

Analysis of Artificial Intelligence in Internet Security

Wutong Dong

*Case School of Engineering, Case Western Reserve University, Ohio, United States
wxd184@case.edu*

Abstract: With the rapid development of artificial intelligence (AI) this year, it has become a key technology that can improve Internet security. Regarding network security, AI technology and its related tools have been widely used in intrusion detection, malware identification, and phishing attack defense. After reviewing the latest relevant literature, this paper reviews the current technical standards and uses the methods mentioned in the literature to analyze the advantages and limitations of AI in network security, and focuses on the challenges it faces in the face of adversarial attacks, data privacy leakage, and interpretable models. According to the research reviewed, although the current AI model has excellent performance in an ideal experimental environment, its reliability in the face of real environments or hostile conditions still has room for improvement. To this end, this paper deeply explores optimization directions such as adversarial training, federated learning, zero-shot learning, and interpretable AI, and puts forward future development trends and governance recommendations. Finally, according to the research, it can be concluded that the technological innovation of AI and network security needs to be paralleled with ethical supervision to achieve AI's reliable application in network security.

Keywords: Artificial Intelligence (AI), Intrusion Detection, Data Privacy, Federated Learning, Internet Security.

1. Introduction

Artificial intelligence (AI) is increasingly integrated into digital technology structures, including Internet security. As cyberattacks or threats become increasingly complex and pervasive, as computer algorithms and computing power grow, intelligent, adaptive machine learning algorithms that can learn and summarize patterns from big data are replacing traditional security measures based on basic network rules. These AI-driven technologies can detect malware in networks more quickly, identify anomalies more accurately, and respond more effectively to new types of threats [1]. As a result, AI has become an indispensable tool for defending against the rapidly evolving strategies of cybercriminals.

However, while AI revolutionizes cybersecurity, its adoption also introduces a new class of threats. While these algorithms strengthen digital defenses, adversaries can also exploit them. Malicious actors increasingly employ adversarial machine learning to deceive detection systems, launch targeted model inversion and membership inference attacks that extract sensitive data, and generate convincing phishing content with large language models. This dual-use nature of AI—its capacity to serve as both a protector and an attacker—creates a complex security landscape where defensive capabilities improve offensive abilities, as both rely on the same technologies. In this new paradigm,

defenders and adversaries are powered by similar technologies, leading to a competitive escalation in AI capabilities within internet security.

In addition, AI improves network security performance and addresses special vulnerabilities that traditional systems do not have. The most prominent of these is the impact of adversarial samples on machine learning models. Specifically, even though AI-designed algorithms seem impeccable and efficient, they can cause the models to produce errors or even harmful results in experiments. For example, when tested in a black box environment, it turns out that attackers do not need direct access to the underlying model of the network architecture to carry out attacks, which makes it very difficult to detect and mitigate attacks. Therefore, in the application areas of machine learning (such as fraud detection, intrusion prevention, or biometric authentication), just one adversarial input may endanger system integrity or leak sensitive information.

In addition to the abovementioned concerns, the challenges of combining AI with cybersecurity are not limited to technical vulnerabilities. Still, it can also reduce the threshold for sophisticated attacks. For instance, under certain conditions, attackers can use deep fake technology to avoid authorization of administrators, then spread false information, or bypass recognition systems [2, 3]. In addition, chatbots can be hijacked or manipulated through an injection attack, and large-scale social engineering campaigns can be carried out through highly personalized automation. These tools help attackers craft more deceptive messages cheaply, imitate legitimate behavior, and break through human defenses.

Moreover, it is also essential to study and analyze the ethical issues and the corresponding solutions created by AI and internet security. As algorithms become increasingly autonomous and opaque, concerns about accountability, fairness, and subjective unintended biases arise. Many models operate as “black boxes,” and explaining or justifying their decisions is difficult. This issue makes compliance with data protection laws such as the General Data Protection Regulation (GDPR) more complicated than it would otherwise be. In addition, the widespread deployment of AI-driven surveillance and predictive policing systems raises questions about civil liberties and the balance between security and privacy. In some cases, poorly regulated AI tools may inadvertently reinforce discrimination or be abused by authoritarian regimes in the name of national security [4, 5].

Given these trends and possibilities, there is a need for a multidimensional and nuanced understanding of the role of AI in Internet security. When developing combined systems, it is necessary to analyze how the two can coexist and enhance each other. Therefore, a comprehensive review of network systems' defensive and offensive capabilities and a rigorous examination of ethical issues and legal frameworks are essential to building strong and reliable AI security systems. In addition, the cybersecurity community must promote collaboration among researchers, policymakers, and industry stakeholders to establish standards for responsible AI development and deployment.

Thus, this essay will explore related papers and articles to illustrate the comprehensive relationship between AI and internet security. It investigates the opportunities AI offers to enhance digital protection and the risks it introduces when adversaries exploit it. This study aims to provide a balanced understanding of how AI functions as a shield and a vulnerability in the modern cybersecurity environment by reviewing the latest academic research, real-world threat scenarios, and cutting-edge mitigation strategies. The goal is to illuminate the current state of AI in internet security and anticipate the evolving challenges it may bring in the years ahead.

2. Methodology

2.1. Data explanations

To further explore the relationship between AI and Internet security, this study adopted a method based on a systematic literature review and secondary data analysis. Instead of relying directly on

original experimental data, it referred to published academic papers, institutional reports, and widely cited public datasets in cybersecurity and machine learning. This approach helps to fully understand the advantages and potential drawbacks of AI-driven cybersecurity solutions.

The research is based on ten high-quality professional papers. These papers were selected based on their high relevance to the research topic, the time of publication, whether they have passed peer review, and the frequency of citations in the field. The selected materials are diverse in form, covering systematic reviews, experimental studies, and forward-looking analyses. They show the specific application of AI in security systems and the risks of abuse they may encounter.

In addition, the study conducted a full-text review of each paper, focusing on its research methods, experimental design, evaluation criteria, chart data, and conclusions. This in-depth analysis helped to clarify common vulnerabilities in AI architecture and training data exposed by issues such as membership inference attacks and model inversion [6, 7]. In addition to academic materials, the study also referred to multiple recognized datasets. Although these datasets are not directly used in experiments, they form the basis for training and evaluating many existing security models. For example, Network Security Laboratory - Knowledge Discovery and Data Mining (NSL-KDD), Intrusion detection evaluation dataset (CICIDS2017), and Knowledge Discovery and Data Mining Cup 1999 (KDD'99) are classic intrusion detection system research datasets, while DREBIN and Virus Share datasets provide standard samples for malware classification [8, 9]. The reference to these datasets makes this study reliable in theory and practice.

To further enhance the study's real-world connection, the method section also includes the analysis of real cases and public incident reports. These cases cover AI-driven phishing, deepfake attacks, and adversarial behaviors that evade terminal protection systems [10]. These events verify that the risks raised in the literature are happening in the real world.

This study builds a deep understanding of the relationship between AI and network security through multi-level data integration, including academic research, standard datasets, and real cases. This multi-source cross-validation method enhances the study's credibility and lays a solid foundation for the subsequent design, evaluation, and critical analysis of AI systems. The following section will further introduce the specific methods and techniques used in the study and explore their applicability and contribution in this field.

2.2. Method explanations

After identifying the data sources that this study relies on, the next step is to explore the methodologies used in the reviewed literature to build, test, and evaluate practical applications of AI in Internet security. Although the specific methods vary greatly depending on the threat type, model architecture, and research objectives, several key technical paths can be identified through analysis, which depict the mainstream methodological framework of AI-driven network security research.

First, a large number of studies rely on supervised learning models. Such methods require labeled datasets to complete classification tasks such as distinguishing normal from malicious network traffic, identifying phishing emails, or detecting malware based on static and behavioral features. In terms of model selection, the research covers both traditional algorithms, such as decision trees, support vector machines, and logistic regression, as well as deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [8,9]. The choice of model is closely related to the data's complexity and the task's characteristics. For example, CNN is often used to classify malware images or bytecodes due to its ability to recognize spatial patterns.

In addition to supervised learning, some studies use unsupervised learning strategies to identify abnormal system behavior patterns or network traffic patterns. These methods do not rely on labeled data, but instead aim to capture behaviors that deviate from the norm, such as unauthorized access, data leakage, or exploitation of unknown vulnerabilities. Clustering algorithms (such as k-means and

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)) and autoencoders are widely used in this type of analysis, where autoencoders can learn standard patterns and detect abnormal inputs based on them [5,9]. These methods are instrumental in real-world environments, especially in scenarios where high-quality labeled data is challenging to obtain.

In adversarial machine learning, researchers focus on developing and testing attack algorithms to reveal potential vulnerabilities of existing models. Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner attack are standard techniques that aim to construct adversarial examples - inputs that appear normal to humans but can mislead AI models [4, 6, 7]. These attacks have been tested in white-box (where the attacker knows the model structure) and black-box (where the attacker has limited information) environments. The former evaluates extreme vulnerabilities and is closer to realistic threat scenarios.

At the same time, defense mechanisms have become a core component of many research methods. Adversarial training is a common strategy involving standard and adversarial samples for model training. Other defenses include distillation defense (reducing sensitivity through “soft label” learning), feature compression (reducing input complexity), and input purification (preprocessing data to remove potential attack components) [6-8]. The effectiveness of these defenses is usually evaluated through robustness indicators, such as accuracy under attack, confidence reduction, and attack success rate.

Furthermore, some literature proposes an analysis framework based on the system life cycle to identify the attack surfaces that AI systems may encounter during the data collection, model training, deployment, and reasoning stages. For example, the life cycle model established by Wu et al. reveals that security risks run through the entire machine learning process rather than being limited to specific links, thereby expanding the research perspective from point risks to systemic defense [7].

To enhance the practical application value, researchers generally introduce evaluation methods such as cross-validation and use indicators such as precision, recall, F1 value, and area under the ROC curve (AUC) to measure model performance. These indicators reflect the prediction accuracy and the stability of the model under adversarial conditions. Some studies even simulated real-world deployment environments to test models' response speed, resource usage, and adversarial adaptability in real-time systems [1, 4, 9].

In summary, the methods used in the analyzed literature demonstrate the multi-dimensional progress of research on integrating cybersecurity and artificial intelligence. These methods cover the entire technical chain from threat detection and model building to attack demonstration and defense optimization. The following section will further analyze how these methodological choices affect research results and provide a more strategic understanding of the dual role of AI in cybersecurity.

2.3. Additional methodological details

After discussing the data types and basic methods used in the previous section, this section will further introduce some typical experimental assumptions, modeling methods, and evaluation criteria in the selected literature. Although these details are sometimes not particularly emphasized in the article, they are crucial to understanding the practical significance of the results and the applicability of AI systems in security scenarios.

Regarding practical operations, most of the literature uses some widely used Python machine learning frameworks, such as TensorFlow, PyTorch, Scikit-learn, or Keras. These tools support flexible model design and training and can be easily combined with visualization or parameter adjustment tools. The training process usually adopts supervised learning, and the dataset is also divided into a training set, a validation set, and a test set for model evaluation. Regarding dataset division, some studies use the traditional Hold-out method. In contrast, some studies use cross-validation (especially K-Fold) strategies, which can more comprehensively test the stability of the

model. As for the data itself, tasks such as intrusion detection often encounter the problem of class imbalance. Therefore, many studies use stratified sampling methods to keep the label distribution in each data subset consistent with the original data [5, 9].

In addition to basic data partitioning and model training methods, many studies have also emphasized the impact of hyperparameter tuning on model performance. Grid search or random search is often used to test various parameter combinations, such as learning rate, number of network layers, activation function type, dropout rate, etc. Especially in deep learning tasks, controlling model capacity to avoid overfitting is a key issue. In terms of optimizers, Adam and Root Mean Square Propagation (RMSProp) are more commonly used; of course, many studies also choose to use the more traditional stochastic gradient descent (SGD).

In security-related research, adversarial attacks are a standard part. Especially in white-box attacks, the attacker has full access to the model structure and parameters, so methods such as FGSM, PGD, or Carlini-Wagner are often used in research to generate adversarial samples. These samples look the same as the original input, but as long as a small perturbation is applied (limited by the ℓ_0 , ℓ_2 , or ℓ_∞ norm), the model may make wrong judgments [4, 6]. In contrast, black-box attacks are more restricted because the attacker cannot access the internal information of the model. Many studies simulate the target model by training a proxy model and then use the generated adversarial samples for transfer testing. This approach is more valuable in the real world because most attackers do not know what model researchers are using and can only guess based on the relationship between input and output.

In terms of defense, different studies have different strategies. Adversarial training is a standard method that puts adversarial and normal samples into the training set to train the model to improve its robustness. In addition, there are methods such as feature compression, input denoising, or defense distillation. The core idea of these practices is to reduce the model's sensitivity to perturbations. When evaluating the effectiveness of these defense measures, researchers usually focus on the accuracy under attack, the success rate of adversarial samples, or the change in model prediction confidence [6-8].

In addition to these classic modeling and testing methods, some studies have tried introducing simulation methods closer to the real network environment. For example, some studies use tools such as CICFlowMeter or Wireshark to analyze network traffic and extract traffic features (such as source address, destination address, protocol type, number of bytes, etc.). These features are then provided as input to the model to determine whether there is abnormal behavior. In malware detection, some studies run sample programs in a sandbox environment, extract their dynamic behavior logs or frequency distribution of opcodes, and then construct training samples [5].

Some studies not only focus on the model itself, but also model threats from a more macroscopic system lifecycle perspective. For example, Wu et al. proposed a relatively complete framework that maps possible attack paths to each link from data collection to model deployment [7]. The advantage of this approach is that it can help researchers discover security issues that may be caused in the early data processing or training stage, rather than the inference stage.

In addition, in recent years, more studies have emphasized the importance of model interpretability, especially in high-risk application fields such as finance, medical care, or law enforcement. Shapley Additive Explanations (SHAP), Low Light Image Enhancement (LIME), and some attention visualization methods are widely used to analyze how the model makes each decision.

3. Results and discussion

3.1. Results analysis

AI has a dual nature in the field of Internet security. It is like a double-edged sword. On the one hand, it can demonstrate impressive capabilities in detecting threats, classifying malware, and filtering

phishing attacks; on the other hand, it is still vulnerable to manipulation and privacy violations. This section analyzes the performance metrics, adversarial effects, and defense effectiveness reported in the reviewed studies.

In most standardized evaluation environments, AI models generally accurately perform security tasks. For example, in experiments using the DREBIN dataset for malware classification, deep belief networks (DBNs) achieved classification accuracy exceeding 96% [5]. Similarly, in intrusion detection tasks based on the CICIDS2017 dataset, studies using deep learning methods such as CNNs or RNNs reported precision and recall exceeding 95% [1, 5]. Although these numbers show that the prediction performance of the models under clean data conditions is quite ideal, it should be noted that these results are often obtained without adversarial perturbations.

The vulnerability will gradually become apparent once the model faces a more complex adversarial environment. According to the National Institute of Standards and Technology (NIST) research report [6]. Figure 1 (provided in the original text) shows a Pareto efficiency diagram to express the trade-off between model accuracy and adversarial robustness. Figure 1 shows that models performing well under non-perturbation conditions tend to show weaker defense capabilities after introducing adversarial inputs. This curve intuitively reflects a core problem: achieving a complete balance between improving predictive ability and enhancing robustness in model development is often difficult. Only those models that maintain a good balance in both aspects will be close to the Pareto frontier. In contrast, those insufficient in either element will fall into the suboptimal feasible.

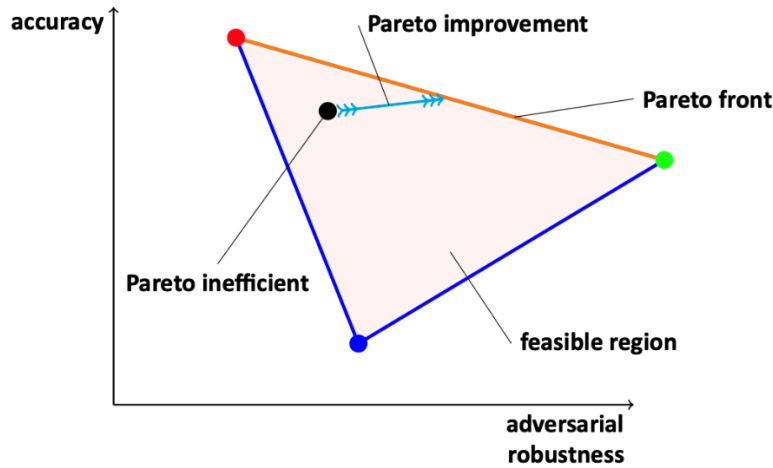


Figure 1: Trade-Off between clean accuracy and robustness (photo credit: original).

To characterize these vulnerabilities, Wu et al. proposed a general framework for adversarial machine learning (AML) [7]. The framework defines the adversarial problem space using precise notations for datasets, model states, and attack conditions. These components are combined in mathematical conditions to quantify model behavior. Specifically, Wu et al. define three primary conditions governing adversarial learning:

1) Stealthiness $\mathcal{S}(x_0, x_\epsilon; w_0, w_\epsilon)$

Enforces minimal perturbation between original and adversarial inputs $(x_0, x_\epsilon; w_0, w_\epsilon)$, or between model weights. The goal is to ensure that adversarial modifications remain undetectable to human observers.

2) Benign Consistency $\mathcal{C}(x_0, x_\epsilon; w_0, w_\epsilon)$

Ensures that both the benign model and its adversarially modified version produce the correct output y_0 for a clean input x_0 .

3) Adversarial Inconsistency $\mathcal{L}(x_0, x_\epsilon; w_0, w_\epsilon)$

This drives the model to misclassify adversarial samples as a target label y_ε , differing from their original label.

These conditions are instantiated through specific loss functions such as $\mathcal{L}_{CA}$, $\mathcal{L}_{CB}$, $\mathcal{L}_{IA}$ and $\mathcal{L}_{IB}$, which guide the optimization objectives during attack generation or model hardening. Collectively, these mathematical structures allow researchers to simulate and evaluate attacks across the entire AI lifecycle rigorously. For example, Wu et al. showed how backdoor attacks affect the training phase, while adversarial examples perturb reasoning. Each attack phase is governed by a set of optimization objectives derived from these core conditions. Experimental results based on these formulas show that most models not trained with adversarial defenses experience a significant drop in accuracy even for modest perturbation norms (e.g., $l_\infty < 0.03$). In contrast, adversarial training (a technique for retraining models using perturbed examples) achieves significantly better robust accuracy despite higher false positive rates and longer training times [6, 7]. The result is shown in Figure 2.

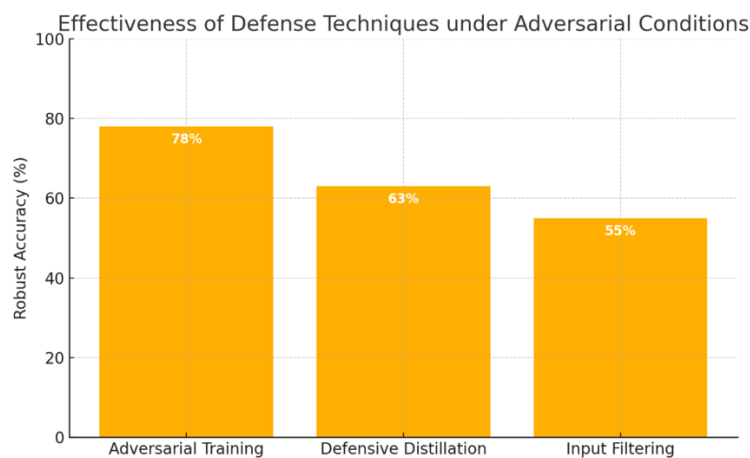


Figure 2: Effectiveness of defense techniques under adversarial conditions (photo credit: original).

These findings demonstrate that adversarial machine learning frameworks are theoretical constructs and practical tools for benchmarking and improving AI security systems. By embedding AML principles directly into evaluation pipelines, developers can preemptively identify failure modes and strengthen model resilience before deployment.

3.2. Thematic discussion (revised with AI usage)

The results above highlight an essential but complex reality: while AI has dramatically enhanced the performance of Internet security systems, it has also brought new technical and ethical challenges. Compared with traditional rule-based detection systems, AI-based models are more effective in discovering security threats, identifying malware, and blocking phishing attacks. However, these models show obvious vulnerabilities when facing adversarial attacks. In other words, the powerful capabilities of AI systems can also become their potential weaknesses. This "advantage is risk" paradox has led to several issues worthy of in-depth discussion on the role of AI in cybersecurity.

One of the core issues is the trade-off between accuracy and adversarial robustness. As shown in the Pareto front diagram in the NIST report, no model can guarantee high accuracy on clean data while also having strong adversarial interference capabilities. This balance is not an abstract theoretical issue, but a choice that many organizations must face when deploying AI security systems in reality. For example, some organizations prefer to pursue high sensitivity to threats, even if this may cause the model to fail when encountering adversarial samples. In contrast, others may sacrifice some detection accuracy to enhance the system's stability in actual attack environments. Although

adversarial training is generally considered a more effective defense method, it often increases the consumption of computing resources, and the model performance when processing ordinary inputs may also decrease as a result [6, 7].

This brings up another topic worth paying attention to: the contradiction between system robustness and practical deployment feasibility. In an experimental environment, researchers can try different defense strategies many times, constantly adjust parameters, and tolerate the time cost of model training. However, in real systems, especially in scenarios that require real-time response, such as intrusion detection or automated threat response systems, computing resources and response speed often constitute key limiting factors. Some lightweight models that perform well in theory may have a significant performance drop when faced with complex and unpredictable real attacks. The pursuit of low latency and high stability in real needs often conflicts with the complex mechanisms required to improve the model's adversarial capabilities.

Another issue closely related to robustness is the interpretability of the model and the resulting attribution of responsibility. The AI systems widely used in the current cybersecurity field, especially those based on deep neural networks, are often criticized as "black boxes". Even if the prediction results are accurate, the system's decision-making process is complicated for the outside world to understand. This situation is particularly prominent in some scenarios that require high interpretability of the results, such as financial fraud detection, biometric authentication, and security management of critical infrastructure. In these applications, wrong judgments can have serious consequences. Without methods such as SHAP, LIME, or attention mechanism visualization, it will be difficult for organizations to clearly explain the decision logic of AI systems to regulators or users, and it may also be challenging to meet the basic requirements of "explainable decisions" in regulations such as the GDPR.

In addition, privacy and ethical risks have also been a common theme in recent research, especially when discussing adversarial attacks. For example, Shokri et al.'s study of membership inference attacks showed that deep learning models may inadvertently leak information about whether a user's information is included in the training data [4]. This risk is even more serious if the model is trained on sensitive data, such as medical records, social media behavior, or user behavior logs. Further attacks, such as model inversion or data extraction, may even be able to reverse restore sensitive features of the input data without triggering security alerts, thereby causing potential privacy leaks.

At a more macro level (shown in Figure 3), the widespread use of AI in cybersecurity has also raised concerns about automation and social control. Brundage et al. proposed in their scenario prediction research that AI systems, especially those with generative or autonomous decision-making capabilities, may be used to manipulate public opinion, monitor society, and even interfere with the operation of critical infrastructure [10]. As facial recognition, predictive law enforcement, and AI-driven surveillance tools continue to expand, society faces a series of ethical controversies regarding privacy, voluntariness, and digital civil liberties.



Figure 3: AI-enhanced crowd surveillance visualization (photo credit: original).

The various topics discussed above reveal a core issue: the credibility of AI is still obviously vulnerable when facing complex network threats. Although AI systems have shown stronger detection and response capabilities than traditional security mechanisms in dealing with static or known threats, their stability and ethical reliability still face many challenges in dynamic and unpredictable real environments. Future research needs to continue to improve the accuracy and robustness of models, but also focus on building AI architectures with explainability, fairness, and security, so that these systems can combat potential risks from the beginning of their design.

In addition, the development of technology must also be coordinated with the evolution of the governance system. Algorithm optimization alone is challenging to solve all risks, so it will become imperative to formulate corresponding regulatory frameworks and ethical norms only when technological innovation and institutional guarantees are promoted simultaneously that the application of AI in the field of cybersecurity truly serves the interests of humanity, rather than weakening individual freedom or undermining the responsibility mechanism of organizations.

3.3. Challenges, future optimization, and trends in AI and security

Although AI has made significant progress in cybersecurity in recent years, some fundamental problems remain unsolved. In controlled experimental environments, many models have demonstrated high threat detection accuracy, but this does not directly translate into actual reliability in the real world. In particular, the generalization capabilities of many models are still limited when dealing with unknown or changing attacks. This limitation is even more pronounced in noisy or adversarial environments, which affect the model's deployment and provide opportunities for experienced attackers.

The trade-off between accuracy and robustness is a long-standing and unresolved core issue among these challenges. As shown in the studies above, especially in the Pareto analysis diagram, the better the performance of most current models on clean data, the more difficult it is to ensure their stability

under adversarial samples. Although adversarial training is considered a strategy that can enhance robustness, it often brings additional computational overhead and significantly reduces the model's accuracy under regular inputs [6, 7]. Therefore, how to improve the system's defense against adversarial behavior while maintaining detection capabilities remains an unavoidable problem in AI system design.

At the same time, attackers constantly evolve, forming a dynamic game with defense systems. With the development of AI technology, attackers are increasingly inclined to exploit the model's weaknesses to achieve their goals. Technical means include model inversion, data poisoning, and adversarial sample generation, allowing attackers to interfere with or mislead the originally well-trained model system. With the development of generative adversarial networks (GANs) and large language models, the concealment and controllability of phishing attacks, fraudulent behaviors, and malware have also significantly improved [4, 10]. This trend further confirms the dual nature of artificial intelligence: it can be used for defense and attack.

Another problem that has not been fully addressed is the high dependence on high-quality annotated data. Many current models still rely on manually annotated training datasets, which tend to become ineffective in the face of evolving threats. In particular, for rare types or emerging forms of cyber attacks, existing data may not provide sufficient representativeness. This deficiency is more serious when facing multilingual, multi-regional, or global cyber threats, weakening the practicality of pre-trained models in diverse environments.

To address these challenges, future research directions should focus more on developing learning mechanisms with adaptive capabilities. For example, few-shot learning and zero-shot learning provide potential feasible paths for identifying new attacks, and they require a low number of samples. In addition, federated learning, as a method that supports cross-device collaborative training, is also considered promising for achieving a wider range of collaborative cybersecurity defense capabilities while protecting user privacy. Introducing these technologies not only improves the system's scalability but also helps reduce reliance on centralized data.

The interpretability of the model is also an issue that cannot be ignored. Since most current deep learning models are "black box" structures and lack a transparent decision-making basis, accountability mechanisms are often complex in security, law, and critical infrastructure scenarios. Explainable AI (XAI) tools such as SHAP and LIME provide a solution that can help developers and users understand the behavioral logic of the model, thereby improving the system's auditability and meeting relevant compliance requirements.

At the same time, the pace of policy, regulation, and technological development also needs to be better coordinated. Although organizations such as NIST have begun to develop an AI risk framework [6], there is still a lack of a widely accepted and universal AI resilience testing standard in cybersecurity. Strengthening the construction of regulatory systems and improving the transparency of information disclosure will be key measures to promote the safe implementation of AI.

Looking ahead to the next decade, several prominent trends may dominate the evolution of cybersecurity. First, more and more autonomous defense agents will be introduced into security systems, which can significantly improve response speed and scalability, but also put higher requirements on regulatory mechanisms. Second, the combination of AI and biometric monitoring is triggering more and more ethical discussions, especially in high-security or high-level state control occasions, where concerns about privacy and civil rights are rising. Finally, the construction of human-machine collaborative systems has become increasingly critical, and they will play a regulatory role in mitigating the risks of automatic decision-making and preventing accidental injuries.

In short, although AI has brought unprecedented possibilities to the digital security system, it must be coordinated from multiple levels, such as technology, policy, and ethics, to realize its potential in infrastructure protection. Only by forming a joint force in these aspects can AI safely and efficiently

expand its scope of application and provide solid protection capabilities for the future network environment.

4. Conclusion

The widespread use of AI in Internet security has driven significant progress in threat detection, anomaly identification, and response automation. Compared with traditional rule-based systems, AI models, intensive learning-based methods, have shown stronger capabilities in identifying known attack patterns and adapting to new threats. However, as these technologies develop, some new security risks have gradually been exposed, including vulnerability to adversarial attacks, potential privacy leakage risks, and model deficiencies in interpretability. As discussed above, the "double-edged sword" nature of AI means that it can be used for defense and malicious purposes. Technical means such as adversarial machine learning, model inversion attacks, and generative forgery are being used by attackers to bypass detection mechanisms and tamper with system outputs. In this process, although defenders try to use adversarial training and input filtering for protection, these strategies often fail to achieve an ideal balance between practicality, efficiency, and transparency.

This study summarizes several current limitations of AI in network security by analyzing recent research literature, mainstream datasets, and model frameworks. These include: insufficient generalization of models in real application environments, strong dependence on labeled training data, and the lack of a unified robustness evaluation standard. Nevertheless, with the continuous development of new methods such as federated learning, few-shot modeling, and XAI, there is still great potential for improving model resilience in the future. In addition to technical paths, the widespread application of AI in network security has also triggered more macro-level reflections, such as social issues, surveillance ethics, and governance mechanisms. In the context of governments and institutions increasingly relying on AI technology for digital defense and behavior monitoring, ensuring the system's legitimacy regarding accountability, fairness, and civil liberties has become unavoidable. These challenges remind authors that future development depends on breakthroughs in engineering technology and requires cross-disciplinary cooperation between policymakers, system developers, and ethical researchers. In short, although AI redefines the future of Internet security, it is not a "universal solution". Its actual effect often depends on the design quality of the model, the deployment method, and the complexity of the environment in which it is located. To leverage the advantages of AI technology while controlling risks, it needs to build a system that is intelligent, secure, explainable, and ethical, and this depends on the co-evolution and coordination of technology and institutions.

References

- [1] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153-1176.
- [2] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy*, 305-316.
- [3] Zhang, Y., & Paxson, V. (2019). Detecting and analyzing automated activity on the internet. *Communications of the ACM*, 62(6), 78-85.
- [4] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *the 2017 IEEE Symposium on Security and Privacy*. 3-18.
- [5] Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. *Information*, 10(4), 122.
- [6] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., & Davies, X. (2025). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. National Institute of Standards and Technology (NIST), AI 100-2e2025.
- [7] Wu, B., Zhu, Z., Liu, L., Liu, Q., He, Z., & Lyu, S. (2023). *Attacks in Adversarial Machine Learning: A Systematic Survey from the Life-cycle Perspective*. arXiv preprint, 9457.

- [8] Li, L. (2024). *Comprehensive Survey on Adversarial Examples in Cybersecurity: Impacts, Challenges, and Mitigation Strategies*. 12217.
- [9] Atadoga A, Sodiya E O, Umoga U J, et al. (2024). *A comprehensive review of machine learning's role in enhancing network security and threat detection*. *World Journal of Advanced Research and Reviews*, 21(2), 877-886.
- [10] Djenna A, Bouridane A, Rubab S, et al. (2023). *Artificial intelligence-based malware detection, analysis, and mitigation*, 15(3): 677.