

Challenges and Defense Technologies in Cybersecurity Based on Artificial Intelligence

Zikang Nie

*Yangzhou University, Yangzhou, China
17712674309@163.com*

Abstract: With the rapid advancement of network technologies, emerging tools of the new era have increasingly drawn attention to issues of cybersecurity. This evolving landscape presents both unprecedented challenges and significant opportunities for cyber defense. Among the most transformative forces is artificial intelligence (AI), which, while driving remarkable progress across various sectors, has also profoundly impacted the cybersecurity domain. Notably, AI-enabled cyberattacks are becoming increasingly sophisticated, dynamic, and difficult to detect, outpacing the capabilities of traditional security defense mechanisms. These legacy systems often lack the intelligent responsiveness required to counteract the rapidly evolving threat landscape, making them inadequate in the face of complex, large-scale attacks originating globally. In response to these developments, this paper provides a comprehensive overview of the interplay between AI and cybersecurity. It outlines the defining characteristics and developmental trends of AI technologies in the context of cybersecurity and conducts a dual-perspective literature review—exploring both the ways in which AI enhances cyberattacks and how it can be harnessed to strengthen cyber defense. The paper examines state-of-the-art AI-driven attack techniques as well as contemporary AI-based defensive solutions, offering a comparative analysis of their strengths and limitations. Finally, it addresses current developmental constraints and proposes directions for future research. The study underscores the growing importance of AI in cybersecurity and offers strategic insights into mitigating AI-enabled cyber threats.

Keywords: Artificial intelligence, Network attack, Network defense technology

1. Introduction

1.1. Background

1.1.1. Introduction of artificial intelligence

Artificial intelligence (AI) refers to the theories, methodologies, technologies, and application systems that utilize digital computers or machines controlled by digital computers to simulate, extend, and augment human intelligence. By mimicking human cognitive processes, AI is capable of learning, reasoning, and processing information. It primarily involves abilities such as environmental perception, knowledge acquisition, and decision-making aimed at achieving optimal outcomes. In recent years, AI has experienced rapid development and gained widespread attention, becoming integrated into various aspects of human life, such as healthcare, autonomous driving, and more.

However, its application potential in the field of cybersecurity surpasses that in many other domains. Malicious actors have exploited AI to develop increasingly sophisticated cyberattack techniques capable of bypassing traditional security defenses. Yet, the influence of AI on cybersecurity is inherently dualistic; just as AI empowers attackers, it also enables the transformation of passive defense into proactive defense. This allows for improved threat identification and faster incident response times.

1.1.2. Introduction to cybersecurity

Cybersecurity threats refer to malicious attacks carried out within cyberspace, and correspondingly, cybersecurity defense refers to the set of technical and managerial strategies employed to counter such threats. With the widespread proliferation of the internet, cybersecurity has become a matter of growing importance to the public. However, the development of cybersecurity defense technologies faces significant challenges. One of the most critical issues is the diversity and unpredictability of cyber threats, which often outpace the evolution of defense mechanisms. Attackers continuously adopt and iterate novel techniques to circumvent existing and widely known security measures, rendering many traditional defenses obsolete.

1.1.3. Artificial intelligence as both a challenge and a defense technology in cybersecurity

The complexity of modern network environments and the massive volume of data they generate have significantly increased the difficulty of effective defense. Under constraints, such as limited funding, it is often challenging to implement sufficient security measures, leading to data breaches, financial losses, and threats to public safety. While AI has undeniably empowered cyberattacks, it also introduces new mechanisms for cybersecurity defense. By leveraging AI, security systems can evolve beyond static, reactive models to more dynamic, intelligent, and proactive defense strategies.

1.2. Research significance

This study provides an overview of the characteristics and developmental trends of artificial intelligence and cybersecurity in the contemporary era. Through a comprehensive literature review, it examines how AI enables cyberattacks and simultaneously supports cybersecurity defense. It compares various AI-based defense techniques in terms of their strengths and limitations, and further explores current developmental bottlenecks and future research directions. By building AI-driven cybersecurity defense mechanisms, organizations can enhance their ability to combat cyber threats, reduce associated risks, and establish more robust and comprehensive security frameworks. Ultimately, this contributes to maintaining stability and safety in cyberspace, improving the public's sense of security, satisfaction, and well-being, and further reinforcing national security.

2. Literature review

2.1. Challenges of artificial intelligence in cybersecurity

Artificial intelligence (AI) presents multifaceted challenges in the field of cybersecurity, stemming not only from technological limitations but also from broader concerns, including legal, ethical, and data privacy issues.

2.1.1. Data privacy and leakage

In the digital age, individuals are increasingly exposed online, making privacy protection more critical than ever. The leakage of personal information can lead to severe financial losses. AI

applications in cybersecurity typically involve the processing of large volumes of sensitive data and personal information. Malicious use of AI to extract unlawful information from databases can lead to significant data breaches. Ensuring the secure storage and transmission of such data has therefore become a primary concern.

2.1.2. Partial uncontrollability of AI

Cyber attackers may manipulate AI systems by introducing subtle perturbations through input data, causing incorrect judgments or decisions. For instance, compromising autonomous driving systems could have catastrophic consequences. Additionally, due to incomplete or biased training datasets, AI models may exhibit discriminatory behavior or produce unfair outcomes. Furthermore, current legal frameworks struggle to define accountability when AI systems make erroneous decisions that lead to adverse consequences.

2.2. Empowerment of cyberattacks through artificial intelligence

In today's cyber landscape, threats are increasingly diverse and sophisticated. The advent of AI has accelerated this trend by enabling automation, scalability, stealth, and evasion in cyberattacks.

2.2.1. Autonomous and scalable Denial-of-Service attacks

The rapid proliferation of Internet of Things (IoT) devices has greatly increased the number of internet-exposed endpoints. In such environments, attackers can deploy various methods, including Distributed Denial-of-Service (DDoS) attacks, malware, and phishing campaigns. AI has significantly enhanced the scale and automation of these attacks, making them more difficult to detect and mitigate. For instance, botnets have been used to launch massive DDoS attacks that severely disrupt target websites, marking a trend toward increasingly autonomous and large-scale cyber threats [1].

2.2.2. Intelligent and targeted malware threats

Attackers are increasingly using highly customized malware to enhance the precision, success rate, and efficiency of their attacks [1]. AI has significantly improved the speed and accuracy of malware generation, surpassing traditional methods. These AI-enhanced techniques are now being employed in sophisticated Advanced Persistent Threat (APT) campaigns.

2.2.3. Intelligent social engineering threats

AI techniques such as machine learning and neural networks are being used to automate the entire process of social engineering attacks—phishing, virus dissemination, and spam distribution [1]. Automated phishing systems represent a particularly prevalent form of attack, where attackers analyze victims' interests to generate highly targeted phishing scripts and messages.

2.2.4. Threats from Generative Adversarial Networks

Generative Adversarial Networks, consisting of a generator and discriminator, can produce realistic synthetic content through adversarial training. Proposed in 2014, GANs have become a core model in deep learning. They are capable of generating deceptive content such as fake text, images, audio, and video [2]. The widespread dissemination of such falsified content poses a significant threat to cybersecurity by misleading users and facilitating misinformation campaigns.

2.3. Common AI technologies in cybersecurity defense

2.3.1. Situation awareness prediction using least squares Support Vector Machines

Predicting cybersecurity posture requires extensive collection and analysis of security information. Long-term evaluation relies heavily on both quantitative and qualitative data. Support Vector Machines (SVM), particularly in their non-linear form, offer robust modeling capabilities for such tasks. The situation prediction mechanism involves: training on selected data samples using a sample selection algorithm based on information criteria; (2) filtering evaluation data to retain elements with strong predictive value; (3) validating the sample quality through simulated network attacks [3].

2.3.2. Web Application Firewall based on Recurrent Neural Networks

Web Application Firewalls (WAFs) use rule-based systems to detect and filter HTTP traffic. Given the complexity of modern malicious traffic, WAF rules must be continuously updated [4]. A WAF reinforcement solution based on Recurrent Neural Networks enables automated strengthening without reliance on expert-defined rules.

2.3.3. Malware detection with deep learning

Traditional malware detection techniques rely on predefined virus signatures, which are often ineffective against rapidly evolving threats. They also require continuous manual rule updates. Deep learning-based approaches leverage the representational power of neural networks to detect malware with higher accuracy [5]. These models can autonomously learn features from code samples and generalize well to detect unknown variants, thereby enhancing overall system security [6].

2.3.4. Malicious behavior detection using Convolutional Neural Networks

Convolutional Neural Networks (CNNs) can be used to detect malicious behavior through steps including data preprocessing, model construction, convolution and pooling operations, and final classification [7]. CNNs help identify anomalies by distinguishing between normal and abnormal data behaviors, supporting intrusion detection and analysis tailored to specific application needs.

2.3.5. Agent-based rootkit detection

The foundational model for Intrusion Detection Systems (IDS) was first proposed by Denning, emphasizing rule-based updates in synchronized environments [8]. OSSEC is a widely used open-source Host-based IDS (HIDS), which collects data through agents installed on hosts. These agents support log analysis and rootkit detection through integrity checking [9]. Administrators can customize configuration files as needed. Integrating AI and neural networks into IDS design significantly enhances the defensive capability of modern computer systems.

2.3.6. Simulated network attack modeling

This framework provides a comprehensive method for modeling and analyzing network attacks. It abstracts critical network system information for security analysis, modeling both network elements and attacker capabilities. The generated models are then integrated into vulnerability analysis workflows [10]. Based on attack chains, this framework helps network administrators understand multi-stage attacks and system vulnerabilities, directly supporting strategic cybersecurity management.

3. Advantages and disadvantages of several defense techniques

The situation awareness prediction mechanism based on Least Squares Support Vector Machines (LS-SVM) offers significant advantages. After analyzing a large volume of security information, the model gains evolutionary capability, allowing existing cybersecurity defenses to become more adaptive and efficient. It enhances the interception and detection of intrusions, and enables proactive intervention by combining security policies with threat intelligence in anticipation of potential cyberattacks. However, its disadvantages include the risk of suboptimal performance due to improper initialization parameters, which may lead the model to fall into a local optimum. Additionally, since the prediction process requires participation of the entire training dataset each time, it results in increased computational complexity.

The Web Application Firewall (WAF) based on Recurrent Neural Networks (RNNs) has the advantage of utilizing scoring functions to identify critical strings and generate reinforced signatures. This helps protect the WAF from subsequent similar attacks, and overcomes the traditional reliance on expert knowledge. Moreover, through training, the model can distinguish between normal and malicious traffic patterns, thus improving detection accuracy. However, its main limitation is performance degradation under high traffic volumes, where the WAF may become a bottleneck in the system.

The malware detection technique based on deep learning excels in its ability to automatically learn feature representations from input data without manual intervention. This enables the system to better adapt to continuously evolving malware scripts and enhances detection accuracy. Nonetheless, the wide variety and rapid evolution of malware pose significant challenges for model training and updates. Constant expansion of the malware sample library and repeated retraining of the model require substantial time and human resources.

The detection of malicious behavior using Convolutional Neural Networks (CNNs) offers several strengths. It can automatically learn and extract key features from data, making it particularly well-suited for analyzing complex datasets. CNNs also demonstrate robustness to noise and distortion in input data, and possess strong generalization capabilities, which facilitates the detection of novel malicious behaviors. On the downside, CNNs are vulnerable to adversarial attacks—attackers can craft deceptive inputs to trick the model into making incorrect predictions or classifications.

The use of agents to collect data for rootkit detection presents a key advantage: as these agents operate directly within the system, they can obtain more accurate and detailed performance metrics and anomalous events, which provides strong support for effective rootkit detection. However, this approach carries potential security risks. If the agent software contains vulnerabilities or is exploited by malicious actors, it may itself become a security threat to the system.

The simulation-based network attack modeling technique has the advantage of enabling the development of integrated analytical tools for mitigating threats. These tools can uncover vulnerabilities in enterprise security architectures, correlate events, and support the modeling, visualization, and prediction of multi-stage attacks. As a result, enterprises and organizations can familiarize themselves with potential attack scenarios and prepare corresponding incident response procedures before real attacks occur. This contributes to rapid and accurate responses during an attack, minimizing potential damage. However, its limitations include the inability to cover all possible attack types and scenarios, and constraints in technology that may prevent fully accurate simulation of real-world attacks. Consequently, the models may exhibit some inaccuracies and limitations in practical applications.

4. Conclusion

The application of artificial intelligence (AI) in cybersecurity holds significant potential for future development. However, the effectiveness of AI-based defense techniques is heavily dependent on the quality of algorithms and the training data used. If algorithm parameters are not optimally tuned or if high-quality training data is lacking, AI systems may fail to accurately identify threats or may generate false positives. Although AI technologies excel at rapidly detecting abnormal behaviors and potential threats within networks, inherent limitations in their algorithms and operational mechanisms still make them susceptible to both false alarms and missed detections.

While automation is a notable advantage of AI, excessive reliance on automated systems may reduce the role of human intervention. This overdependence can result in the neglect of critical security issues that require human judgment and response.

Looking ahead, future research should focus on developing more robust algorithms and models to enhance the resilience of AI systems against adversarial attacks. Additionally, improving the interpretability of AI systems is crucial—making the decision-making process more transparent and understandable can significantly enhance trust and usability. Furthermore, advancements in data encryption, access control, and privacy-preserving technologies should be pursued to ensure that AI systems can securely handle sensitive user data.

Moreover, integrating AI with other emerging security technologies—such as blockchain and quantum computing—may lead to the creation of more comprehensive and intelligent defense frameworks. Interdisciplinary collaboration should also be encouraged, leveraging expertise from diverse fields to provide AI systems with deeper and more holistic insights.

Finally, as technology continues to evolve, AI systems should aim to achieve higher levels of autonomy and intelligence. They should be capable of proactively identifying potential threats, implementing preventive security measures in real time, and predicting future cyberattack trends to enable preemptive defense strategies.

References

- [1] Fang, B., Shi, J., Wang, Z., et al. (2021). *Security threats and countermeasures of AI-empowered cyberattacks*. *Strategic Study of CAE*, 23(3):e60-66. DOI:10.15302/J-SSCAE-2021.03.002 .
- [2] Wang, W., Li, Y., Guo, F., et al. (2022). *A survey on generative adversarial networks and their application in text-to-image synthesis*. *Computer Engineering and Applications*, 58(19):14-36. DOI:10.3778/j.issn.1002-8331.2205-0119.
- [3] Peng, X. (2023). *An analysis of the application of artificial intelligence technology in cybersecurity defense*. *Cybersecurity Technology and Application*. DOI:10.3969/j.issn.1009-6833.2023.09.012 .
- [4] Zhu, S., Du, R., Chen, J., et al. (2022). *A reinforcement scheme for Web application firewalls based on recurrent neural networks*. *Computer Engineering*, 48(11):120-126. DOI:10.19678/j.issn.1000-3428.0063581.
- [5] Wang, M., Sun, Z., Yang, H. (2024). *Threats and countermeasures of artificial intelligence in cybersecurity*. *Cybersecurity Technology and Application*, (1):163-165. DOI:10.3969/j.issn.1009-6833.2024.01.063.
- [6] Feng, R., Gou, Y., Zhang, K., et al. (2024). *Cybersecurity technology and applications based on artificial intelligence*. *Cybersecurity Technology and Application*, (3):118-121. DOI:10.3969/j.issn.1009-6833.2024.03.044.
- [7] Zhai, X. (2025). *Research on cybersecurity defense technology based on artificial intelligence*. *Information Recording Materials*, 26(1):29-31.
- [8] Li, Y., Huang, Gq., Wang, Cz. et al. (2019) *Analysis framework of network security situational awareness and comparison of implementation methods*. *J Wireless Com Network*, 205 .
- [9] Bian, L. (2023) *Design of Computer Network Security Defense System Based on Artificial Intelligence and Neural Network*. *Wireless Pers Commun*.
- [10] Dawkins J., Hale J. (2004) *A systematic approach to multi-stage network attack analysis*, *Second IEEE International Information Assurance Workshop Proceedings.*, Charlotte, NC, USA, pp. 48-56e. doi: 10.1109/IWIA.2004.1288037.