ISSN:2790-1688 Volume-14-(2025)

A Review of the Impact of Adversarial Attacks on Intrusion Detection Systems Based on Machine Learning and Deep Learning

Yanlong Li

China University of Mining and Technology, Xuzhou, China

362331696@qq.com

Abstract. As cybersecurity threats grow increasingly sophisticated, intrusion detection systems (IDS) have become pivotal in safeguarding networks and hosts. Machine learning (ML) and deep learning (DL) techniques have markedly enhanced the detection capabilities of host-based IDS (HIDS), particularly in addressing unknown and zero-day attacks. However, the opaque nature of these models renders them susceptible to adversarial attacks. This paper systematically reviews the adversarial attacks targeting ML/DL-based HIDS, their impacts on system performance (e.g., accuracy and efficiency), and available defense measures. Findings reveal that evasion, poisoning, and exploratory attacks pose significant threats, leading to reduced detection accuracy, increased false positive rates, and compromised model integrity. While defense strategies, including adversarial training, feature squeezing, and ensemble methods, demonstrate potential, their practical applicability remains to be validated. This study offers a comprehensive perspective on HIDS vulnerabilities to adversarial attacks and proposes future research directions, such as developing dedicated datasets and real-world validation, to bolster HIDS robustness.

Keywords: Intrusion Detection Systems; Machine Learning; Deep Learning; Adversarial Attacks;

1. Introduction

With the rapid advancement of computer technology, network security threats have become increasingly numerous and sophisticated, making intrusion detection systems (IDS) essential for safeguarding networks and hosts [1]. Host-based intrusion detection systems (HIDS) monitor host activities, such as system calls and log files, to detect malicious behaviors, and the integration of machine learning (ML) and deep learning (DL) techniques has significantly improved their detection efficiency, particularly for unknown and zero-day attacks [1]. However, these techniques' "black box" nature renders them vulnerable to adversarial attacks [2, 3]. Adversarial attacks, through carefully crafted perturbed inputs or poisoned training data, can mislead ML/DL models, resulting in reduced detection accuracy, increased false positive rates, and compromised system integrity. For instance, evasion attacks can disguise malicious activities as benign, while poisoning attacks can corrupt the training process of models [4, 5].

Although the impact of adversarial attacks on IDS has gained increasing attention, research specifically targeting HIDS remains limited, and comprehensive reviews are particularly scarce [2]. Therefore, this paper aims to systematically review the vulnerabilities of ML/DL-based IDS, with a focus on HIDS, to adversarial attacks, addressing the following research questions: (1) What types of adversarial attacks target ML/DL-based IDS? (2) What are the impacts of these attacks on IDS performance, such as accuracy and efficiency? (3) What defense measures can mitigate the effects of these adversarial attacks? By synthesizing relevant literature, this study provides in-depth insights into the characteristics, impacts, and defense strategies of adversarial attacks, offering theoretical support for enhancing the adversarial robustness of IDS and identifying directions for future research.

2. Types of Adversarial Attacks on ML/DL-based HIDS

2.1 Evasion Attacks

Evasion attacks occur during the host-based intrusion detection systems (HIDS) inference phase. Attackers generate adversarial examples by introducing subtle perturbations to malicious inputs, causing them to be misclassified as benign. These perturbations typically do not alter the functionality of the input but are sufficient to deceive the model. For instance, in HIDS, attackers may modify system call sequences to make malicious processes appear benign.

Common evasion techniques include gradient-based attacks such as the Fast Gradient Sign Method (FGSM) and Jacobian-based Saliency Map Attack (JSMA). Alotaibi and Rassam noted that adversarial examples generated by Generative Adversarial Networks (GANs) can reduce detection rates to near 0% on the CICIDS2017 dataset [2]. Ayub et al. demonstrated that JSMA attacks can degrade the accuracy of multilayer perceptron (MLP) models from 99.5% to 77.48% [5].

Evasion attacks pose a significant threat to HIDS, as these systems rely on precisely classifying system calls or logs. Satılmış et al. [1] highlighted that HIDS, commonly used to detect anomalies in system call sequences, can be significantly disrupted by adversarial examples.

2.2 Poisoning Attacks

Poisoning attacks compromise the training data of machine learning and deep learning models, undermining the learning process and leading to degraded performance or biased decision-making. These attacks can be targeted, causing the model to misclassify specific inputs, or non-targeted, reducing overall performance. The training data of host-based intrusion detection systems (HIDS), such as system call sequences in the ADFA-LD dataset, may be contaminated with maliciously injected data, impairing the model's ability to detect malicious behavior accurately. Corona et al. [4] noted that poisoning attacks can significantly increase the false negative rate, thereby reducing the reliability of HIDS.

Such attacks are particularly insidious as they fundamentally undermine the model, affecting its long-term detection capabilities. Apruzzese et al. [3] emphasized that while poisoning attacks may be constrained in real-world scenarios due to limited access to training data, they still pose a significant threat.

2.3 Exploratory Attacks

Exploratory attacks aim to extract model information, such as parameters or decision boundaries, to support subsequent attacks. A standard method is model extraction, where attackers construct a substitute model to mimic the host-based intrusion detection system (HIDS) behavior.

In HIDS, attackers may query APIs to extract model parameters, enabling the design of more effective evasion attacks. Akhtar and Mian [6] discussed model extraction techniques in computer vision, which can be extended to the HIDS domain. While exploratory attacks do not directly impair detection performance, they may leak sensitive information, increasing the likelihood of successful future attacks.

2.4 Other Attack Types

Other attack types include denial of service attacks (DoS attacks), where attackers generate many false alerts or exhaust system resources to disrupt the normal operation of host-based intrusion detection systems (HIDS). DoS attacks can slow down HIDS processing through algorithmic complexity attacks; Response manipulation attacks, where attackers exploit the response mechanisms of HIDS to trigger erroneous security measures, such as mistakenly blocking legitimate traffic; and hijacking attacks, where attackers leverage HIDS responses to target other systems or users, manipulating the system's behavior to facilitate secondary attacks.

Corona et al. provided a detailed classification of adversarial attacks on intrusion detection systems (IDS), encompassing evasion, poisoning, exploratory, DoS, response manipulation, and hijacking attacks, laying a foundation for understanding the threat landscape [4]. Apruzzese et al. further emphasized the importance of realistic attack modeling, highlighting that attacks must adhere to practical constraints (e.g., protocol consistency) to ensure feasibility [3].

Table 1: Summary of Attack Types

Attack Type	Description	Relevance to HIDS	Example Studies
Evasion Attacks	Injection of Adversarial Examples During the Testing Phase, Leading to Misclassification of Malicious as Normal	Perturbation of System Call Sequences or Logs	[2], [5]
Poisoning Attacks	Contaminate training data to degrade model performance or induce biased decision-making.	Compromise HIDS by polluting system call training data, impairing the model's ability to accurately detect malicious behavior.	[2], [4]
Exploratory Attacks	Extract model information, such as parameters or decision boundaries, to support subsequent attacks.	Enable attackers to extract HIDS model parameters, facilitating the design of targeted attacks like evasion.	[2], [6]
DoS Attacks	Generate a large volume of false alerts or exhaust system resources.	Disrupt HIDS functionality, rendering it unable to operate normally.	[4]
Response Manipulation Attacks	Exploit HIDS response mechanisms to trigger erroneous security measures.	Cause HIDS to mistakenly block legitimate traffic, undermining system reliability.	[4]
Hijacking Attacks	Leverage HIDS responses to attack other systems or users.	Exploit HIDS response mechanisms to facilitate secondary attacks on other systems.	[4]

3. Adversarial Attack Impacts on HIDS Performance

Adversarial attacks can significantly impair the performance of machine learning (ML) and deep learning (DL)-based host-based intrusion detection systems (HIDS), affecting key metrics such as detection accuracy, false positive rate, and model generalization capability.

3.1 Reduction in Detection Accuracy

Evasion attacks reduce the true positive rate (TPR) by generating adversarial examples. Apruzzese et al. [3] demonstrated that evasion attacks based on realistic modeling can reduce the detection rate of NIDS from 99.9% to 70%, and HIDS may face similar threats. For instance, perturbing system call sequences may cause malicious processes to go undetected. Ayub et al. [5] showed how JSMA attacks can reduce the accuracy of MLP models from 99.5% to 77.48%.

Poisoning attacks contaminate training data and render models unable to distinguish between benign and malicious activities, thereby reducing overall accuracy. Alotaibi and Rassam [2] pointed out that poisoning attacks may cause the detection rate to drop to nearly 0%.

3.2 Increasing False Positive Rate

Adversarial attacks may cause HIDS to misclassify normal activities as malicious, thereby increasing the false positive rate (FPR). A high FPR can lead to alert fatigue, reduce system credibility, and improve the cost of manual investigation. Satılmış et al. [1] mentioned that the high FPR of HIDS is a common challenge, and adversarial attacks may exacerbate this problem. For example, overstimulation attacks interfere with security administrators' operations by generating false alerts [4].

3.3 Decline in Model Generalizability

Poisoning attacks may cause models to overfit to contaminated data, thereby reducing their generalization ability to new attacks. This is particularly critical for HIDS, as it needs to detect zero-day attacks. Corona et al. [4] pointed out that poisoning attacks may significantly undermine the long-term effectiveness of models.

3.4 The destruction of model integrity

Exploratory attacks may leak information about model parameters or training data, supporting more sophisticated attacks. For example, extracting the parameters of HIDS models may help attackers design targeted evasion attacks. Akhtar and Mian [6] mentioned that model extraction attacks can generate more effective adversarial examples.

Table 2:Performance Impact Summary

Impact	Description	HIDS Relevance	Example Studies
Reduced detection accuracy	Evasion and poisoning attacks decrease TPR	Malicious system calls go undetected	[2], [3], [5]
Increased false positive rate	Normal activities are misclassified as malicious	Increases alert fatigue	[1], [2], [4]
Diminished generalization ability	Poisoning attacks lead to overfitting	Inability to detect new attacks	[2], [4]
Compromised integrity	Information leakage supports subsequent attacks	Parameter leakage threatens security	[2], [6]

4. Defense Measures Against Adversarial Attacks

Various defense strategies have been developed to mitigate the threats posed by adversarial attacks and enhance the robustness of ML/DL-based HIDS. These strategies can be categorized into two types: modifying the training process or input data, and adding additional networks.

4.1 Modification of Training Process or Input Data

Modifying the training process or input data aims to strengthen HIDS by addressing vulnerabilities. Incorporating adversarial examples into the training process, known as adversarial training, enhances the model's resistance to attacks. Alotaibi and Rassam [2] demonstrated that this approach, exemplified by the ZK-GanDef defense, can improve detection accuracy by up to 49.17%, though it often requires significant computational resources. Similarly, feature compression reduces the precision of input features to smooth the input space, making it harder for attackers to generate effective perturbations, particularly for high-dimensional data like system call sequences. To improve model generalization, data augmentation techniques, such as the MGAN module in Def-IDS proposed by Wang et al. [7], generate additional training data using GANs, enabling better handling of diverse intrusion patterns.

The Matrix Estimation Network (ME-Net) enhances robustness by removing noise through matrix estimation. At the same time, Deep Image Prior Defense (DIPDefend) reconstructs inputs to eliminate

adversarial perturbations, offering a viable solution for scenarios without pre-training [2]. These methods fortify HIDS by improving resilience during training and data processing, though their computational demands necessitate careful consideration for real-time applications.

4.2 Additional Networks

Incorporating additional networks provides a robust defense by detecting or mitigating adversarial inputs. For instance, the Adversary Detection Network trains a binary classifier to distinguish between regular and adversarial inputs, offering a proactive approach to identifying threats [2]. Ensemble methods, such as Def-IDS proposed by Wang et al. [7], combine multiple models to make decisions through a voting mechanism, enhancing robustness against attacks like FGSM and JSMA. Similarly, APE-GAN leverages GANs to remove adversarial perturbations from inputs, providing an effective defense without requiring detailed model knowledge [2]. Another approach, dropout, randomly discards neural network units to detect adversarial examples, proving particularly effective against CW attacks [2]. These network-based strategies enhance HIDS resilience by diversifying decision-making processes and isolating adversarial inputs, though their complexity may challenge resource-constrained environments.

4.3 Other Defense Strategies

Beyond modifying training processes and adding networks, complementary strategies further strengthen HIDS defenses. Restricting attacker access to training data and model parameters, such as by deploying HIDS on secure remote servers, reduces the feasibility of white-box and gray-box attacks [3]. Selecting robust metrics that capture invariant features of intrusion activities improves detection reliability, while incorporating contextual information from hosts and services helps verify alerts, thereby reducing false positives [4]. These approaches enhance the overall security posture of HIDS by addressing technical and operational vulnerabilities, ensuring a more comprehensive defense against adversarial threats.

Table 3: Summary of Defense Measures

Defense Measures	Description	HIDS Relevance	Example Studies
Adversarial training	Incorporating adversarial examples to enhance robustness	Improving the robustness of system call classification	[2]
Feature compression	Reducing feature precision to smooth the input space	Mitigating the impact of system call perturbations	[2]
Def-IDS	Integrating multiple models for decision-making via voting mechanisms	Enhancing HIDS robustness	[7]
APE-GAN	Using GAN to remove adversarial perturbations	Applicable to log data	[2]
Access restriction	Protecting training data and model parameters	Reducing white-box attacks	[3]
Robust metrics	Capturing invariant features	Improving detection reliability	[4]
Contextual information	Verifying alerts to reduce false positives	Enhancing detection accuracy	[4]

5. Limitations and Future Directions

5.1 Limitations of current research

Current research on adversarial attacks against ML/DL-based HIDS faces significant limitations, which can be grouped into algorithmic, practical, and theoretical challenges. Algorithmically, studies

often focus on white-box attack scenarios, assuming full attacker knowledge of model structure and parameters, while neglecting black-box or gray-box scenarios, which limits applicability to real-world settings where attacker knowledge varies [3]. Additionally, defense mechanisms struggle to generalize across diverse attack types, with Wang et al.'s research indicating that some defenses fail against new or modified attacks [7]. The lack of standardized evaluation metrics and datasets further complicates comparisons, as different experimental setups lead to inconsistent results [2]. Practically, high computational costs hinder real-time applications; for instance, Alotaibi and Rassam's analysis showed that Def-IDS retraining requires 115.1 seconds, making it unsuitable for time-sensitive environments [2]. Moreover, reliance on outdated datasets like ADFA-LD and CICIDS2017, as noted by Satılmış et al. [1], disconnects research from modern threat landscapes, reducing practical relevance. Theoretically, game-theoretic approaches face challenges, as calculating Nash equilibria is computationally infeasible and relies on unrealistic assumptions about optimal attacker strategies, limiting their utility in dynamic adversarial scenarios [1]. Other limitations, such as insufficient focus on HIDS-specific data characteristics, further exacerbate these challenges, underscoring the need for more targeted research.

5.2 Potential Future Research Directions

To address these limitations, future research should focus on targeted directions that align with the identified algorithmic, practical, and theoretical challenges. Developing robust defense mechanisms that generalize across black-box, gray-box, and white-box attack scenarios is critical to tackle algorithmic limitations. For instance, Wang et al.'s work suggests designing adaptive defenses, such as enhanced ensemble methods, to counter diverse attack variants [7]. Establishing standardized evaluation frameworks, including unified datasets like CSE-CIC-IDS2018 and consistent metrics, will address the lack of uniformity, enabling reliable comparisons across studies [2]. For practical challenges, optimizing computational efficiency is essential to enable real-time HIDS applications; improving algorithms like JSMA or Def-IDS to reduce iterations, as proposed by Wang et al. [7], could make defenses more viable.

Additionally, developing HIDS-specific datasets that reflect current host-level threats, such as system call sequences, will bridge the gap between research and real-world applications, addressing the outdated dataset issue noted by Satılmış et al. [1]. For theoretical challenges, exploring computationally feasible game-theoretic methods, particularly for repeated adversarial scenarios, can overcome the limitations of Nash equilibrium calculations and unrealistic assumptions, as highlighted by Satılmış et al. [1]. Broader applications of adversarial classification frameworks to fields like fraud detection can further enhance generalizability [1]. These directions directly address the identified limitations, ensuring that advancements in HIDS research are robust and practical.

6. Conclusion

This paper systematically reviews the vulnerabilities and potential of machine learning (ML)- and deep learning (DL)-based intrusion detection systems (IDS), particularly host-based IDS (HIDS), under adversarial attacks. By analyzing the types of adversarial attacks, their impacts on IDS performance, and existing defense measures, we reveal the important role of IDS in network security and the challenges they face. Research shows that although defense mechanisms such as Def-IDS have improved robustness through dataset augmentation and adversarial retraining, they are still limited by issues such as the singularity of attack scenarios, insufficient generality of defenses, high computational costs, lack of unified evaluation methods, and inadequate testing in practical applications.

To address these challenges, future research should focus on expanding attack scenarios, developing robust defense mechanisms, standardizing evaluation methods, optimizing computational efficiency, testing in real-world scenarios, deeply understanding adversarial examples, and applying

adversarial classification frameworks to a wider range of fields. These efforts will help enhance the adversarial robustness of IDS and provide more reliable guarantees for network security.

This paper provides an important reference for research on the adversarial robustness of IDS and lays the foundation for the development of more secure and reliable intrusion detection systems. As network threats continue to evolve, continuous innovation and cross-disciplinary collaboration will be key to ensuring the security of the digital environment.

References

[1] Satilmiş H, Akleylek S, Tok Z Y. A systematic literature review on host-based intrusion detection systems[J]. IEEE Access, 2024, 12: 27237-27266.

[2] Alotaibi A, Rassam M A. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense[J]. Future Internet, 2023, 15(2): 62.

[3] Apruzzese G, Andreolini M, Ferretti L, et al. Modeling realistic adversarial attacks against network intrusion detection systems[J]. Digital Threats: Research and Practice (DTRAP), 2022, 3(3): 1-19.

[4] Corona I, Giacinto G, Roli F. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues[J]. Information sciences, 2013, 239: 201-225.

[5] Ayub M A, Johnson W A, Talbert D A, et al. Model evasion attack on intrusion detection systems using adversarial machine learning[C]//2020 54th annual conference on information sciences and systems (CISS). IEEE, 2020: 1-6.

[6] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410-14430.

[7] Wang J, Pan J, AlQerm I, et al. Def-ids: An ensemble defense mechanism against adversarial attacks for deep learning-based network intrusion detection[C]//2021 International Conference on Computer Communications and Networks (ICCCN). IEEE, 2021: 1-9.