AI-Based Mitigation Strategies for Evasion Attacks in Intrusion Detection Systems

Intisar Adnan Hasan[1] and Abdulbasit AL Azzawi[2]
[1]*Department of Computer Science, University of Diyala, 32001 Baqubah, Iraq*
[2]*Department of Computer Science, College of Science, University of Diyala, 32001 Baqubah, Iraq*
antasaralansary@gmail.com, dr.abdulbasit@uodiyala.edu.iq

Keywords: Evasion Attacks, Robustness, Adversarial Machine Learning, Advanced AI, Mitigation.

Abstract: The high rate of artificial intelligence (AI) has had a major impact on Enhancing the intrusion detection systems (IDS). However, these systems remain highly vulnerable to adversarial evasion attacks, in which adversarial inputs are intentionally designed to avoid detection systems. This research paper is a systematic review of recent studies on evasion attacks on intrusion detection systems, and aiming to identify to the popular attack methodology, most widely datasets, and up-to-date defences. The results demonstrate that gradient-based attacks, in particular Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), are the most widely used methods of testing system robustness. In addition, it can be seen that there is a strong reliance on standard benchmark datasets, while real-world evaluation scenarios are minimal. Despite the wide range of defences mechanisms, most of them have a hard time keeping pace with adaptive and advanced adversarial threat. This review points out the gaps in current studies that are critical such as a lack of a variety of evaluation conditions and more adaptive and robust detection strategies. The research provides on the existing research issues and outlines key trends to pursue by future research, which could help create more robust and resilient cybersecurity.

1 INTRODUCTION

Intrusion detection systems made on AI have experienced tremendous growth [1]. Because of its significance in identifying the malicious properties and protecting cyber systems against intrusions [2], Evasion attacks are extremely harmful, and they threaten the adversarial machine learning in which an attacker purposely distorts malicious input to induce errors in classification models to label malicious activities as lawful and harmless [3].

In the majority of such attacks, the attackers will design examples where the desired class is purposefully manipulated so as to generate an example that is expected to have a specific class [4]. Practically, the evasion attacks are usually based on exploiting the mutable input properties, especially in network-based attacks where the attacker alters the sizes of packets, the timing aspects of data traffic, or the level of flow patterns in such a manner that enables the malicious input traffic to wriggle around the ML-based intrusion detection systems. These minor adjustments change the images of the properties that the model identifies and carries the original malicious behavior of the attack, therefore, deceiving the classifier without interfering with the very attack logic [5].

Besides interfering with feature levels, malicious evasion attacks take advantage of the vulnerabilities of ML models, including sensitivity to decision boundaries and generalization limitations. With the help of input data statistics manipulation, an attacker can mislead the reasoning processes of the model without destroying the integrity of the system and altering the internal elements of the model [6].

As the number of modern AI-driven applications is on the rise, the question of the strength and reliability of AI systems has been raised. Of all the other security threats, evasion attacks are the most alarming since they are performed at the inference stage and they modify the input environment without modifying the training set or the model architecture [7]. Consequently, a number of AI-driven systems, such as IDS [8]. the malware detection systems [9],The fraud detecter systems [10]. The autonomous driving systems [11], And the medical image analysis systems [12], Have been more susceptible to evasion attacks, in which

malicious actors use model vulnerabilities to avoid detection systems [13].

Evasion attacks pose a serious threat to the ML-based IDS by adversely affecting the detection performance by deploying adversarial-designed inputs [14]. These externally introduced inputs are not retrained and reconfigured into the system affect the decision-making process of the model, and eventually results in classifying malicious activities as legitimate traffic, Although the research in adversarial machine learning has advanced at a very fast pace, the current AI-based security detection systems are not robust enough to with stand the practical evasion conditions. Most of the suggested defenses are tested on limited assumptions, and tend to deteriorate when dealing with adaptive attackers who can adapt perturbations to them [15]. This gap between the strength of experiments and reliability in the real world reveals a fundamental weakness of existing evasion mitigation strategies and demonstrates the importance of developing more efficient and robust mitigation strategies. In the successful mitigation of evasion attacks using AI-based security systems at real-world settings, The recent progress in the use of AI in security-critical settings has further emphasized the maliciousness of evasion attacks since such attacks can disrupt the reliability of the system by causing the model to make erroneous decisions without altering the model parameters or training data [16]. Besides, evasion attacks are not specific to a particular data modality, but have been shown to be effective on network traffic, textual and visual inputs. In these areas, invisible noise is able to fool the DL models because its relies on the high dimensional feature representations [17].

This paper seeks to undertake a systematic review of the latest literature that looks at mitigation techniques that can be used to counter evasion attacks in AI-based IDS. The review aims at discussing different types of defensive technologies, comparing the performance of the models in application, and the existing research problems and the future of this area.

1.1 Theoretical Background

The Intrusion Detection Systems IDS based on AI are a significant element of contemporary cybersecurity. Conventional intrusion detection methods, which include rule-based or signature-based IDS, are the basic elements of cybersecurity.

Nevertheless, they heavily depend on established rules or signatures. prevents them from identifying new or advanced attacks, and leaves them. companies susceptible to new risks. As cyber threats grow more complicated, the necessity of more sophisticated, AI-based solutions emerges imperative[18]. There has been an evolution of IDS. drastically as time passes. The first one is based on the simple rule-based. mechanisms, now they implement the latest ML. Detect patterns, anomalies and threats in ML algorithms network traffic [19]. Due to the rapid development of the AI discipline, ML and DL tools have become prevalent in augmenting the capabilities of the IDS. Support vector machines (SVMs), k-nearest neighbors (k-NNs), decision trees, and random forests are some of the ML algorithms that have been used to categorize network traffic, as well as identify malicious activity. DL architecture, including CNN, LSTM ana Transformer – based model have proven particulary effective at learning intricate representation whithin large network datasets [20], [21].

However, despite noteable progress in AI based IDS, Evasion attacks remain acritical challenge, When attackers make carefully changes to network data in order to bypass detection.

Several techniques have been developed to generate these malicious examples, such as FGSM, PGD, BIM, and C&W, which exploit the models' sensitivity to minor data changes [7], [22].

To address this challenge, numerous studies have proposed mitigation strategies designed to enhance the robustness of AI-driven IDS against evasion attacks. The mitigation strategies proposed in the literature include adversarial training, ensemble learning, feature transformation techniques, robust optimization approaches, and hybrid DL models. Adversarial training is one of the most prominent defense mechanisms, where adversarial examples are incorporated during the training process to improve the model’s resistance to such attacks [23].

This will help in the creation of more effective cybersecurity measures that are in a better position to evolve with the sophisticated threats.

1.2 Research Questions

During a Systematic Literature Review (SLR), research questions are the key element as they define the field of study and provide a roadmap to adhere to when collecting, analysing, and synthesising the relevant studies. In accordance with the goals of this review, the following research questions will be created:

- RQ1. What are the most prevalent evasion attacks on the AI-based Intrusion Detection Systems?

- RQ2. What machine learning and deep learning methods detect or defend.
- RQ3. Which datasets and evaluation measures are popular to evaluate the efficiency of IDS models against evasion attacks?
- RQ4. What are the key challenges and research gaps in protecting AI-based IDS to evasion attacks?

1.3 Aim of the Review

This study aims to systematically analyze recent research on evasion attacks against AI-based intrusion detection systems, focusing on identifying common attack techniques, employed datasets, evaluation methods, and existing research gap.

2 METHODOLOGY

This paper will involve a systematic review of the recent studies in the context of mitigating evasion attacks in the AI-based intrusion detection systems. The reviewed articles are dated between 2021 and 2026, and the discussion is aimed at examining the current trends in machine learning and deep learning algorithms to find the most efficient techniques of detecting cyberattacks and counteracting adversarial sampling.

2.1 Search Strategy

The keywords that were used in the search of the various databases had the following forms:

1) IDS;
2) Evasion Attack or Adversarial Attack;
3) AI, ML, or DL;
4) Cybersecurity or Network Security.

2.2 Inclusion and Exclusion Criteria

Studies were selected based on predefined inclusion and exclusion criteria. Eligible studies were those published between 2021 and 2026 that focused on adversarial evasion attacks against AI-based intrusion detection systems and employed ML, DL, or hybrid techniques. Only peer-reviewed journal and conference papers written in English were considered.

Studies were excluded if they were duplicates, unrelated to intrusion detection systems, did not address evasion attacks, or lacked sufficient experimental evaluation. After the screening process, 20 studies were selected from the initial 42 identified records and included in the final review.

2.3 Data Analysis

The extracted data underwent qualitative analysis to explore modern research trends, the most commonly used datasets, the most popular ML techniques, and proposed mitigation methods to counter evasion attacks. This analysis helped clarify the advantages and limitations of current methods, as well as reveal research gaps that require further exploration in future studies.

2.4 Reporting

The results of this systematic review were organized and presented based on the PRISMA guidelines, which provide a structured methodological framework for identifying, examining, and selecting relevant studies in a transparent and systematic way [24].

2.5 Study Results

The initial search identified 42 studies. After screening titles and abstracts, 22 studies were excluded. The remaining 20 studies satisfied the inclusion criteria and were included in the final review. Figure 1 illustrates the corresponding PRISMA-based study selection process, including the number of records excluded at each stage.

Figure 2 presents the distribution of the reviewed studies by publication year. The number of publications increased gradually over the study period, rising from one study in 2021 and one study in 2022 to five studies in 2023 and eight studies in 2024. A high level of research activity continued in 2025 with seven studies, while two studies were published in 2026. This trend reflects the growing research interest in adversarial evasion attacks and their mitigation in AI-based IDS.

Figure 3 illustrates the distribution of adversarial evasion attack methods used in the reviewed studies. FGSM was the most frequently investigated attack, appearing in 13 studies, followed by PGD and C&W attacks, each reported in 9 studies. BIM and JSMA were used less frequently, appearing in 4 studies each. These results indicate that gradient-based attacks remain the dominant approach for evaluating the robustness of AI-based IDS.

Figure 4 shows the distribution of datasets used in the reviewed studies. CIC-IDS2017 was the most frequently used dataset, appearing in 10 studies,

followed by CIC-IDS2018 in 6 studies. NSL-KDD, UNSW-NB15, and CIC-DDoS-2019 were each used in 5 studies, while Kitsune appeared in 2 studies. Other datasets, including InSDN, ECU-IoFT, ISCX-2014, Smart Grid, ICS/SCADA, MNIST, CIFAR-10, and ImageNet, were used less frequently. These findings indicate a strong reliance on benchmark datasets for evaluating IDS against adversarial evasion attacks.

Figure 5 shows the distribution of ML and DL models used in the reviewed studies. CNN was the most commonly used model, appearing in 9 studies, followed by DNN and LSTM in 8 and 6 studies, respectively. Other models, including MLP/FNN, Random Forest, Decision Tree, and SVM, were also frequently employed. In contrast, approaches such as Transformer, GNN, XGBoost, and Naive Bayes appeared only in a limited number of studies. These results highlight the widespread use of DL models in intrusion detection and evasion attack mitigation research.

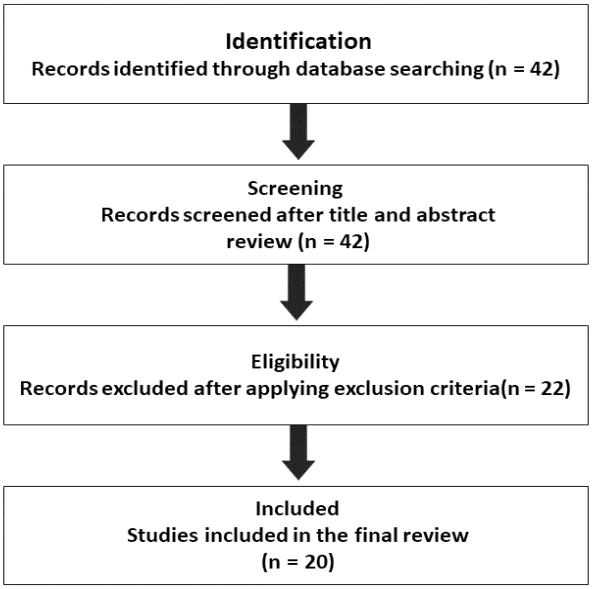

Figure 1: PRISMA flow diagram of study selection process.

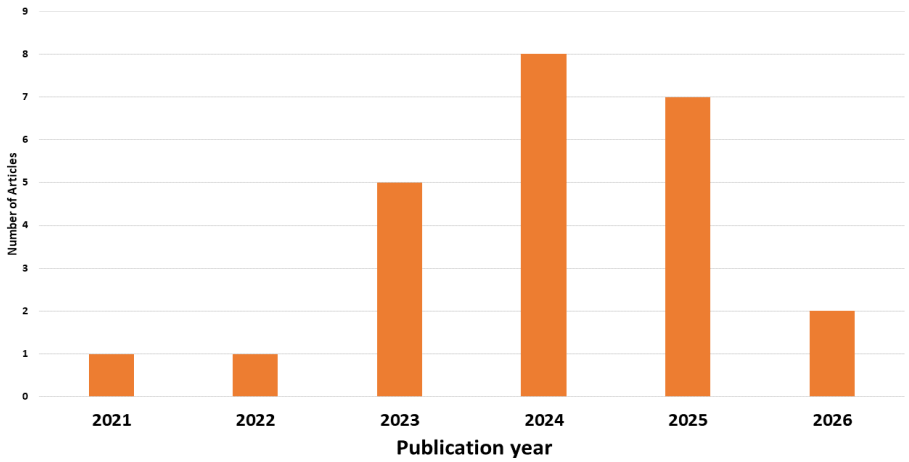

Figure 2: Temporal distribution of selected articles by publication year (2021_2026).

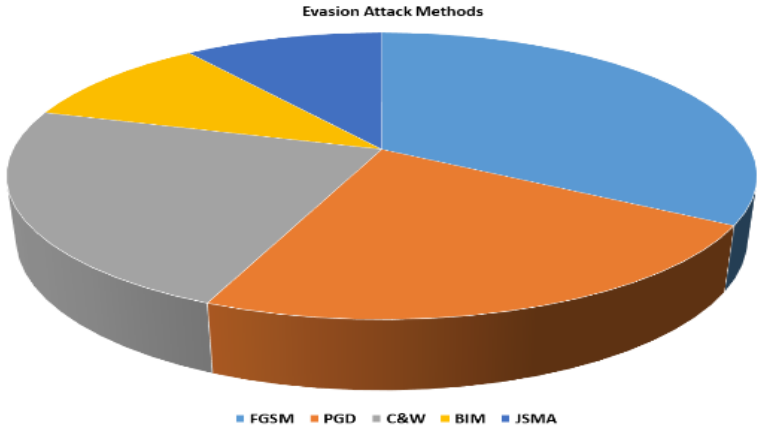

Figure 3: Distribution of adversarial evasion attack methods across studies.

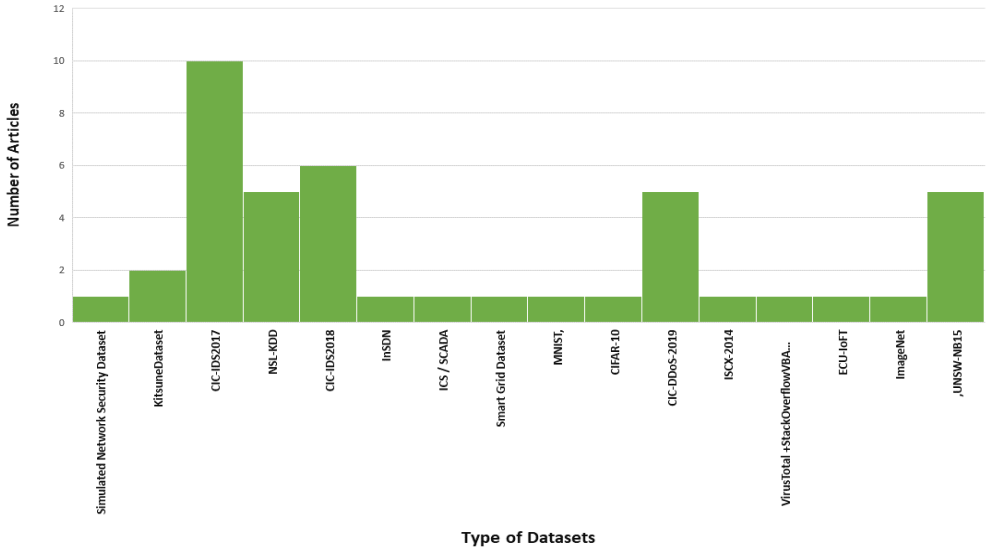

Figure 4: Comparative of dataset types used in the literature.

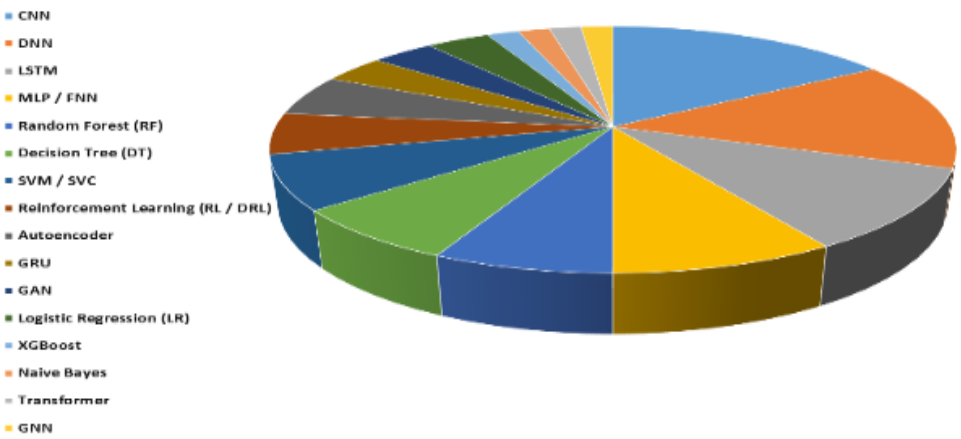

Figure 5: Classification of ML and DL models across reviewed studies.

3 STUDY GOAL

The objective of this systematic review is to review and summarize the literature that has been done on evasion attacks against security systems based on artificial intelligence.

The studies reviewed are aimed at defining popular methods of adversarial attacks, popular datasets, and the most significant issues which relate to the literature. Besides, this review also brings out the current research trends and offers insights towards creating a stronger and more reliable cybersecurity system.

3.1 Dataset Characteristics

The studies in analysis implement various datasets in terms of size, structure and design of experiments. Normal and malicious network traffic is utilized to simulate real-life situations in most of the research. However, the labeling of data, form, and categories may vary and this may affect the comparability and generalizability of the results.

3.2 Evasion Attack Methods

Research on adversarial evasion attacks has investigated various techniques, including FGSM, PGD, BIM, Carlini and Wagner (C&W), DeepFool, JSMA, ZOO, HSJA, GAN-based attacks, and timing-based evasion methods. These approaches aim to modify input data or network traffic characteristics in a way that deceives intrusion detection systems while maintaining the original attack objectives (RQ1).

Among the reviewed studies, FGSM and PGD are identified as the most commonly applied gradient-based evasion attacks, as demonstrated in Figure 3. Their popularity is mainly attributed to their simplicity, effectiveness, and adaptability across different machine learning and deep learning models. However, more advanced techniques, including decision-based attacks, GAN-based approaches, and timing-based evasion strategies, have also gained attention due to their ability to generate more adaptive adversarial examples.

3.3 Analysis of Detection Models and Defense Approaches

The reviewed studies demonstrate that deep learning models are the dominant approaches for intrusion detection and adversarial robustness evaluation. CNN, DNN, and LSTM architectures are among the most frequently applied models, followed by hybrid approaches that combine multiple deep learning techniques. Figure 4 illustrates the distribution of machine learning and deep learning models used in intrusion detection research.

In addition to conventional deep learning architectures, several studies employ advanced approaches, including GANs, reinforcement learning, autoencoders, attention mechanisms, and Transformer-based models, to improve detection performance and enhance resistance against adversarial attacks. Furthermore, multi-model and ensemble frameworks indicate a transition toward more adaptive and robust intrusion detection architectures.

Overall, the literature shows a clear movement from traditional single-model solutions toward more complex hybrid defense strategies. The findings presented in Table 1 address RQ2 by demonstrating that CNN, LSTM, DNN, and hybrid deep learning models represent the most commonly used detection and mitigation approaches.

3.4 Analysis of Datasets

This section analyzes the datasets used in the reviewed studies to identify the most common benchmarks, their characteristics, and their suitability for evaluating adversarial evasion attacks against AI-based intrusion detection systems. Several standard cybersecurity datasets are frequently employed, including CIC-IDS2017, NSL-KDD, UNSW-NB15, and CIC-DDoS2019. In addition, some studies utilize domain-specific datasets related to IoT and smart grid environments, reflecting the diversity of experimental settings (RQ3).

Table 2 analysis reveals that CIC-IDS2017, NSL-KDD, UNSW-NB15, and CIC-DDoS2019 are the most frequently used because they are available and have a standard format. Nevertheless, there are few studies that are based on domain or real world data, including ECU-IoFT, IEC 60870-5-104-based data and Irish Smart Energy Trials, which suggests that there is a lack of realistic availability. Also, there are non-network datasets which are used to test adversarial robustness in general settings, not general intrusion settings, e.g. MNIST, CIFAR-10, ImageNet. The majority of the studies are aimed at enhancing IDS resilience to adversarial evasion using hybrid models, adversarial training, and GAN based methods.

In general, the results demonstrate that benchmark datasets are a more preferred choice, and more realistic and domain-specific assessments are required.

4 RESEARCH GAPS

Although the number of studies researching evasion attacks against AI-based intrusion detection systems is rather high, a range of critical research gaps still exist. To begin with, the majority of the available literature discusses smaller models of attacks, including FGSM and PGD, and pays little or no attention to the adaptive and more advanced evasion attacks. Second, most studies are based on benchmark data sets and little has been done to test them on real world and domain specific settings. Third, single-model architecture is the basis of most of these approaches with more sophisticated hybrid models that integrate multiple techniques that can further improve performance of detection. Lastly, the effectiveness of a variety of the defense strategies can be demonstrated to be robust in the controlled experimental environment but their stability in the conditions of the real world and dynamic environment is not explored enough. This section focuses on (RQ4).

5 RESULTS AND DISCUSSION

The results highlight the importance of addressing adversarial evasion attacks in AI-based intrusion detection systems. The obtained results suggest that most of the existing literature is focused on typical attack techniques, such as FGSM and PGD, and therefore the study is focused on gradient-based attacks. More dynamic and better attacks are however less mature and this may limit the extrapolation of the current defensive measures.

Table 1: Summary of adversarial evasion attacks and detection models in the reviewed studies.

Ref	Author(s)&year	Attack Method	Model
[25]	Yam Sharon et al., 2021	TANTRA Timing-based evasion attacks	LSTM-based of DNN
[26]	Islam Debicha et al., 2022	FGSM, PGD, Deep Fool, C&W	Deep Neural Network IDS + Transfer Learning + Multiple Adversarial Detectors
[27]	Yuan et al., 2023	FGSM, BIM, Deep Fool, C&W	DL model, FNN,CNN ML model, Label Spreading Adversarial detector, LID + SVM classifier
[28]	Sarıkaya et al., 2023	PGD, WGAN, CTGAN	RAIDS framework
[29]	Asimopoulos et al., 2023	FGSM , CTGAN adversarial attacks	DRL-DDQN (CNN, GRU, FFNN)
[30]	El-Toukhy et al., 2023	DRL Evasion + FGSM	DRL-DDQN (CNN, GRU, FFNN)
[31]	Fukuda et al.2023	FGSM, PGD	MLP, CNN (VGG-11)
[32]	Roshan et al., 2024	Carlini &Wagner (C&W)	(DNN)
[33]	Khushnaseeb et al.,2024	FGSM	DNN
[13]	Paya et al., 2024	ZOO, HSJA, W-GAN based attack	Apollon (multiple classifiers + MAB with Thompson Sampling + K-Means clustering)
[34]	Randhawa et al.2024	GAN + Deep Reinforcement Learning	RELEVAGAN(EVAGAN + DRL
[35]	Barik et al., 2024	FGSM, PGD, JSMA, C&W	WCSAN + PSO + SVC
[36]	Mimura & Kurashina, 2024	Benign feature insertion into malicious macros	DNN, RNN, LSTM, Attention-LSTM, Decision Tree, Random Forest, SVM
[37]	Kousik Barik et al., 2025	JSMA, FGSM, C&W	CNN-based NIDS + Hybrid Defense (PGD +PIOA + SS)
[38]	Alazzawi et al., 2025	FGSM, BIM, C&W	RF, DT, SVM, LR + GAN
[7]	Raja et al., 2025	FGSM, PGD, C&W	ResNet50V2
[39]	Omer Fawzi et al., 2025	Adversarial Evasion Attack	ADCEN (DTCN + LSTM + GRU + Attention)
[40]	Yang et al., 2025	Adversarial ML attacks	Auto ML + ML/DL Models (CNN, LSTM, ML)
[41]	Mohammed Mustafa 2025	FGSM, PGD	CNN + Auto encoder (Hybrid DL)
[42]	Liyakat et al.2025	FGSM, PGD + synthetic attacks (GAN / AE)	Reinforcement Learning + Adversarial ML + Multi-Agent System

Table 2: Summary of datasets and study goals in the reviewed studies.

Ref	Author(s)&year	Datasets	Goal
[25]	Yam Sharon et al., 2021	Kitsune, CIC-IDS2017	Bypass NIDS by design a packet-time evasion attack.
[26]	Islam Debicha et al., 2022	NSL-KDD, CIC-IDS2017	Detect evasion attacks-based IDS using transfer learning.
[27]	Yuan et al., 2023	NSL-KDD, CICIDS2018	Enhance the resilience of deep learning-based IDS against evasion attacks by employing a hybrid detection framework.
[28]	Sarıkaya et al., 2023	InSDN, CICIDS2017	Develop a robust IDS resistant to an evasion attack.
[29]	Asimopoulos et al., 2023	IEC 60870-5-104 based dataset	To investigate the impact of FGSM and CTGAN
[30]	El-Toukhy et al., 2023	Irish Smart Energy Trials dataset	Develop a DRL-based model to detect electricity theft and defend against evasion attacks
[31]	Fukuda et al.2023	MNIST, CIFAR-10	Improve the robustness of quantized DL models against evasion attacks
[32]	Roshan et al., 2024	CIC-DDoS-2019	Improve robustness of NIDS against evasion attacks
[33]	Khushnaseeb et al.,2024	CICDDoS-2019	Impact of adversarial attacks on DL-based cyber-attack detection systems and evaluate attack transferability between models.
[13]	Paya et al., 2024	CIC-IDS-2017, CIC-IDS-2018, CIC-DDoS-2019	The objective of the research is to protect ML-based IDS from adversarial ML attacks that attempt to evade detection.
[34]	Randhawa et al.2024	ISCX-2014, CIC-2017, CIC-2018	Develop a DRL-GAN framework to generate adversarial traffic that evades ML-based botnet detection systems.
[35]	Barik et al., 2024	CIC-IDS2017	Detect adversarial attacks on IDS using the WCSAN-PSO framework
[36]	Mimura & Kurashina, 2024	VirusTotal +StackOverflowVBA macros	Evaluate the vulnerability of ML/DL-based malware detection systems to evasion attacks by inserting benign features into malicious samples.
[37]	Kousik Barik et al., 2025	CSE-CIC-IDS2018	Improve the robustness of DL-based NIDS against adversarial attacks.
[38]	Alazzawi et al., 2025	ECU- IoFT, CICIDS2018	Improve IoFT IDS robustness using GAN and adversarial training to detect evasion attacks.
[7]	Raja et al.,2025	ImageNet	Analyzing evasion attacks on ML models and proposing defense methods
[39]	Omer Fawzi Awad et al.,2025	CIC-IDS-2017	Develop a robust IDS capable of detecting attacks and improving resistance against evasion attacks in network traffic.
[40]	Yang et al., 2025	CICIDS2017, NSL-KDD, UNSW-NB15	Development of Automated cybersecurity frameworks for Zero Touch Networks using of Automated ML and defending against adversarial attack.
[41]	Mohammed Musthafa2025	CIC-IDS2017, UNSW-NB15	Enhancing the strength of AI-based IDS against evasion attacks and simultaneously maintaining high detection accuracy in real-time settings.
[42]	Liyakat et al.2025	Cyber-Range simulated environment	Design a machine learning-powered platform to mimic pseudo hacking in order to upgrade cyber-protection.

The deep learning models such as CNN and LSTM are also quite popular in detection models since they can extract the spatial as well as temporal characteristics of network traffic. Recent evidence also suggests the growing appeal of hybrid models where a large number of architectures are combined to increase resiliency. Despite these improvements, a great number of models are tested in an artificial environment and may not always be applicable in a real attack.

The analysis of the dataset suggests that the use of such benchmark datasets as CIC-IDS2017, NSL-KDD, and UNSW-NB15 is regular. Inasmuch as such datasets do offer a general assessment configuration, they may not accurately reflect the complexity of systems in the real world, specifically, in new domains such as IoFT and UAV networks.

More than that, although various mitigation techniques such as adversarial training, GAN-based models, and ensemble models have been proposed, their susceptibility to adaptive adversaries has been a significant problem. Majority of the defense mechanisms are highly efficient in countering the attacks that have been experienced before and are inefficient in countering attacks that are unknown or their strategies are changing dynamically.

On the whole, the findings suggest that despite the fact that much has been done, it is still necessary to develop stronger, versatile, and dynamic systems of defense that could work in practice in the real world.

6 CONCLUSIONS

The systematic review has analyzed the recent works regarding the study of adversarial evasion attacks to AI-based intrusion detection systems. The findings demonstrate that gradient-based attacks such as FGSM and PGD are the most popular approaches to evaluating model robustness. Moreover, there is also active usage of deep learning models, in particular, CNN, LSTM and hybrid due to their ability to reveal a certain complexity in network traffic. It may also be observed that the enormous majority of research employ benchmark datasets, such as CIC-IDS2017, NSL-KDD, and UNSW-NB15, whereas relatively less research assume real-world settings, namely, IoFT and smart grid systems. This may be a shortcoming of applying proposed models to real world.

In addition, even with several mitigation strategies, including adversarial training, GAN-based techniques and conditional models being proposed, many of these solutions currently do not handle adaptive and unseen adversarial attacks.

Overall, this review has demonstrated that there is the necessity of having stronger, flexible and hybrid intrusion detection systems that can be useful in real-life conditions. The outcomes are useful in enhancing the understanding of the current research trends and to establish a foundation that will enable the development of more credible AI-based cybersecurity systems.

7 FUTURE WORK

The research needs to be done in future to come up with more resilient and adaptive intrusion detection systems that can resist sophisticated and dynamic adversarial evasion attacks. The necessity is to go beyond restricted evaluation conditions and take into consideration more varied and closer to reality attack conditions.

Also, further research must focus on the applicability of real-world and domain-specific data, specifically to new systems including UAV networks and smart grid systems. Such an assessment of systems will give a better idea about the performance of systems under real-life situations.

Furthermore, in future work, there are better ways to develop more sophisticated adversarial training designs and dynamic attack generation mechanisms that can enhance system resilience against unknown and evolving threats.

Last but not least, the real-time detection mechanisms and effective solutions that should be designed to be working efficiently in resource-constrained environment are worth investigating to maintain the balance between the performance and the practicability.

REFERENCES

[1] Z. Xu, Y. Wu, S. Wang, J. Gao, T. Qiu, and Z. Wang, “Deep learning-based intrusion detection systems: A survey,” Journal of the ACM, vol. 1, no. 1, pp. 1-38, 2025, [Online]. Available: https://doi.org/10.48550/arXiv.2504.07839.

[2] V. F. Santos, C. Albuquerque, D. Passos, and S. E. Quincozes, “Assessing machine learning techniques for intrusion detection in cyber-physical systems,” Energies, pp. 1-18, 2023, [Online]. Available: https://doi.org/10.3390/en16166058.

[3] U. Ahmed, J. C. W. Lin, and G. Srivastava, “Mitigating adversarial evasion attacks of ransomware using ensemble learning,” Computers & Electrical Engineering, vol. 100, p. 107903, 2022, [Online]. Available: https://doi.org/10.1016/j.compeleceng.2022.107903.

[4] Oprea, M. Hamin, and A. Fordyce, “NIST AI 100-2e2025,” 2025.

[5] Tafreshian, “A defensive framework against adversarial attacks on machine learning-based network intrusion detection systems,” in Proceedings of IEEE TrustCom, 2024, [Online]. Available: https://doi.org/10.1109/TrustCom63139.2024.00337.

[6] P. K. Jha, “Adversarial machine learning: Attacks, defenses, and open challenges,” 2025, [Online]. Available: https://doi.org/10.48550/arXiv.2502.05637.

[7] R. Muthalagu, J. Malik, and P. M. Pawar, “Detection and prevention of evasion attacks on machine learning models,” Expert Systems with Applications, vol. 266, p. 126044, 2025, [Online]. Available: https://doi.org/10.1016/j.eswa.2024.126044.

[8] M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model evasion attack on intrusion detection systems using adversarial machine learning," in IEEE Conference on Information Sciences and Systems (CISS), 2020, [Online]. Available: https://doi.org/10.1109/CISS48834.2020.1570617116

[9] Li, S. Cui, Y. Li, J. Xu, F. Xiao, and S. Xu, "PAD: Towards principled adversarial malware detection against evasion attacks," IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 2, pp. 920-936, 2024, [Online]. Available: https://doi.org/10.1109/TDSC.2023.3265665.

[10] J. Lum, Q. Zeng, S. Chen, O. Fawkes, and H. Chen, "Foe for fraud: Transferable adversarial attacks in credit card fraud detection," in IEEE International Conference on Web Services, 2025, [Online]. Available: https://doi.org/10.1109/ICWS67624.2025.00043.

[11] A. Wainakh et al., "Defenses against evasion attacks in the automotive industry," IEEE Open Journal of Vehicular Technology, vol. 6, pp. 2272-2300, 2025, [Online]. Available: https://doi.org/10.1109/OJVT.2025.3595705.

[12] U. Ahmed, J. C. W. Lin, and G. Srivastava, "Mitigating adversarial evasion attacks by deep active learning for medical image classification," Multimedia Tools and Applications, vol. 81, no. 29, pp. 41899-41910, 2022, [Online]. Available: https://doi.org/10.1007/s11042-021-11473-z.

[13] A. Paya and S. Arroni, "Apollon: A robust defense system against adversarial machine learning attacks in intrusion detection systems," Computers & Security, 2025, [Online]. Available: https://doi.org/10.1016/j.cose.2023.103546.

[14] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," Future Generation Computer Systems, vol. 110, pp. 148-154, 2020, [Online]. Available: https://doi.org/10.1016/j.future.2020.04.013.

[15] J. Cortellazzi and C. London, "Intriguing properties of adversarial machine learning attacks," 2023.

[16] M. Pawlicki, A. Pawlicka, R. Kozik, and M. Choraś, "A meta-survey of adversarial attacks against artificial intelligence algorithms," Neurocomputing, vol. 653, p. 131231, 2025, [Online]. Available: https://doi.org/10.1016/j.neucom.2025.131231.

[17] A. Abomakhelb, K. A. Jalil, A. G. Buja, and A. Alhammadi, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," Technologies, 2025, [Online]. Available: https://doi.org/10.3390/technologies13050202.

[18] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI," Frontiers in Computer Science, 2025, [Online]. Available: https://doi.org/10.3389/fcomp.2025.1520741.

[19] A. Mari, D. Zinca, and V. Dobrota, "Development of a machine-learning intrusion detection adversarial network," 2023.

[20] P. Dini, A. Elhanashi, A. Begni, S. Saponara, Q. Zheng, and K. Gasmi, "Overview on intrusion detection systems design exploiting machine learning for cybersecurity," Applied Sciences, 2023, [Online]. Available: https://doi.org/10.3390/app13137507.

[21] L. Ali, K. Thakur, S. Schmeelk, J. Debello, and D. Dragos, "Deep learning vs. machine learning for intrusion detection in computer networks: A comparative study," Applied Sciences, 2025, [Online]. Available: https://doi.org/10.3390/app15041903.

[22] V. Heydari, "Enhancing adversarial robustness in network intrusion detection," Electronics, 2025, [Online]. Available: https://doi.org/10.3390/electronics14163249.

[23] S. Sharma and Z. Chen, "A systematic study of adversarial attacks against network intrusion detection systems," Electronics, vol. 13, no. 24, 2024, [Online]. Available: https://doi.org/10.3390/electronics13245030.

[24] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," 2021, [Online]. Available: https://doi.org/10.1136/bmj.n71.

[25] D. Berend and Y. Elovici, "TANTRA: Timing-based adversarial network traffic reshaping attack," IEEE Transactions on Information Forensics and Security, 2018, [Online]. Available: https://doi.org/10.1109/TIFS.2022.3201377.

[26] I. Debicha, R. Bauwens, T. Debatty, J. Dricot, T. Kenaza, and W. Mees, "Transfer learning-based multi-adversarial detection of evasion attacks," Future Generation Computer Systems, 2022, [Online]. Available: https://doi.org/10.1016/j.future.2022.08.011.

[27] X. Yuan, S. Han, W. Huang, H. Ye, X. Kong, and F. Zhang, "A framework to enhance adversarial robustness of intrusion detection systems," 2023, [Online]. Available: https://doi.org/10.1016/j.cose.2023.103644.

[28] M. Demirci, "RAIDS: Robust autoencoder-based intrusion detection system," Computers & Security, 2025, [Online]. Available: https://doi.org/10.1016/j.cose.2023.103483.

[29] Christos, "Investigating FGSM and CTGAN adversarial attacks on intrusion detection systems," 2026, [Online]. Available: https://doi.org/10.1145/3600160.3605163.

[30] A. T. El-Toukhy et al., "Countering evasion attacks for smart grid reinforcement learning detectors," IEEE Access, vol. 11, pp. 97373-97390, 2023, [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3312376.

[31] Y. Fukuda and K. Yoshida, "Evaluation of model quantization for mitigating adversarial examples," IEEE Access, vol. 11, pp. 87200-87209, 2023, [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3305264.

[32] A. Zafar, "Boosting robustness of intrusion detection systems against adversarial attacks," Expert Systems with Applications, 2025, [Online]. Available: https://doi.org/10.1016/j.eswa.2024.123567.

[33] K. Roshana and A. Zafar, “Black-box adversarial transferability in cybersecurity,” Computers & Security, 2024, [Online]. Available: https://doi.org/10.1016/j.cose.2024.103853.

[34] R. H. Randhawa et al., “Deep reinforcement learning-based GAN for botnet detection,” Future Generation Computer Systems, vol. 150, pp. 294-302, 2024, [Online]. Available: https://doi.org/10.1016/j.future.2023.09.011.

[35] K. Barik, S. Misra, and L. Fernandez-Sanz, “Adversarial attack detection framework based on optimized networks,” International Journal of Information Security, vol. 23, no. 3, pp. 2353-2376, 2024, [Online]. Available: https://doi.org/10.1007/s10207-024-00844-w.

[36] M. Mimura and K. Kurashina, “Practical evasion attack against neural network-based malware detection,” Cluster Computing, vol. 28, no. 2, pp. 1-16, 2025, [Online]. Available: https://doi.org/10.1007/s10586-024-04825-5.

[37] K. Barik and S. Misra, “A comprehensive defense approach of deep learning-based NIDS,” 2025.

[38] T. Gaber, T. Ali, M. Nicho, and M. Torky, “Robust attacks detection model for Internet of flying things,” IEEE Internet of Things Journal, vol. 12, no. 13, pp. 23961-23974, 2025, [Online]. Available: https://doi.org/10.1109/JIOT.2025.3555202.

[39] O. F. Awad, M. Çevik, and H. Mutlag, “An enhanced attention-based model for intrusion detection under evasion attacks,” Peer-to-Peer Networking and Applications, 2025, [Online]. Available: https://doi.org/10.1007/s12083-024-01859-9.

[40] L. Yang, M. El Rajab, and A. Shami, “Enabling AutoML for zero-touch network security,” 2025.

[41] M. Musthafa, “Adversarial robustness in AI-driven cybersecurity solutions,” 2025.

[42] K. Kutubuddin et al., “AI-driven adversarial threat simulation for cyber-defense training,” 2025.