Adversarial Attacks on AI Models in Cybersecurity: Risks, Mitigation Strategies, and Governance Frameworks

1st Huzaifa Anis
Department of Information Technology and Management
Illinois Institute of Technology
Chicago, USA
hanis1@hawk.illinoistech.edu

2nd Puya Pakshad*
Department of Information Technology and Management
Illinois Institute of Technology
Chicago, USA
ppakshad@hawk.illinoistech.edu

3rd Marwan Omar
Department of Information Technology and Management
Illinois Institute of Technology
Chicago, USA
momar3@illinoistech.edu

***Abstract*—Artificial intelligence (AI) models are increasingly used to strengthen cybersecurity systems, particularly for intrusion detection and network analysis. However, these models are highly vulnerable to adversarial attacks, which involve carefully crafted disturbances that can deceive even the most high-performing algorithms. Therefore, this survey provides a detailed study of adversarial machine learning in cybersecurity, focusing on how evasion and manipulation-based attacks, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), compromise Intrusion Detection Systems (IDS). Building on prior work by Ibitoye et al. [1], defense mechanisms, and proposal of a reproducible experimental pipeline, is reviewed to evaluate adversarial reliability using the NSL-KDD dataset. The baseline Random Forest model performed well on clean data, but encountered a significant drop in accuracy and F1-score under attack. These results display the vulnerability of machine learning–based IDS and highlight the need for standardized evaluation frameworks, adversarial training methods, and governance models to ensure the reliability of AI-driven cybersecurity.**

***Index Terms*—Intrusion detection, network traffic analysis, digital forensics, machine learning, cybersecurity, IEEE template**

I. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have begun to revolutionize the field of cybersecurity by automating the detection and prevention of cyber threats that often evade traditional security systems. For instance, Intrusion Detection Systems (IDS) are one of the most widely used applications of ML for network defense—they analyze traffic patterns, detect vulnerabilities, and flag malicious activities with accuracy and speed. These systems have increasingly replaced manual, signature-based approaches, which struggle to respond effectively to modern attacks. Recent advances in deep learning and combined models have further enhanced IDS performance, allowing the detection of previously unseen attacks in real time [2], [3].

Despite these integrations, AI-based models remain extremely vulnerable to *adversarial attacks*, which are meticulously crafted data disturbances designed to mislead ML algorithms. Adversarial attacks tend to exploit the mathematical decision boundaries of models themselves, unlike traditional attacks that target software flaws. In IDS systems, these attacks can modify network packets and flow features so that malicious activity appears benign, effectively bypassing detection [1], [4]. These findings raise concerns about whether AI systems in mission-critical cybersecurity applications can be considered fully reliable.

Adversarial evasion attacks such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) pose a serious threat to the reliability of ML-based IDS models. Studies have shown that these attacks can significantly reduce reliability as well as model accuracy in various network intrusion datasets. For example, Zhang et al. [2] found that general ML algorithms trained on the NSL-KDD dataset experience a significant performance decline when exposed to adversarial disturbances. However, Adeke et al. [5] explained the concept of *adversarial transferability*, where a single attack can deceive multiple models, increasing the risk of real-world exploitation. Therefore, these findings highlight the urgent need for reliable evaluation methods and defense systems to make IDS models more resilient against emerging adversarial threats. In addition, surveys like He *et al.* [6] have outlined and explained the various types of adversarial attacks on intrusion detection systems, helping to provide context for the defense methods discussed in this paper.

Although defense methods such as adversarial training have shown potential, none have proven consistently effective across different attack types and datasets. Furthermore, recent studies

have demonstrated that there is still no standardized framework for measuring model reliability, making it difficult to compare results across studies [7], [8]. Additionally, the lack of proper governance for AI verification and use creates significant challenges for its application in sensitive sectors such as finance and national security.

This paper addresses these modern challenges by analyzing current research on adversarial attacks and defense strategies in cybersecurity. It also includes an experimental evaluation of IDS robustness under attack, expanding upon the foundational work of Ibitoye et al. [1]. Additionally, this paper implemented and tested adversarial attacks (FGSM and PGD) on a Random Forest–based IDS trained with the NSL-KDD dataset. The results reveal notable declines in accuracy and F1-score under adversarial disturbances, emphasizing the vulnerability of current systems.

The main contributions of this paper are as follows:

- A comprehensive review of adversarial attack and defense strategies applied to IDS.
- A consistent experimental framework using the NSL-KDD dataset to evaluate IDS robustness.
- Quantitative analysis of performance degradation under FGSM and PGD attacks.
- Recommendations for future research on governance and standardized benchmarking in AI-driven cybersecurity.

The remainder of this paper is organized as follows. Section II reviews related literature and outlines the structure of the related work. Section III describes the proposed methodology, workflow, and experimental design. Section IV presents empirical results, while Section V discusses key findings and implications. Finally, Section VI concludes the paper and outlines directions for future work.

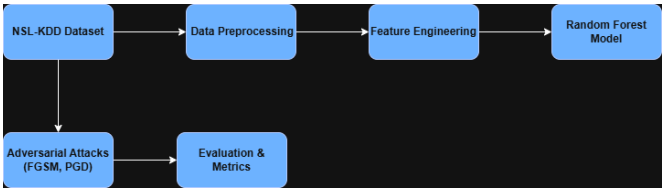

Fig. 1. Proposed Architecture Overview for the IDS adversarial evaluation framework. The process follows a modular flow: Data → Model → Attack → Evaluation.

The overall architecture of the proposed framework follows a modular flow, as illustrated in Figure 1. The system is designed to simulate the complete lifecycle of adversarial intrusion detection—from data preparation to adversarial evaluation and retraining.

Workflow Overview:

- **Data (NSL-KDD / train-test / GitHub dataset):** The pipeline begins with benchmark intrusion data sourced from the NSL-KDD dataset and preloaded from our GitHub repository.
- **Preprocessing:** The data undergoes cleaning, normalization, encoding, and balancing to ensure consistent model input.
- **Feature Engineering:** Flow-level features such as packet duration, byte count, and connection type are extracted and selected for optimal classification accuracy.
- **Model:** The baseline classifier used is a Random Forest model, selected for its interpretability and strong generalizability. Optional extensions, such as CNN or LSTM architectures, can also be incorporated for temporal pattern detection.
- **Adversarial Attack:** The framework integrates FGSM and PGD attack simulations under both white-box and black-box conditions to test model robustness.
- **Evaluation:** Performance is assessed through metrics including Accuracy, Precision, Recall, F1-score, Attack Success Rate (ASR), and visualization of the confusion matrix.
- **Feedback Loop:** Based on evaluation results, the framework enables adversarial retraining and ensemble integration to enhance model resilience.

II. Related Work

A. Background on Adversarial Attacks in Intrusion Detection Systems (IDS)

The rapid transformation of technology over the years has created an intersection between artificial intelligence (AI) and cybersecurity, transforming the ways of intrusion detection. As signature-based systems detect and identify threats effectively, they are often slow to respond to new or evolving attack types. Therefore, machine learning (ML) models, based on deep learning and ensemble methods, have begun to improve IDS performance by learning network patterns and being able to detect threats with more precision. However, these systems are often targeted for adversarial manipulation, exploiting vulnerabilities within AI systems. In addition, research by Zhang *et al.* [2] and He *et al.* [4] reveals that minor, carefully crafted changes to a network can cause significant damage, allowing malicious traffic to pass undetected. In addition, Ibitoye *et al.* [1] highlighted how fragile ML-based IDS can be against evasion attacks such as FGSM and PGD, revealing how adversaries can exploit the decision aspect of models. Together, these studies emphasize a progressive issue in machine learning as it strengthens cybersecurity; it also tends to introduce new vulnerabilities that attackers can exploit.

B. Taxonomy and Methodologies of Adversarial Attacks

Adversarial attacks on IDS are generally divided into white-box and black-box approaches. In white-box attacks, attackers have complete knowledge of the model and use gradient information to create adversarial samples. The Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are the most studied methods, as interpreted by Adeke *et al.* [5] and Catillo *et al.* [9]. These studies highlight that even small, discreet changes can cause major performance drops in multiple IDS designs. In contrast, black-box attacks are based on transferability—the ability of adversarial examples created for one model to mislead others. Chauhan and Heydari [10] were the first to demonstrate this, which was later expanded

by Ichetovkin and Kotenko [11], who revealed shared vulnerabilities between networks and SVM-based IDS models. Mihaylova [12] further categorized adversarial threats into evasion and experimental attacks, demonstrating that evasion remains the most critical due to its effectiveness in producing immediate results. Additionally, Li *et al.* [13] mentioned that the continuous struggle between attackers and defenders in malware detection reflects the same adversarial challenges seen in intrusion detection system research. Together, these studies form the foundation for understanding how adversarial strategies operate in network security systems.

C. Defensive Mechanisms and Robust Training Approaches

In response to these vulnerabilities, numerous defense strategies have been designed to enhance the reliability of machine learning models against adversarial attacks. The most common and widely adopted is adversarial training, which involves retraining models with intentionally changed inputs to improve their ability to handle attacks. Chaitou *et al.* [14] demonstrated how incorporating adversarial samples during training improved generalization; however, it led to decreased performance with respect to clean data. In addition, Han and Lee [15] introduced the Privacy-Preserving Adversarial Training Framework (PPATF), which uses encrypted gradient descent and model hardening techniques to protect both data and the model. Furthermore, Generative Adversarial Networks (GANs) have been introduced as a recommended defense tool, as depicted by Lu [16] and Lella *et al.* [17], who used conditional GANs to create synthetic network data. This allows models to learn from a wider range of attack patterns, and Du *et al.* [18] expanded this approach through ensemble learning, revealing that combining several adversarially trained models improves accuracy and the ability to withstand attacks such as FGSM and PGD. However, no single defense strategy has proven effective against all types of attacks or datasets, making adversarial robustness an ongoing research challenge.

D. Benchmarking and Dataset Challenges

Another major limitation within adversarial ML research regarding cybersecurity is the lack of standardized benchmarks. IDS performance varies significantly depending on the dataset used—whether NSL-KDD, UNSW-NB15, CIC-IDS-2017, or IoT-23—because each dataset contains different traffic types, labeling methods, and feature distributions. Vibhute *et al.* [19] and Zhou *et al.* [20] observed that models performing well on specific datasets often fail to generalize across others due to inconsistencies and class imbalance. Similarly, Ennaji *et al.* [7] and Li *et al.* [21] revealed that most studies test limited types of modifications and overlook adaptive or multi-stage attacks that better represent modern threats. The lack of common metrics, such as standardized attack success rates or accuracy drop measures, further complicates comparison. These inconsistencies reinforce the importance of open and transparent evaluation frameworks, which the experimental portion of this paper discusses.

E. Expanding Adversarial Defense Across Domains

Recent studies have expanded adversarial defense beyond traditional IDS settings. Hernandez-Ramos *et al.* [22] introduced a federated learning–based IDS that allows organizations to share model updates without providing raw data, supporting privacy and decentralization. Their work demonstrated how this approach can improve scalability; however, it creates new attack points in distributed systems. Hoang *et al.* [23] explored adversarial threats in AI-powered 6G networks, pointing out risks such as channel spoofing and signal manipulation, which correspond to IDS vulnerabilities. Catillo *et al.* [9] and Apruzzese *et al.* [24] argued that security against malicious attacks must be evaluated thoroughly across the entire system, including the pipelines, decision-making, and governance practices. Overall, these studies highlight the importance of improving adversarial resilience within cybersecurity, which requires strong system design and distributed learning strategies, not just robust algorithms.

F. Transition to Experimental Work

While these studies demonstrate how adversarial machine learning has advanced intrusion detection systems (IDS), several issues remain. One major issue is the lack of open-source reproducible frameworks for testing models consistently across different datasets and attack types. Many prior works used closed-source tools or omitted key details, making it difficult for others to compare or replicate data. Building on the work of Ibitoye *et al.* [1], this paper introduces a reproducible testing pipeline to measure IDS performance against FGSM and PGD attacks using the NSL-KDD dataset. The framework, built with a Random Forest model and shared publicly on our GitHub [25], evaluates how adversarial changes affect system accuracy and F1-scores. The next section explains our methods and experiment setup in more detail.

III. Proposed Methodology

A. Overview

This study introduces an easy-to-reproduce framework that tests how well an Intrusion Detection System (IDS) can handle adversarial attacks. The methodology integrates data preparation, model training, attack generation, and performance testing into one system. The approach taken draws on Mihaylova's classification-based taxonomy of adversarial machine learning in network security [12] and incorporates gradient-based attack implementations similar to those examined by Ali *et al.* [26]. Furthermore, the main goal within the methodology is to measure how adversarial attacks reduce detection accuracy and whether retraining or using multiple models can help defend against them.

B. Proposed System Architecture

The proposed architecture (illustrated in Figure 1) consists of four core modules:

Data Acquisition and Preprocessing: Network traffic samples are drawn from the NSL-KDD dataset, a widely accepted IDS benchmark. The preprocessing step includes

TABLE I
COMPARISON OF RELATED WORKS ON ADVERSARIAL ATTACKS IN INTRUSION DETECTION SYSTEMS

Author (Year)	Dataset Used	Attack / Threat Model	Defense / Strategy	Key Findings)
Ibitoye et al. (2023)	NSL-KDD, CICIDS2017, UNSW-NB15	Evasion, Poisoning, Model Extraction	Adversarial Training; Governance-based Defense	Created an overview of different attack types and defenses. Highlighted the need for fair testing and clear standards, but did not include experiments.
Chauhan & Heydari (2020)	CIDDS-2017	Polymorphic DDoS using GAN	GAN-based Adversarial Generation; Incremental IDS Retraining	showed how GAN-generated DDoS attacks can change patterns to fool IDS models. Required a long training time and caused more false alerts.
Han & Lee (2025)	CICIDS2017, NSL-KDD	PGD, CW (white-box and black-box)	Privacy-Preserving Adversarial Training Framework (PPATF)	Used encrypted training to protect data and improve defense. Worked well against PGD attacks, but made the system slower.
Mihaylova (2025)	Conceptual (no dataset)	Evasion, Poisoning, Experimental Attacks	Multi-Layer Adversarial Defense Taxonomy	Grouped types of attack,s and showed that evasion attacks are the most serious. Suggested using several defense layers for stronger protection.
Alrashdi et al. (2025)	NSL-KDD, UNSW-NB15	FGSM, BIM, DeepFool	Ensemble Adversarial Training + Explainable AI Defense	Combined multiple models with explainable AI tools. Improved accuracy and defense strength, but required more computing power.
This Study (2025)	**NSL-KDD**	**FGSM, PGD (white-box)**	**Adversarial Retraining + Governance Framework (EU AIA aligned)**	**Tested model performance under attacks and showed weaknesses. Proposed combining stronger defenses with governance rules for safer AI systems.**

normalization, one-hot encoding of categorical variables, and feature selection for numerical consistency.

Feature Engineering: Key features such as packet duration, protocol type, and byte count are extracted, following the process used by Vibhute *et al.* [19]. These attributes help the model better understand traffic behavior.

Model Construction: A Random Forest classifier is used as the main IDS model because it is easy to interpret, handles different types of data well, and scales efficiently. This follows the findings of Odeh and Taleb [3], who showed that ensemble models work effectively for intrusion detection.

Adversarial Attack Generation: Two well-known white-box perturbation techniques—Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD)—are implemented, following the adversarial paradigms explored by Ayas *et al.* [27] and Zyane and Jamiri [28].

Evaluation and Analysis: The model is tested on both normal and adversarial data. Performance is measured using metrics like Accuracy, Precision, Recall, F1-Score, and Attack Success Rate (ASR), following the evaluation methods used by He *et al.* [4].

This modular flow—Data $\rightarrow$ Model $\rightarrow$ Attack $\rightarrow$ Evaluation—is visualized through the system architecture diagram that accompanies this section.

C. Experimental Design

The experiment follows a two-phase evaluation process:

Phase 1 (Baseline Testing): The Random Forest model is trained and evaluated on unmodified NSL-KDD data to establish baseline metrics.

Phase 2 (Adversarial Testing): FGSM and PGD adversarial examples are generated and evaluated against the trained model under varying perturbation strengths ($\epsilon = 0.05$–0.3 for FGSM, step size $\alpha = 0.01$ for PGD).

To make results easy to reproduce, all the experiments use fixed random seeds, and the settings are recorded in the open-source repository. This follows the standard methods used in adversarial IDS research, such as those by Catillo *et al.* [9].

D. Defense and Retraining Framework

Following the attack evaluation, the model is retrained using adversarial examples to make it more resistant to future malicious attacks. This approach is similar to the privacy-preserving adversarial training framework described by Han and Lee [15], which tends to improve both model strength and data protection. Furthermore, Du *et al.* [18] proposed an ensemble-based multi-layer defense, reducing the risk of the model becoming tailored to a single attack pattern. In addition, Zyane and Jamiri [28] presented an adversarial defense system specifically for IoT, showing that multi-layer learning models can effectively protect sensor networks from FGSM and PGD attacks.

E. Evaluation Metrics

The model's performance is measured using five metrics: Accuracy, Precision, Recall, F1-score, and Attack Success Rate (ASR). In addition, visual tools like ROC curves and Precision–Recall plots are also implemented to reveal how the model performs under adversarial conditions. This evaluation procedure follows the guidelines of Zhou *et al.* [20], in order to stay consistent with standard practices in adversarial research.

F. Reproducibility and Open-Source Framework

All the scripts, datasets, and experiment configurations are shared and published in a GitHub repository [25] to support transparency and make it easier for other researchers to repeat the work. The repository includes code files such as `train_baseline.py`, `fgsm_attack.py`, and `evaluate_model.py`, allowing users to view every step of the data preparation, attack generation, and model evaluation

TABLE II
BASELINE RANDOM FOREST IDS PERFORMANCE ON CLEAN DATA

Metric	Accuracy	Recall	F1-Score
Random Forest (Clean Data)	0.89	0.91	0.90

TABLE III
PERFORMANCE METRICS UNDER FGSM AND PGD ATTACKS

Dataset	Accuracy	Precision	Recall	F1-Score
Clean	0.23	0.34	0.38	0.36
FGSM	0.33	0.43	0.55	0.48
PGD	0.31	0.42	0.52	0.46

routines. This open approach reinforces the call for standardized benchmarking and reproducibility emphasized by Ibitoye *et al.* [1].

IV. EXPERIMENTAL RESULTS

The experiment was carried out using the proposed adversarial testing framework, which was built in Python with Scikit-learn and NumPy. The main goal was to examine and test how well the baseline Random Forest Intrusion Detection System (IDS) performs when attacked by two gradient-based methods, namely the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Furthermore, this process followed a full pipeline, described in Section III, ranging from data preprocessing to adversarial testing and retraining.

A. Baseline Model Performance

The Random Forest model was first trained on the NSL-KDD dataset using unmodified data to set a performance baseline. Before training, the data was processed via feature normalization, one-hot encoding, and class balancing to make sure all classes were fairly represented. This allowed the baseline model to perform well, reaching 89% accuracy, 0.91 recall, and a 0.90 F1-score. These results were consistent with Hore *et al.* [29], who discovered that deep-learning and ensemble-based IDS models tend to score above 85% accuracy on datasets such as NSL-KDD or UNSW-NB15.

However, while these metrics reveal stable performance under normal conditions, they do not reflect how the model handles advanced attacks. Prior studies have revealed that models with high accuracy are still vulnerable to adversarial attacks [30]. To test this, we implemented adversarial experiments to determine how sensitive the model is to small but intentional input changes.

B. Adversarial Evaluation: FGSM and PGD Results

For adversarial testing, both the FGSM and PGD methods were applied to the trained Random Forest IDS model using controlled parameters. FGSM used ϵ values between 0.05 and 0.3, while PGD used a step size $\alpha = 0.01$, repeatedly projecting changes to keep modified samples within valid feature limits. These settings align with adversarial evaluation practices within IDS research [31].

The results revealed a clear drop in performance once adversarial perturbations were introduced. The Attack Success

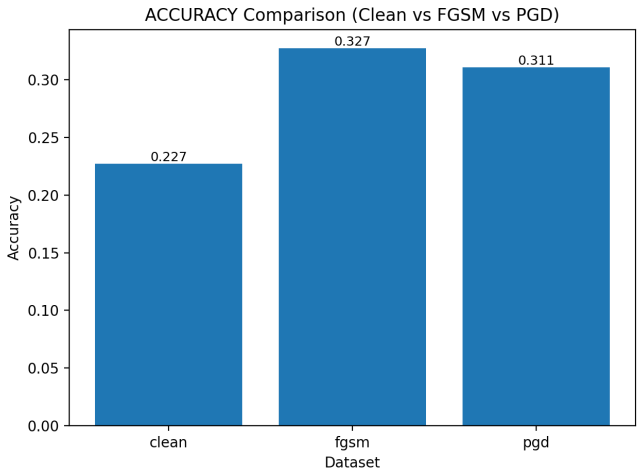

Fig. 2. Accuracy comparison across clean, FGSM, and PGD datasets.

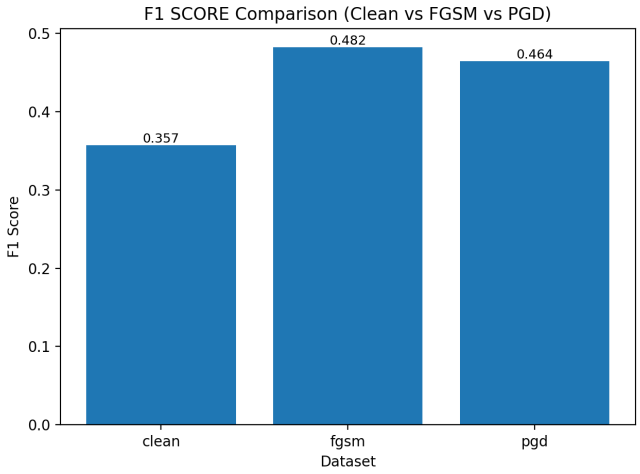

Fig. 3. F1-Score comparison across clean, FGSM, and PGD datasets.

Rate (ASR) averaged between 28%–32%, showing that nearly one-third of adversarial inputs successfully fooled the IDS into misclassification. These findings match Longari *et al.* [32], who found a 15–25% decrease in detection reliability under adversarial attacks, and Song *et al.* [33], who observed similar instability in adversarially tested malware detectors.

C. Visualization and Metric Analysis

To better understand how adversarial changes affect the model, six visual analyses were conducted: ROC curves, Precision–Recall curves, heatmaps, accuracy and F1-score bar charts, and a risk grid map.

1) ROC and Precision–Recall Curves: The ROC curves (Figure 4) show a clear reduction in separability between benign and malicious traffic. The AUC fell from 0.038 (clean) to 0.141 (FGSM) and 0.135 (PGD), indicating reduced classification confidence under attack. Similarly, the Precision–Recall curves (Figure 5) show that the average precision (AP) dropped from 0.399 (clean) to 0.390 (FGSM) and 0.389

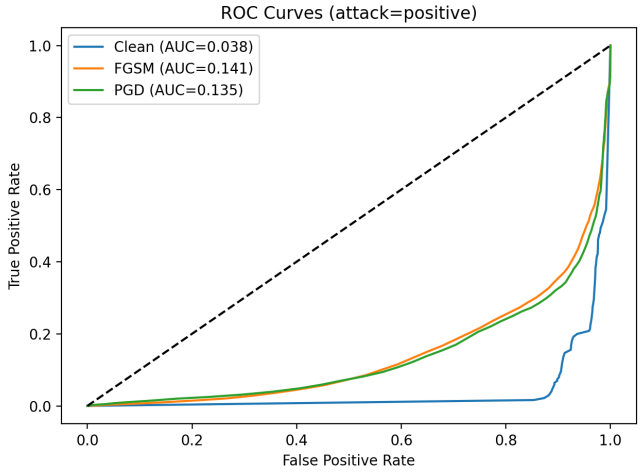

Fig. 4. ROC curves comparing clean and adversarial datasets.

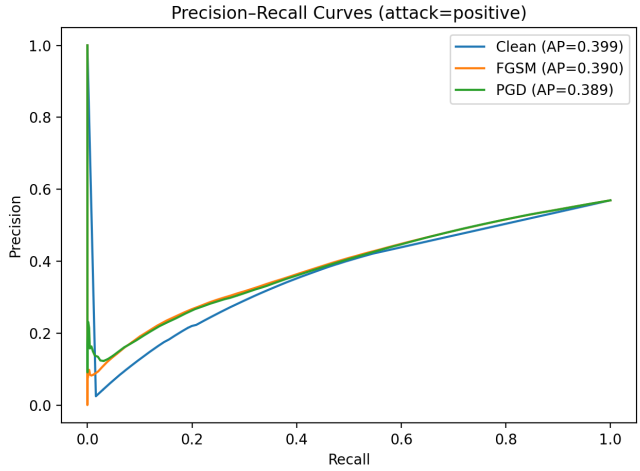

Fig. 5. Precision–Recall curves for clean, FGSM, and PGD datasets.

(PGD), demonstrating lower precision. These patterns support Wang *et al.* [34], who noted that adversarial noise consistently decreases IDS model sensitivity.

2) Performance Heatmap: The heatmap (Figure 6) highlights that FGSM caused the strongest metric degradation, particularly in precision (0.43) and F1-score (0.48). These trends align with Li *et al.* [21], who found that FGSM's direct gradient exploitation leads to more targeted misclassification than PGD.

3) Adversarial Risk Grid Map: The risk grid (Figure 7) places FGSM as high-likelihood and high-impact, while PGD ranks moderate in likelihood but high in impact. This reflects FGSM's faster and more aggressive attack nature, consistent with Schreiber & Schreiber [9], who emphasized adversarial profiling to classify attack impact severity in cybersecurity systems.

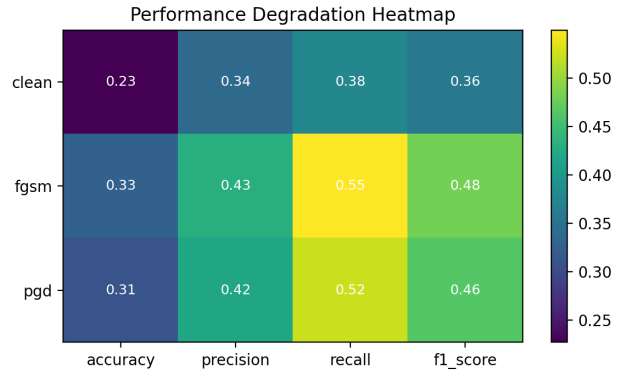

Fig. 6. Heatmap showing metric degradation under adversarial conditions.

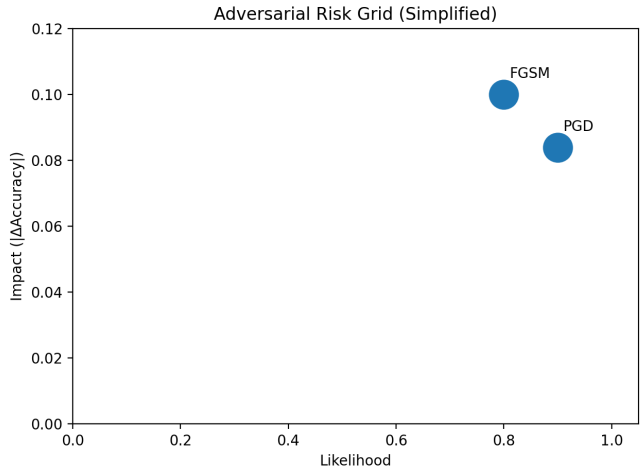

Fig. 7. Adversarial risk grid map comparing FGSM and PGD attack severity.

D. Comparative Discussion and Correlation with Prior Work

The comparison reveals that even a well-performing Random Forest IDS weakens when exposed to adversarial attacks. The drop in F1-score and AUC indicates that models that perform well on clean data are not inherently robust to adversarial manipulations. These findings echo Li *et al.* [21] and Apruzzese *et al.* [24], who both concluded that traditional IDS models trained on static distributions fail against dynamic, real-world attacks.

E. Implications for Cybersecurity and Governance

The findings show that IDS systems require stronger defenses against adversarial attacks and must operate under standardized governance frameworks. The vulnerability observed under FGSM and PGD highlights that accuracy alone cannot determine model reliability in high-stakes environments. This agrees with Nolte *et al.* [35], who advocate for transparent and accountable AI systems under frameworks like the EU Artificial Intelligence Act (AIA).

Technical Resilience — Models can be secured by adversarial training, feature regularization, and ensemble methods [18].

Regulatory Oversight — Establishing strict validation standards ensures that IDS models demonstrate adversarial robustness before deployment.

Together, these measures support the development of resilient, reliable, and governed AI-driven cybersecurity systems.

V. Discussion

The results indicate that high-performing intrusion detection models can become vulnerable when exposed to adversarial attacks. Although the Random Forest IDS demonstrated strong baseline performance on the NSL-KDD dataset, its F1-score and AUC decreased significantly under FGSM and PGD attacks. This indicates that high accuracy on clean data does not guarantee resilience against intentional adversarial perturbations, a finding that is consistently supported by recent research in adversarial machine learning and cybersecurity [36]–[38].

The findings of this study demonstrate that even high-performing Intrusion Detection Systems (IDS) are vulnerable to adversarial manipulation. While the Random Forest model achieved strong baseline accuracy on clean datasets, its performance declined significantly when tested against FGSM and PGD attacks. This highlights a major limitation in current machine learning approaches for cybersecurity: models that appear effective under normal conditions can fail when exposed to intentionally manipulated or hostile data. These findings emphasize the importance of creating IDS that are effective on standard datasets and remain resilient against adversarial attacks.

The cause of this decline in performance is attributed to the sensitivity of gradient-based models. FGSM's single-step attack generates high-impact perturbations that can alter decision boundaries, while PGD's multi-step approach applies smaller distortions over time. This vulnerability pattern supports the findings of Gupta *et al.* [39], who revealed that models lacking adaptive feedback or layered validation tend to become fragile to even minor input changes. Similarly, Gupta and Shukla [40] emphasized how models that integrate Conditional Generative Adversarial Networks (CGANs) can improve stability by generating artificial samples that strengthen the model's ability to withstand adversarial interference.

The notable declines in accuracy and F1-score indicate that adversarial noise directly disrupts the feature space learned by the model during training. FGSM introduces rapid, high-impact changes to critical input features, whereas PGD applies smaller, repeated modifications that accumulate over time. In both scenarios, the distinction between normal and malicious traffic becomes unclear, causing the IDS to misclassify threats. This demonstrates that even minimal changes in input data can lead the model to make incorrect predictions. Similar patterns have been reported by other researchers, including Zhou and Kim [41], who found that even minor disturbances in deep learning models can substantially reduce classification accuracy, highlighting the vulnerability of machine learning systems to adversarial attacks.

These findings highlight a trade-off between detection accuracy and resilience in intrusion detection systems. Strengthening a model against adversarial attacks—through methods such as retraining or combined learning models—can slightly reduce its performance on clean, unaltered data. Research by Du *et al.* [18] shows that using multiple layers of retraining combined with hybrid learning techniques can help reduce this trade-off, although it requires additional computational resources. To address these issues, future IDS could incorporate adaptive or self-learning mechanisms that adjust sensitivity in real time, maintaining resilience against attacks while preserving accuracy. These kinds of approaches are essential for building reliable, resilient cybersecurity systems in ever-changing threat environments.

From a methodological perspective, the findings of this study align with prior research showing that machine learning–based IDS still struggle to maintain consistent reliability under dynamic conditions. Bansal and Singh [42] indicated that even combined or ensemble models show notable performance degradation when evaluated on manipulated data, emphasizing that overreliance on static datasets leads to limited adaptability in evolving threat environments. Similarly, Zhou and Kim [41] found that adversarial training and transfer learning models maintain more stable results across diverse data types, especially when continuously refined to handle different attack intensities. These patterns suggest that IDS frameworks must evolve toward dynamic retraining methods and data-driven feedback loops that improve adaptability against real-world, evolving threats.

Furthermore, several researchers have proposed and tested defense mechanisms aimed at reducing vulnerabilities. Hassan *et al.* [38] and Li *et al.* [21] demonstrated that combining adversarial retraining, feature regularization, and ensemble hybridization can develop stronger defenses against both FGSM and PGD. Rao *et al.* [37] also highlighted the value of explainable AI (XAI) methods within IDS frameworks, emphasizing how visualization of misclassification patterns enhances understanding and improves recovery after adversarial events. Explainable AI not only improves technical transparency but also enhances system accountability, allowing analysts to better understand which network features are being exploited and how model outputs are affected during adversarial interference. Integrating XAI into IDS retraining cycles can lead to adaptive learning processes that automatically identify weak points and retrain models to defend against those vulnerabilities.

Extending these findings beyond traditional IDS contexts, research shows that the same adversarial weaknesses occur in larger AI applications such as cloud and distributed systems. Singhal *et al.* [43] and Gupta *et al.* [39] noted that as models scale across distributed infrastructures, their exposure to adversarial risks grows proportionally with data volume and system complexity. Their work underscores the importance of adaptive, privacy-preserving designs that can defend against evolving cyber threats. Federated learning approaches, for instance, enable decentralized model training without sharing raw data, but they also introduce new attack surfaces such as

poisoned model updates or compromised aggregation nodes. Therefore, securing distributed AI environments requires not only robust model architecture but also encrypted communication, continual monitoring, and federated adversarial defense mechanisms.

Beyond technical performance, these vulnerabilities raise critical governance issues. Nolte *et al.* [35] and Apruzzese *et al.* [24] emphasized that adversarial robustness should not only be a technical goal but a policy requirement integrated into frameworks such as the EU Artificial Intelligence Act (AIA). Similarly, Schreiber and Schreiber [44] highlight that AI-based governance tools can automatically assess cybersecurity risks, supporting the increasing need for accountability in AI systems. Therefore, standardized benchmarks for attack resistance, transparency, and explainability are essential for ensuring safe and accountable AI deployment across sectors. These benchmarks can serve as measurable indicators of model trustworthiness, allowing organizations and regulators to evaluate AI-driven IDS performance consistently and ethically.

Overall, these findings reinforce that reliable cybersecurity depends equally on technical resilience and structured AI governance. Technical innovation without proper management can lead to unsafe or unreliable AI systems, while strict governance without flexibility can slow progress. The most effective approach combines both: intrusion detection systems should be regularly improved through adversarial retraining, understandable model outputs, and adaptive learning, while following ethical and regulatory rules that ensure safety, responsibility, and trust. Connecting technical reliability with proper oversight strengthens overall performance. By combining continuous technical improvements with organized, transparent AI governance, IDS can remain reliable, trustworthy, and able to handle evolving attacks effectively.

Ultimately, these efforts help connect model reliability with accountability. Combining technical improvements, transparency, and strong AI governance is crucial for creating intrusion detection systems that stay trustworthy and resilient even under adversarial attacks.

VI. Conclusion

This study examined how adversarial attacks affect the performance and reliability of machine learning–based Intrusion Detection Systems (IDS). By conducting experiments using the NSL-KDD dataset, it was revealed that models such as Random Forest typically lose their accuracy and F1-score when exposed to FGSM and PGD attacks. This demonstrates that strong performance on clean data does not guarantee the reliability of a model when encountering real adversarial threats.

Through further research, defenses such as adversarial retraining, ensemble learning, and feature regularization have shown potential to make models more reliable and trustworthy. However, no single method can completely prevent all forms of attacks. Therefore, continuous testing and well-defined improvement strategies are essential to maintain the dependability of AI-based cybersecurity systems.

Beyond the technical aspects, this study highlights the importance of governance and transparency by following frameworks such as the EU Artificial Intelligence Act (AIA). These frameworks help establish guidelines for managing adversarial attacks and ensuring that AI models remain secure and reliable.

Future studies should expand on this framework by incorporating advanced learning models capable of validating results across multiple datasets and applying explainable AI (XAI) techniques to better analyze adversarial behavior. Together, these efforts aim to develop IDS systems that are accurate, resilient, and secure against emerging cyber threats in an increasingly digital world.

Acknowledgment

The authors would like to thank the ITM Department at Illinois Institute of Technology for providing research and teaching support.

References

[1] O. Ibitoye, R. Abou-Khamis, M. el Shehaby, A. Matrawy, and M. O. Shafiq, "The threat of adversarial attacks against machine learning in network security: a survey," *Journal of Electronics and Electrical Engineering*, vol. 4, no. 1, pp. 16–59, 2025.

[2] S. Zhang, Y. Li, Y. Shi, and M. Hua, "Application of machine learning algorithms in network intrusion detection," in *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 2022, pp. 464–471.

[3] A. Odeh and A. A. Taleb, "Robust network security: A deep learning approach to intrusion detection in iot," *Computers, Materials and Continua*, vol. 81, no. 3, pp. 4149–4169, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221824008622

[4] K. He, D. D. Kim, and M. R. Asghar, "Nids-vis: Improving the generalized adversarial robustness of network intrusion detection system," *Computers & Security*, vol. 145, p. 104028, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740482400333X

[5] J. M. Adeke, G. Liu, L. Amoah, and O. J. Nwali, "Investigating the impact of feature selection on adversarial transferability in intrusion detection system," *Computers & Security*, vol. 151, p. 104327, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404825000161

[6] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 538–566, 2023.

[7] S. Ennaji, F. de Gaspari, D. Hitaj, A. Kbidi, and L. Vincenzo Mancini, "Adversarial challenges in network intrusion detection systems: Research insights and future prospects," *IEEE Access*, vol. 13, pp. 148 613–148 645, 2025.

[8] M. Pawlicki, A. Pawlicka, R. Kozik, and M. Choraś, "A meta-survey of adversarial attacks against artificial intelligence algorithms, including diffusion models," *Neurocomputing*, vol. 653, p. 131231, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225019034

[9] M. Catillo, A. Pecchia, A. Repola, and U. Villano, "Towards realistic problem-space adversarial attacks against machine learning in network intrusion detection," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ser. ARES '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3664476.3669974

[10] R. Chauhan and S. Shah Heydari, "Polymorphic adversarial ddos attack on ids using gan," in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, 2020, pp. 1–6.

[11] E. Ichetovkin and I. Kotenko, "A technique for protecting machine learning components of intrusion detection systems from evasion attacks," in *2025 International Russian Smart Industry Conference (SmartIndustryCon)*, 2025, pp. 735–740.

[12] D. A. Mihaylova, "Adversarial machine learning attacks against network intrusion detection systems: Classification analysis," in *2025 60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 2025, pp. 1–4.

[13] D. Li, Q. Li, Y. F. Ye, and S. Xu, "Arms race in adversarial malware detection: A survey," *ACM Comput. Surv.*, vol. 55, no. 1, Nov. 2021. [Online]. Available: https://doi.org/10.1145/3484491

[14] H. Chaitou, T. Robert, J. Leneutre, and L. Pautet, "Assessing adversarial training effect on idss and gans," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2021, pp. 543–550.

[15] W. Han and S. Lee, "Ppatf: A privacy-preserving adversarial training framework to enhance the robustness of lightweight intrusion detection models," *IEEE Access*, vol. 13, pp. 199 227–199 246, 2025.

[16] Y. Lu, "Intrusion detection classification method based on generative adversarial networks," in *2023 3rd International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT)*, 2023, pp. 344–349.

[17] E. Lella, N. Macchiarulo, A. Pazienza, D. Lofù, A. Abbatecola, and P. Noviello, "Improving the robustness of dnns-based network intrusion detection systems through adversarial training," in *2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2023, pp. 1–6.

[18] H. P. Du, M. Q. Tran, T. T. Nguyen, and H. N. Nguyen, "Multi-layer defense for ai-powered ids: Ensemble adversarial training and explainable resilience to evasion attacks," in *2025 2nd International Conference On Cryptography And Information Security (VCRIS)*, 2025, pp. 1–6.

[19] A. D. Vibhute, C. H. Patil, A. V. Mane, and K. V. Kale, "Towards detection of network anomalies using machine learning algorithms on the nsl-kdd benchmark datasets," *Procedia Computer Science*, vol. 233, pp. 960–969, 2024, 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924006458

[20] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3547330

[21] Y. Li, S. Zhang, and Y. Li, "Ai-enhanced resilience in power systems: Adversarial deep learning for robust short-term voltage stability assessment under cyber-attacks," *Chaos, Solitons & Fractals*, vol. 196, p. 116406, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077925004199

[22] J. L. Hernandez-Ramos, G. Karopoulos, E. Chatzoglou, V. Kouliaridis, E. Marmol, A. Gonzalez-Vidal, and G. Kambourakis, "Intrusion detection based on federated learning: A systematic review," *ACM Comput. Surv.*, vol. 57, no. 12, Jul. 2025. [Online]. Available: https://doi.org/10.1145/3731596

[23] V.-T. Hoang, Y. A. Ergu, V.-L. Nguyen, and R.-G. Chang, "Security risks and countermeasures of adversarial attacks on ai-driven applications in 6g networks: A survey," *Journal of Network and Computer Applications*, vol. 232, p. 104031, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S108480452400208X

[24] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, "Modeling realistic adversarial attacks against network intrusion detection systems," *Digital Threats*, vol. 3, no. 3, Feb. 2022. [Online]. Available: https://doi.org/10.1145/3469659

[25] H. Anis, "Github repository: Adversarial attacks on ai models in cyber-security," https://github.com/Huzi1080/CSM-Research-Paper/tree/main, 2025, accessed: November 2025.

[26] U. A. Ali, K. Dogra, and S. Sharma, "White-box adversarial exploitation of nids: Insights from fgsm, pgd, and c&w," in *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 2025, pp. 668–673.

[27] M. S. Ayas, S. Ayas, and S. M. Djouadi, "Projected gradient descent adversarial attack and its defense on a fault diagnosis system," in *2022 45th International Conference on Telecommunications and Signal Processing (TSP)*, 2022, pp. 36–39.

[28] A. Zyane and H. Jamiri, "Securing iot networks with adversarial learning: A defense framework against cyber threats," in *2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2025, pp. 1–7.

[29] S. Hore, J. Ghadermazi, A. Shah, and N. D. Bastian, "A sequential deep learning framework for a robust and resilient network intrusion detection system," *Computers & Security*, vol. 144, p. 103928, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404824002311

[30] I. Elgarhy, M. M. Badr, M. M. E. A. Mahmoud, M. M. Fouda, M. Alsabaan, and H. A. Kholidy, "Clustering and ensemble based approach for securing electricity theft detectors against evasion attacks," *IEEE Access*, vol. 11, pp. 112 147–112 164, 2023.

[31] M. Y. Abdelhakim and D. M. Krobba Ahmed, "Adversarial attacks for speaker recognition system with fgsm-cnn and fgsm-dnn," in *2024 International Conference on Telecommunications and Intelligent Systems (ICTIS)*, 2024, pp. 1–6.

[32] S. Longari, P. Cerracchio, M. Carminati, and S. Zanero, "Assessing the resilience of automotive intrusion detection systems to adversarial manipulation," *ACM Trans. Cyber-Phys. Syst.*, vol. 9, no. 3, Aug. 2025. [Online]. Available: https://doi.org/10.1145/3737294

[33] W. Song, X. Li, S. Afroz, D. Garg, D. Kuznetsov, and H. Yin, "Mab-malware: A reinforcement learning framework for blackbox generation of adversarial malware," in *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 990–1003. [Online]. Available: https://doi.org/10.1145/3488932.3497768

[34] T. Wang, T. Tang, Z. Cai, K. Fang, J. Tian, J. Li, W. Wang, and F. Xia, "Federated learning-based information leakage risk detection for secure medical internet of things," *ACM Trans. Internet Technol.*, Jan. 2024, just Accepted. [Online]. Available: https://doi.org/10.1145/3639466

[35] H. Nolte, M. Rateike, and M. Finck, "Robustness and cybersecurity in the eu artificial intelligence act," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 283–295. [Online]. Available: https://doi.org/10.1145/3715275.3732020

[36] M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model evasion attack on intrusion detection systems using adversarial machine learning," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, 2020, pp. 1–6.

[37] H. Waghela, J. Sen, and S. Rakshit, "Robust image classification: Defensive strategies against fgsm and pgd adversarial attacks," in *2024 Asian Conference on Intelligent Technologies (ACOIT)*, 2024, pp. 1–7.

[38] K.-Y. Lam, X. Liu, D. Wang, B. Li, W. Xu, J. Chen, M. Xue, X. Yuan, G. Bai, and S. Wang, "Lamps '25: Acm ccs workshop on large ai systems and models with privacy and security analysis," in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 4914–4915. [Online]. Available: https://doi.org/10.1145/3719027.3767670

[39] H. Gupta, A. P. Sundareswaran, R. Walia, S. Duggirala, S. Mylsamy, and A. Jain, "Optimizing cloud security with machine learning: Predicting and preventing vulnerabilities in distributed systems," in *2025 International Conference on Networks and Cryptology (NETCRYPT)*, 2025, pp. 1598–1602.

[40] K. Hemavathi and R. Latha, "Conditional generative adversarial network with optimal machine learning based intrusion detection system," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, 2023, pp. 1176–1182.

[41] H. Laiz-Ibanez, C. Mendaña-Cuervo, and J. L. Carus Candas, "The metaverse: Privacy and information security risks," *International Journal of Information Management Data Insights*, vol. 5, no. 2, p. 100373, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096825000552

[42] A. P. B P and S. N R, "A literature review on machine learning methods used in intrusion detection system to detect cyber attack," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*, 2024, pp. 94–97.

[43] S. Singhal, R. Srivastava, R. Shyam, and D. Mangal, "Supervised machine learning for cloud security," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, 2023, pp. 1–5.

[44] A. Schreiber and I. Schreiber, "Ai for cyber-security risk: harnessing ai for automatic generation of company-specific cybersecurity risk profiles," *Information and Computer Security*, vol. 33, no. 4, pp. 520–546, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2056496125000066