack methods have also become increasingly specialized and automated, forming a complete underground industrial chain from vulnerability discovery, weaponization development to attack implementation.

In the meantime, the technology of artificial intelligence, especially machine learning and deep learning, has advanced greatly in the last few years. The fundamental benefits that it has of pattern recognition, anomaly detection, and automated decision-making offer new opportunities in tackling them. The ability of AI to adapt to normal and abnormal patterns by learning them automatically using the massive and multi-dimensional data in the Internet of Things allows detecting various unknown threats; its automation allows responding to security-related incidents in real-time and

managing a large-scale, dynamic, and complex threat environment, which is why its limitations are compensated compared to traditional security systems in terms of scalability and adaptability.

However, technology often has a double-edged effect. The main purpose of this article is not only to systematically review how artificial intelligence, as an enabling technology, enhances the security protection level of the Internet of Things, but also to deeply analyze how artificial intelligence itself is exploited by attackers and the new vulnerabilities it introduces, thereby forming a new type of confrontation in the field of the Internet of Things. Based on this, this article will focus on the application, risks, and games of artificial intelligence in the network and data security aspects of the Internet of Things, and will also cover the closely related edge computing security and some physical security impacts. First, it will explain how AI empowers the security of the Internet of Things, then analyze the confrontation and risks it triggers, and finally summarize the open challenges and future directions.

2. Artificial Intelligence and Internet of Things Security

2.1. Intelligent Intrusion Detection

In the field of intelligent intrusion detection, numerous scholars have conducted systematic research on abnormal behaviors, malicious code, and multi-source data fusion in the context of the Internet of Things, and have verified the practical advantages of AI models. The anomaly detection direction generally adopts device behavior modeling based on traffic. Zhang et al. used LSTM to model the outbound traffic of IoT gateways packet by packet, solving the problem of delayed alerts of traditional threshold rules in low-speed pulse DDoS scenarios. The experiment showed that the detection delay was shortened to 2 seconds, and the false alarm rate was lower than 0.3%, significantly superior to the baseline scheme based on Snort [1]. Kumar and Park deployed a lightweight Autoencoder on resource-constrained industrial gateways specifically for abnormal identification of botnet traffic. They increased the detection rate to 98.4% through a 1.2 MB compressed model, while maintaining CPU usage below 12%, proving the feasibility of deep generative models for real-time detection at the edge. For the malicious code characteristics in firmware and network traffic, lightweight models at the edge have become the mainstream [2]. The Autoencoder scheme does not require uploading the original firmware to complete scanning locally, which not only meets real-time requirements but also avoids privacy leakage. Finally, multi-source data fusion analysis integrates network traffic, device logs, and sensor time series into an "end-user-environment" correlation graph through graph neural network and other algorithms. Chen et al. introduced graph neural networks into threat hunting, integrating network traffic, device logs, and sensor time series data into a three-dimensional "end-user-environment" correlation graph, successfully increasing the detection accuracy of lateral movement attacks from 81% to 96%, and being able to locate the compromised node within 15 seconds, overcoming the limitations of a limited single data source perspective [3]. The research collectively indicates that anomaly detection based on AI, malware identification, and multi-source fusion analysis have become the core technical paths for enhancing the security protection efficiency of the Internet of Things.

2.2. Intelligent Identity Authentication

The traditional IoT identity authentication and access control mechanisms have inherent vulnerabilities such as being vulnerable to theft or bypassing. This makes traditional identity authentication difficult to cope with the dynamic and complex IoT environment. The introduction of artificial intelligence, especially machine learning and reinforcement learning technologies, has made Internet identity authentication contextual and adaptive.

At the identity authentication level, AI has achieved a leap from single, explicit authentication to continuous, implicit authentication. By analyzing the inherent behavioral biometric features of IoT devices or users, as well as the communication patterns of the devices (such as the timing of data packet transmission, protocol interaction sequences), resource usage patterns (such as periodic

fluctuations in CPU/ memory), and even the physical layer signal characteristics (such as radio frequency fingerprints based on channel state information CSI), the system can establish a unique identity authentication baseline. Lightweight AI models deployed at gateways or at the edge can perform real-time comparisons, enabling silent and continuous monitoring of abnormal identities (such as device hijacking, simulation attacks), significantly increasing the difficulty of identity fraud. Marco Ferretti solved the gap of being unable to continuously confirm the identity after login by continuously capturing user interactions with IoT devices and using machine learning to establish a behavioral fingerprint baseline, enabling the system to silently re-authenticate in the background and block session hijacking and identity fraud [4]. Lizhe et al. proposed to construct radio frequency fingerprints using channel state information CSI and combined it with a lightweight BP neural network for real-time comparison at the gateway side, solving the problem of "fake access to the power distribution network" for power wireless terminals, and increasing the illegal device identification rate to over 96.5% [5].

At the access control level, AI is driving the strategy to shift from static rules to dynamic intelligence. The traditional single allows or denies decision has been replaced by an adaptive permission adjustment mechanism based on real-time context risk assessment. For example, the system can use reinforcement learning algorithms to build an intelligent policy engine. This engine will comprehensively analyze multiple real-time information to make decisions, including the access request itself, the current behavior risk score of the device provided by the continuous authentication module, the overall network situation, and specific context information such as time and location. Through continuous interaction with the environment for learning, the engine can automatically determine and dynamically adjust the permission level for each access session. For example, it will decide whether to upgrade the permission from full access to a higher level, maintain the status quo, or downgrade to only read permission or even initiate temporary isolation based on the risk level. This mechanism enables the access control system to not only respond to internal threats and ongoing attacks in real time but also strictly implement the principle of minimum access while ensuring the continuity of core business.

2.3. Privacy Enhancement and Data Security

As a distributed machine learning paradigm, Federated learning has tremendous promise in the privacy-sensitive community of Internet of Things security (IoT). The essence of it is that the various devices or data owners can collectively train a quality global threat model and do so without submitting the underlying local data (network traffic packets or device logs) to a central repository. This aspect is a direct response to the two-way problem of data silos and privacy laws in the Internet of Things. In 2022, Nguyen et al. proposed a federated learning-based intrusion detection system. Such a system trains a recurrent neural network (RNN) model on the local gateway and simply uploads the model parameters to the central server where they can be aggregated and this way efficient detection of network anomalies can be made without returning the original data of the traffic [6]. Also, Wang et al. proved that a feedforward artificial neural network (ANN) can be employed as the foundation model to enhance the results of detection for use in the federated learning setup [7].

On the system architecture level, Wang suggested a federated learning platform (BCS-FL) by using blockchain to support the industrial internet of things (IIoT) scenario to solve the following problems: single-point server failures, data quality variation, and malicious attacks in conventional federated learning [8]. This model will integrate the differentiation privacy technology and introduce noise prior to uploading the model, which will boost the protection ability of sensitive data. Besides, studies of 5G advanced edge computing networks show that by integrating federated learning with hybrid deep learning models (like AE-CNN-BiLSTM), high accuracy (AUC 99.59) at low latency (approximately 0.0476 milliseconds per sample) may be attained in intrusion detection within a gateway or a multi-access edge computing (MEC) node, which is appropriate in large-scale and distributed network security systems. Federated learning as an application to IoT threat detection has ceased its theoretical discussions to become a reality. It can reach detection performance that is similar

to centralized model and it guarantees data privacy which offers a viable direction to creating cross-domain, cross-organizational wide-partnership security security systems.

3. Internet of Things Security Countermeasures

3.1. Offensive AI

In the dimension of Internet of Things security confrontation, artificial intelligence is being transformed by attackers into an efficient attack weapon, significantly enhancing the automation, precision, and concealment of attacks. Firstly, in the generation of attack payloads, attackers can utilize deep learning technologies such as generative adversarial networks to automatically generate malicious software variants targeting specific firmware vulnerabilities or communication protocol flaws of IoT devices. More importantly, AI can autonomously generate adversarial samples, such as subtly perturbing malicious network traffic or code, to "evade" machine learning-based intrusion detection systems in the feature space, achieving escape attacks. This strategy poses a risk of failure for traditional and even some intelligent security defenses. Secondly, in the field of social engineering attacks, the emergence of generative artificial intelligence, especially large language models, has greatly reduced the threshold for large-scale and highly credible phishing attacks. Attackers can use LLM analysis to obtain personal information of IoT administrators from social media or leaked data and automatically generate highly customized phishing emails or messages. Finally, in terms of lateral movement and path planning of attacks, reinforcement learning provides intelligent decision-making tools for attackers. Attackers can train a reinforcement learning agent, modeling the simulated or initially invaded IoT network environment as a state space. The agent aims to obtain control of key systems as its ultimate goal, and through continuous trial and error and learning, autonomously explores the optimal penetration path to bypass internal defense nodes and reach the core target. This automated attack planning capability makes large-scale and multi-stage attacks on complex heterogeneous IoT networks possible, significantly increasing the difficulty for defenders in tracking and responding. In summary, artificial intelligence makes attacks more automatic, more covert, and more efficient, presenting a greater challenge to defense.

3.2. AI Security

As artificial intelligence becomes deeply integrated into IoT security defense systems, the models themselves have also become new attack surfaces, giving rise to a series of adversarial machine learning attacks. These attacks can be mainly divided into four categories, aiming to compromise the confidentiality and integrity of AI systems. The first type is evasion attacks, which occur during the model inference phase. Attackers generate adversarial samples by making carefully designed subtle perturbations to malicious traffic or software samples, deceiving target detection models and causing them to misclassify malicious inputs as normal—for example, disguising intrusion traffic to mimic legitimate communication patterns to bypass anomaly detection. The second type is poisoning attacks, which take place during the model training phase. Attackers inject malicious samples into the training dataset or tamper with model parameters uploaded by clients in federated learning, systematically contaminating the training process. This leads to performance degradation in the deployed model or the creation of specific backdoors, causing it to lose detection capability against certain attacks in the future. The third type of attack aims to steal the intellectual property of models. Attackers repeatedly query AI security APIs deployed on IoT edge devices or in the cloud, analyzing the correspondences between inputs and outputs to low-cost reproduce a functionally similar substitute model. This not only steals core assets but may also allow further analysis of the model’s weaknesses to plan evasion attacks. The fourth type of attack threatens the privacy of training data, mainly including membership inference attacks and attribute inference attacks. Attackers use access to published AI models to infer whether a specific individual's data record exists in the model's original training dataset or even deduce certain sensitive attributes of the data by analyzing outputs or intermediate features. In the IoT context, this could mean inferring from an anomaly detection model whether a specific device

has ever experienced a failure or been attacked, leading to the leakage of sensitive business information or user behavior patterns. These four types of attacks collectively highlight the vulnerabilities of AI models in adversarial environments, posing severe challenges to their reliable deployment in security-critical domains.

4. Challenges and Development

In the implementation of artificial intelligence-enabled IoT security protection, four technical bottlenecks are particularly prominent. First, terminal devices have extremely limited computing power, storage, and power budgets, making them incapable of supporting conventional deep networks. To achieve real-time edge inference while maintaining detection accuracy and robustness, highly compact models must be obtained through structured pruning, quantization, knowledge distillation, or neural architecture search. Second, public security datasets are limited in size, show class imbalance, contain significant noise, and are difficult to cover the full spectrum of real attacks. This leads to supervised models being prone to overfitting and having poor generalization performance. Therefore, it is urgent to build a high-trust, fine-grained, and continuously updated data supply mechanism tailored to IoT traffic characteristics. Third, security response relies on high-confidence decisions. If a model only outputs black-box alerts without feature-level evidence chains, it is difficult for operation and maintenance personnel to verify and trace. Explainable structures or post-hoc explanation techniques must be introduced to make the decision basis transparent and auditable. Fourth, device architectures, communication protocols, and application scenarios are highly fragmented, and threat types vary significantly. A single model cannot operate reliably across scenarios; it is necessary to research general defense frameworks that can adapt and migrate without retraining to achieve scalable deployment of IoT security capabilities.

Future IoT security research is showing a trend of diversification and cross-domain integration, aiming to build more proactive, intelligent, and adaptive defense systems. Embodied security intelligence represents an important direction, with the core being the development of AI agents that can autonomously perceive, analyze, and take actions. These agents will be deeply embedded in IoT environments, continuously monitoring networks like virtual security personnel, capable of identifying complex attack chains in real time and automatically executing a series of responses, from isolating abnormal nodes to adjusting firewall policies, achieving a high degree of automation in security operations. Meanwhile, quantum machine learning brings new opportunities and challenges to IoT security. In the long term, quantum computing may greatly accelerate the training of complex threat analysis models and even break existing encryption systems, but it also fosters new defense research such as post-quantum cryptography, promoting the forward-looking evolution of security paradigms. Bio-inspired security mechanisms draw inspiration from nature, such as mimicking the distributed recognition, memory, and adaptive clearance abilities of biological immune systems to design security models with self-learning, anomaly detection, and collaborative response features, enhancing the resilience and self-healing capabilities of IoT systems against unknown threats. Security operations based on large language models (LLMs) are becoming key to improving efficiency. By leveraging LLMs' powerful natural language understanding and generation capabilities, security systems can automatically process massive threat intelligence reports, generate easy-to-understand summaries and action recommendations, and allow administrators to query and command through natural dialogue, greatly reducing the threshold and response time of security operations. These directions are interconnected and complementary, jointly painting an overall picture of next-generation IoT security technology progressing toward autonomy, intelligence, and human-centric development.

5. Conclusion

This paper explores the dual role of artificial intelligence in the field of IoT security. On one hand, AI significantly enhances the effectiveness of dealing with large-scale and complex attacks through intelligent threat detection, adaptive authentication, and automated responses; on the other hand, attackers are exploiting AI to generate malicious samples that evade detection, conduct precise social engineering attacks, and even carry out adversarial attacks directly targeting AI models. The rapid escalation of this offense-defense game highlights the urgent need to build a secure, reliable, and trustworthy AI-driven IoT security system that is explainable. While academia and industry actively promote technological innovation, they must also systematically assess and manage the associated risks from a forward-looking perspective to guide the healthy development of artificial intelligence in the field of IoT security.

References

[1] Zhang Y, Li J, Wang H. LSTM-based real-time DDoS detection for resource-constrained IoT gateways. IEEE Internet of Things Journal, 2023, 10 (8): 7123 - 7135.

[2] Kumar A, Park S. Lightweight autoencoder for botnet anomaly detection in industrial IoT. Computers & Security, 2021, 109: 102378.

[3] Chen X, Liu M, Ruan N. GNN-driven threat hunting by fusing network traffic and sensor logs in smart factories. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022: 1841 - 1854.

[4] Ferretti M, Nicolazzo S, Nocera A. H2O: Secure interactions in IoT via behavioral fingerprinting. Future Internet, 2021, 13: 117.

[5] Li Z, Wang Q, Zhao Z B, et al. Continuous identity authentication method for power wireless terminals based on CSI radio frequency fingerprint. High Voltage Engineering, 2022, 48 (9): 3447 - 3455.

[6] Nguyen T V, Marchal S, Miettinen M, et al. DIoT-FL: A federated learning framework for distributed intrusion detection in large-scale IoT systems. In: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. 2022: 809 - 823.

[7] Wang H, Wang Z, Yang J. Fed-IDS: Federated learning-enabled intrusion detection scheme for IoT networks. IEEE Internet of Things Journal, 2022, 9 (22): 22455 - 22466.

[8] Wang X Z. BCS-FL: A blockchain-based privacy-preserving federated learning framework for industrial internet of things. Journal of Modeling and Simulation, 2025, 14 (5): 458 - 471.