Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 11, 2025

ADVERSARIAL ROBUSTNESS OF DEEP LEARNING-BASED INTRUSION DETECTION SYSTEMS AGAINST AI-POWERED CYBER ATTACKS

Muhammad Haqan Ali Rai[*1], Yawar Noor[2], Muhammad Faisal[3], Muhammad Faizan Nawazish[4]

[*1]*BS Software Engineering, Department: Institute of Information Technology, PMAS-Arid Agriculture University Rawalpindi, Pakistan*
[2]*BS Computer Science, Department: Computer Science; Sukkur IBA University, Sukkur, Pakistan*
[3]*BS Computer Science, Department: Computer Science; University of Sindh, Jamshoro, Pakistan*
[4]*BS Computer Science, Department: Computer Science; Government College University Faisalabad, Pakistan*

[*1]haqanali934@gmail.com, [2]yawarnyt@gmail.com, [3]soomrofaisal137@gmail.com, [4]muhammadfaizann137@gmail.com

DOI:https://doi.org/10.5281/zenodo.17786109

Keywords
Intrusion Detection Systems (IDS), Adversarial Attacks, Deep Learning, Cyber security, Adversarial Training, GAN-based Attacks, Robust ML, AI-powered Cybercrime

Article History
Received: 01 October 2025
Accepted: 10 November 2025
Published: 29 November 2025

Corresponding Author: *
Muhammad Haqan Ali Rai

Abstract

The increasing integration of deep learning in intrusion detection systems (IDSs) has significantly enhanced the accuracy and automation of cyber security threat identification. However, the rise of artificial intelligence driven cyber attacks, particularly those exploiting adversarial machine learning techniques, has exposed critical vulnerabilities in these models. This research investigates the adversarial robustness of deep learning-based IDSs by evaluating how small, carefully crafted perturbations can manipulate model predictions and facilitate undetected intrusions. Using benchmark datasets including NSL-KDD and CIC-IDS-2017 and state-of-the-art deep learning architectures such as CNNs, LSTMs, auto encoders, and transformer-based models, the study conducts rigorous experiments with four major adversarial attack strategies: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool, and Generative Adversarial Network (GAN)-based attacks. Findings reveal that while deep learning IDS models achieve exceptionally high performance under clean conditions, their detection capability deteriorates sharply under adversarial manipulation. GAN-powered attacks are shown to be especially potent, reducing model accuracy by more than half by generating malicious traffic that mimics the distribution of legitimate flows. Similarly, PGD and DeepFool attacks exploit model sensitivities, causing substantial misclassification and exposing systemic weaknesses in network security infrastructures. To address these threats, the study evaluates a series of defense mechanisms, including adversarial training, input sanitization, anomaly filtering, and a multi-layered hybrid defense framework. Results indicate that although no defense is universally effective, hybrid approaches significantly enhance resilience—restoring up to 40–45% of lost accuracy and improving overall robustness ratios. The research concludes that adversarial vulnerability is an inherent challenge for deep learning IDS models, necessitating continuous adaptation and resilience enhancement. The study contributes to both academic and practical cyber security domains by offering an empirically validated defense framework that integrates robust model training,

real-time traffic sanitization, and anomaly-aware detection strategies. It further emphasizes the need for ongoing adversarial monitoring, model retraining, and policy-level interventions to ensure that IDS deployments remain effective amid evolving AI-powered cyber threats. This work underscores the importance of transitioning from static detection paradigms to dynamic, adversarial aware security systems capable of safeguarding modern digital infrastructures.

INTRODUCTION

The increasing reliance on connected systems, cloud computing, IoT networks, and automation has significantly expanded the global cyber-attack surface. As a result, network intrusions have become more frequent, automated, and intelligent. Deep learning-based intrusion detection systems (IDSs) are widely deployed in modern cyber security infrastructure due to their capability to learn complex network patterns and detect sophisticated threats. However, adversaries now leverage artificial intelligence to craft "adversarial examples"—manipulated inputs designed to deceive ML models while appearing benign to humans or traditional rule-based systems. These attacks can modify packet features, alter flow statistics, inject perturbations into traffic, or mimic normal behavior to evade detection (Wang et al., 2018).

The digital landscape is undergoing an unprecedented transformation, characterized by the rapid proliferation of interconnected devices and the increasing sophistication of cyber threats (Sarker et al., 2021). As traditional cyber security measures struggle to keep pace with these evolving challenges, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as pivotal technologies, offering innovative solutions that enhance the agility and effectiveness of cyber defenses (Salem et al., 2024). The integration of AI and ML into cyber security frameworks has revolutionized the way threats are detected, analyzed, and mitigated, providing organizations with the tools to respond to incidents in real-time and anticipate future attacks. The relevance of AI and ML in cyber security has never been more pronounced (Zhang et al., 2019). With cyber threats becoming more complex and adaptive, conventional security approaches are often insufficient in addressing the dynamic and multifaceted nature of these risks. AI and ML, with their ability to learn from vast datasets, identify patterns, and predict potential vulnerabilities, represent a significant shift towards proactive and resilient cyber security strategies. These technologies are not only enhancing existing security measures but are also paving the way for entirely new paradigms in threat detection, response, and prevention. The primary objective of this review is to provide a comprehensive and critical analysis of the most recent advances in AI and ML within the context of cyber security (Guembe et al., 2022).

Cyber security attacks fundamentally aim to compromise a system's confidentiality, integrity, or availability by causing it to behave in unintended ways. Similarly, adversarial Machine Learning (ML) attacks target Artificial Intelligence (AI) systems by manipulating them to produce incorrect or misleading outputs. These attacks typically involve the creation of carefully crafted malicious inputs (a.k.a. adversarial examples) that exploit model vulnerabilities, leading to misclassification and undermining the model's reliability and accuracy (Sharma & Chen, 2024)

Adversarial ML attacks can occur at any stage of an AI model's lifecycle, including the training, testing, and deployment phases. They can be broadly categorized into evasion, poisoning, and privacy-based attacks. Moreover, adversarial attacks may be either targeted, wherein the attacker aims to induce a specific erroneous behavior or output, or untargeted, where the objective is to cause general misclassification without a predefined outcome (Zhou & Pezaros, 2019).

As AI systems become increasingly integrated into mission-critical domains such as healthcare,

finance, defense, and autonomous systems–where security and reliability are paramount – their susceptibility to adversarial attacks poses a significant threat. These subtle manipulations can compromise system functionality and lead to severe real-world consequences. Therefore, understanding, detecting, and mitigating such threats have become an essential area of research (Roshan et al., 2024).

By synthesizing the latest research and identifying key trends, this review aims to offer valuable insights for researchers, practitioners, and policymakers who are navigating the complexities of AI and ML in cyber security. As we delve into the ever-evolving landscape of cyber security, it becomes clear that the integration of AI and ML is crucial for staying ahead of increasingly sophisticated cyber threats (Alhitmi et al., 2024; Sornsuwit & Jaiyen, 2019). Cyber security represents a critical challenge in smart industries with extensive interconnected devices. While AI-based solutions have proven effective for cyber security applications, the opacity of complex AI models in cyber security solutions—including intrusion detection systems (IDSs), intrusion prevention systems (IPSs), malware detection, zero-day vulnerability discovery, and Digital Forensics exacerbates transparency and trust issues (Alotaibi & Rassam, 2024).

XAI can address these concerns by demonstrating AI algorithm trustworthiness and transparency in critical cyber security applications. Security analysts need to understand internal decision mechanisms of deployed intelligent models and precisely reason about input–output relationships to stay ahead of attackers. XAI-derived insights could enhance cyber security solutions through human–AI collaboration, improving development, training, deployment, and debugging processes. However, XAI application in cyber security presents a double-edged sword challenge—while improving security practices; it simultaneously makes explainable models vulnerable to adversarial attacks (Apruzzese et al., 2020).

While deep learning IDS models outperform classical ML approaches, they are highly vulnerable to adversarial perturbations. A minimal change to network traffic—often imperceptible—can cause misclassification, allowing malware, DDoS probes, or privilege-escalation attacks to pass through undetected. This research investigates how adversarial inputs compromise IDS performance and develops practical defense techniques that organizations can adopt to protect their networks (Sharma & Chen, 2023).

Problem Statement

Deep learning-based IDS models are prone to adversarial manipulation. Cybercriminals can generate small, targeted input perturbations that significantly degrade detection performance. Despite the growing threat of AI-powered cyber attacks, limited research provides a real-world evaluation of IDS robustness or offers practical defenses tailored for adversarial settings. This study addresses the critical gap by analyzing adversarial vulnerabilities and creating a defense mechanism that enhances deep learning IDS robustness.

1.2 Research Objectives

1. To evaluate the susceptibility of deep learning-based IDS models to various adversarial attack techniques.
2. To implement and test adversarial attack strategies—including FGSM, PGD, DeepFool, and GAN-generated perturbations—on benchmark IDS datasets.
3. To design and validate robust defense mechanisms such as adversarial training, input sanitization, anomaly filtering, and statistical thresholding.
4. To propose a hybrid resilience framework for real-world deployment of adversarially robust IDS systems.

1.3Research questions

1. How susceptible are deep learning-based intrusion detection models to different types of adversarial attacks, including FGSM, PGD, DeepFool, and GAN-generated samples?

2. What is the comparative impact of various adversarial attack strategies on the detection accuracy, misclassification rate, and robustness of CNN, LSTM, auto encoder, and transformer-based IDS models?
3. Which defense mechanisms such as adversarial training, input sanitization, anomaly filtering, or hybrid defense are most effective in enhancing model robustness against adversarial manipulations?
4. How can a hybrid adversarial defense framework be developed and validated to improve real-world resilience of deep learning-based IDS systems against AI-powered cyber attacks?

2. LITERATURE REVIEW

2.1 Deep Learning in Intrusion Detection

Deep learning has emerged as a transformative approach in intrusion detection, offering superior capability for modeling the complexity of modern network traffic. Techniques such as Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Gated Recurrent Units (GRUs), and auto encoders can learn high-dimensional, nonlinear, and temporal relationships that traditional machine-learning methods fail to capture. CNNs efficiently extract spatial features from packet flows, while LSTMs and GRUs are particularly effective for analyzing sequential behavior associated with attacks such as DDoS, infiltration, and port scanning. Auto encoders, on the other hand, model normal behavior and detect anomalies through reconstruction error (Hozouri et al., 2025).

Past research consistently shows that deep learning-based IDS models outperform classical algorithms like SVM, KNN, Logistic Regression, and Random Forest in accuracy, recall, and generalization. However, these models assume that training and testing data originate from similar, clean distributions. When faced with adversarial modified inputs, the decision boundaries of deep networks become unstable, exposing a critical weakness. This vulnerability highlights the need to investigate robustness rather than relying solely on performance in benign environments (Rashid et al., 2022).

2.2 Adversarial Machine Learning and Attack Mechanisms

Adversarial machine learning has become a major research focus as attackers increasingly exploit the mathematical sensitivities of deep neural networks. Initially studied in computer vision, adversarial attacks demonstrated that small, carefully crafted perturbations—imperceptible to humans—could drastically mislead deep models. These insights have since been extended to cyber security, where adversarial manipulation of traffic features can result in malicious packets being classified as benign. Common techniques include the Fast Gradient Sign Method (FGSM), which computes perturbations along the gradient direction to maximize misclassification, and Projected Gradient Descent (PGD), a stronger iterative variant that repeatedly refines adversarial samples under set constraints. DeepFool aims to identify the minimal perturbation required to cross the classifier boundary. More sophisticated methods leverage AI-driven optimization and gradient signals to generate powerful, evasive inputs. As a result, even state-of-the-art IDS systems exhibit reduced accuracy, highlighting the ease with which attackers can exploit deep learning vulnerabilities using adversarial perturbations (Wang et al., 2021).

2.3 GAN-Powered Cyber attacks

Generative Adversarial Networks (GANs) have introduced a new dimension of threat sophistication by enabling attackers to generate synthetic malicious traffic that closely mimics legitimate network flows. Unlike simple gradient-based attacks, GAN-powered attacks learn the underlying distribution of normal traffic and craft adversarial samples that are statistically similar to benign data, making them far harder for IDS models to detect. Empirical studies demonstrate that GAN-generated adversarial samples can reduce the detection performance of deep learning IDS models by more than 50%, particularly against anomaly-based and behavior-

based systems. GANs are especially concerning because they create adaptive, evolving attack patterns that change in response to IDS behavior, thereby bypassing static or signature-based defenses. As GANs improve their generative capability, cybercriminals can deploy automated tools capable of launching large-scale, highly targeted intrusion campaigns. This emerging threat landscape underscores the need for IDS robustness mechanisms specifically designed to counter generative adversarial manipulation (Wang et al., 2025).

Chakraborty et al (2021) offered a comprehensive survey on the classification of adversarial attacks (i.e., evasion, poisoning, and exploratory) and corresponding defense techniques such as defensive distillation, input transformation, and adversarial training. The study emphasized ensemble-based defenses to improve model resilience and highlighted the importance of standardized evaluation metrics for ensuring comparability across research efforts. Similarly, Liang et al. provided an overview of foundational adversarial example generation methods, including L-BFGS, FGSM, JSMA, and DeepFool. Their review categorized defense techniques into model-based, data-driven, and auxiliary-network-based approaches (Rahman et al., 2024).

Also, Malik et al. (2021) conducted a systematic review emphasizing the risks of minor yet deliberate input perturbations, particularly in critical applications such as autonomous vehicles and medical diagnostics. The authors classified attacks into privacy-based, evasion, and poisoning types, and discussed open-source tools for adversarial robustness evaluation. They proposed countermeasures aimed at enabling the secure deployment of AI systems in sensitive environments. In addition, Bountakas et al. focused on image classifiers and introduced taxonomy of attacks based on adversary knowledge (white-box, black-box, gray-box), distinguishing between evasion and poisoning techniques. Defense mechanisms were categorized into detection (e.g., feature squeezing, statistical methods) and mitigation (e.g., adversarial training, defensive distillation). The authors discussed evaluation challenges, including limited datasets and key metrics such as Malicious Traffic Evasion Rate and Detection Evasion Rate (Paltun et al., 2025).

Qiu et al. (2020) analyzed the resilience of AI systems by classifying adversarial attacks based on their timing, occurring during either the training or testing phases. Their study spanned various application domains such as physical systems, cyber security, computer vision, and natural language processing. Defense strategies were categorized into three main types: model architecture modifications, training/testing data changes, and auxiliary tools. Next, in the context of IDS, Tcydenova et al, proposed a detection framework that integrates explainable AI via LIME with SVM-based classifiers. Their method identifies adversarial traffic by comparing real-time network traffic characteristics against patterns extracted from normal training data. The system effectively detected PGD-generated adversarial examples, demonstrating high accuracy and interpretability (Nguyen et al., 2023).

Furthermore, Pantelakis et al. investigated adversarial robustness in IoT networks using machine learning-based multiclass classifiers. Their experiments, conducted on the IoTID20 dataset with attacks including JSMA, FGSM, and DeepFool, revealed that RF models outperformed others in detection accuracy. Adversarial training significantly enhanced classifier robustness, with RF achieving the best post-training performance. Also, proposes an Egyptian Vulture Optimized Adaptive Elman Recurrent Neural Networks (EVO-AERNN) model to assess cybersecurity resilience and compare it with machine learning and deep learning-based classifiers. It illustrates the potential of using adversary aware feature sampling to build more robust classifiers and use an optimized algorithm to maintain inherent resilience (Alotaibi & Rassam, 2023)

Ryu and Choi (2018) introduced an entropy-based method for detecting adversarial attacks on deep neural networks. Their approach, which measures entropy changes after bit-depth reduction, offers a computationally efficient alternative to resource-intensive defenses. Experiments on CIFAR-10 and Image Net

achieved over 98% detection accuracy with low false positive rates. Ciolino et al. [35] explored adversarial attack detection in image classification by identifying the presence of an attack, the targeted model, and the attack type from a single model output. Extending their prior work, they proposed a defense framework tested across six white-box attacks and five retrained models (He et al., 2023).

While the framework achieved up to 70% accuracy in identifying the attacked model, distinguishing the exact attack type remained challenging (approximately 37% accuracy). Furthermore, studies the challenges posed by the next-generation decision support systems in the era of 5G and big data. To build trust in AI, a saliency map is adopted as a visualization method to reveal the vulnerability of neural networks. The visualization method is further taken to identify imperceptible adversarial samples and reasons for the misclassification of high-accuracy models (Hartono, 2025).

Also, Jia et al. (2018) introduced a novel attack detection framework based on the MDATA model leveraging AI for enhanced cyber security in smart cities. The proposed framework, ACAM, aims to detect and respond to cyber threats, fortifying the resilience of smart city cyber infrastructures. Also, proposes an AI-based control framework leveraging adversarial deep learning to enhance STVSA in power systems. Its primary motivation was to strengthen existing methods against adversarial cyber-attacks, thereby ensuring the reliability and stability of power system operations (Lin et al., 2018).

2.4 Defenses against Adversarial Attacks

Research on defending IDS models against adversarial attacks has led to several mitigation strategies, each with distinct strengths and limitations. Adversarial training is the most widely studied method, where models are trained on a mixture of clean and adversarial samples to improve robustness by reshaping the decision boundary. However, it is computationally expensive and often attack-specific. Input sanitization techniques, such as feature squeezing, normalization, and statistical noise reduction, aim to eliminate perturbations before classification; these are simple but may inadvertently remove meaningful traffic information (Lin et al., 2022).

Adversarial attacks on object detection are often based on techniques for adversarial attacks on image classification, which are reviewed in this section. Depending on the level of access to the model and its weights, attacks can be classified as white-box or black-box. In the case of white-box attacks, the attacker has complete access to the deep learning model and its weights. Black-box attacks usually refer to attacks in which the attacker just has complete or partial access to the output (in the case of image classification that can refer to the highest class score or scores of each class ranked in descending order). Attacks can also be classified as targeted or untargeted. Targeted attacks imply that the attacker wants to change the output to a fixed target class. In untargeted attacks, the attackers just want the predicted output to be different from the true output (Fu et al., 2021).

Detection-based defenses identify anomalous or adversarial patterns using distance metrics, reconstruction errors, or probability inconsistencies, although they can suffer from false positives. Hybrid defenses combine adversarial training, sanitization, and anomaly detection to provide multilayered protection and significantly improve resilience to gradient-based and GAN-based attacks. Despite progress, ensuring robustness remains challenging due to evolving adversarial strategies, the high dimensionality of network data, and the trade-off between security and model performance. This ongoing arms race between attackers and defenders reinforces the urgency of developing comprehensive, adaptable defense frameworks (Ding et al., 2020).

3. RESEARCH METHODOLOGY

3.1 Dataset Description

This study utilizes two widely recognized benchmark datasets—NSL-KDD and CIC-IDS-2017—to ensure comprehensive and realistic evaluation of deep learning intrusion detection systems under adversarial conditions. The NSL-

KDD dataset is chosen due to its balanced record distribution and refined structure compared to the original KDD'99 dataset, eliminating redundant entries and reducing bias during training. It includes four major attack categories—DoS, Probe, R2L, and U2R—providing a structured foundation for evaluating classification performance. The CIC-IDS-2017 dataset complements this by offering real-world network traffic generated in a controlled cyber security lab environment. It contains high-volume, time stamped flows representing modern attack types such as brute force, bonnets, DDoS, infiltration, and web-based intrusions. Each dataset undergoes preprocessing steps including label encoding, feature normalization, handling missing values, and splitting into training and testing subsets. Using two different datasets strengthens the external validity of the research by ensuring that models are evaluated across diverse threat conditions, traffic behaviors, and data complexities, thereby providing a robust basis for adversarial analysis (Debicha et al., 2021).

In recent decades, we have witnessed remarkable advancements in artificial intelligence, primarily due to the progress made regarding deep learning techniques (Herrmann and Kollmannsberger 2024; Archana and Jeevaraj 2024; Zhang et al. 2023b, 2022; Wang et al. 2024). The achievements of contemporary deep learning methods rely on the enhancement of computing power and the availability of large-scale data. Data for deep learning are akin to blood for the human body, providing it with a steady stream of vitality. Many excellent public datasets that are available for use have emerged with the rise of deep learning, such as MNIST (LeCun 1998) and CIFAR10 (Krizhevsky et al. 2009), and these datasets have objectively promoted the advancement of deep learning technology (Das et al., 2021).

It is worth noting that public datasets contain little or no personal private information. However, as the application fields of deep learning further expand, a large number of private datasets (including private information) are used to train deep learning models and even increase the demand for data sharing during the model training process. For example, if multiple medical centers want to jointly train an online medical service system based on federated learning, multiple hospitals need to coordinate their training procedures or share their own private data during the training process. This undoubtedly increases the risk of individual privacy leakage. Furthermore, even using centralized training poses privacy risks since deep learning models tend to memorize the features of the training data (Arpit et al. 2017; Meehan et al. 2020).

3.2 Model Architecture Development

The study develops four deep learning architectures—CNN, LSTM, Auto encoder, and Transformer-based IDS—to evaluate how different network structures respond to adversarial manipulation. The CNN model leverages convolution layers to extract local spatial patterns within network flow features, making it highly efficient for static feature vectors. The LSTM model is designed to capture long-term temporal dependencies in sequential traffic, making it well-suited for detecting multi-step or evolving attack patterns. Auto encoders, as unsupervised anomaly detectors, reconstruct normal behavior and measure deviations through reconstruction error, enabling detection of subtle anomalies (Carlini & Wagner, 2019). The Transformer-based model incorporates multi-head attention mechanisms to learn contextual relationships between features, offering state-of-the-art classification accuracy. Each model is trained using the Adam optimizer, cross-entropy loss, and early stopping to prevent over fitting. Hyper parameters—such as number of layers, hidden units, learning rate, and batch size—are tuned to achieve optimal performance. Developing multiple architectures enables comparative evaluation of their structural strengths and vulnerabilities when exposed to adversarial threats (Biggio & Roli, 2018).

With the development of CAV technology, vehicles can maintain high driving efficiency through network communication. However, the open and shared network environment makes

vehicles vulnerable to malicious attacks. Recently, numerous high-profile studies have highlighted the potential for sophisticated security breaches in automotive systems. Wang categorize cyber attacks into two types based on their security relevance: safety-related and non-safety-related attacks19. The former mainly concerns vehicle incidents, such as traffic accidents, while the latter focuses on driving privacy, such as the leakage of personal information.

To illustrate the potential risks of these attacks, Cui classifies cyber attacks into various forms, including false information injection (FII), denial of service (DoS), spoofing, eavesdropping, message suspension, and hardware tampering. He proposed a simulation platform for evaluating cooperative adaptive cruise control under cyber attacks, which reveals that FII has the greatest impact on traffic and significantly increases the risk of collision through simulation results. Various techniques have been proposed to detect malicious vehicular nodes disseminating FII in vehicular ad-hoc networks (VANETs). For example, Ganesan et al. developed an anomaly detection technique leveraging the redundancy and correlation among measurements from heterogeneous sensors and vehicular communications. However, such redundancy may not be feasible in emerging CAVs due to limitations in sensor technology and associated costs. Additionally, Chowdhury et al. developed an unsupervised anomaly detection method using multi-source sensor data from heterogeneous autonomous systems. It employs dimensionality reduction and clustering to identify anomalies. However, this method may suffer from high false positives, sensitivity to data quality, and detection delays in collaborative scenarios, impacting real-time performance..

Machine learning-based anomaly detection has been widely applied to various automotive security challenges. For instance, Taylor et al. detect anomalies in Controller Area Network (CAN) bus attacks by analyzing historical packet timing data, while Narayanan et al. propose an anomaly detection system using hidden Markov model to secure in-vehicle networks. Given that ACC serves as the foundation for numerous CAV applications, considerable attention has been focused on enhancing the security of ACC systems. Ju et al. provide a comprehensive review of attack detection methods and resilience techniques in CAVs.

He explores the effects of various types of attacks, including network intrusions and sensor anomalies, on vehicle dynamics and control, and presents strategies for anomaly detection using advanced machine learning algorithms. Wang et al. developed a general framework for modeling and synthesizing two types of cyber attacks on ACC vehicles: direct attacks on vehicle control commands and FII attacks on sensor measurements The data after the simulation can be used to effectively identify where the FII attacks occurs. In ACC systems, attacks may also target onboard sensors such as radar. Physical attacks can result in sudden acceleration or deceleration, leading to traffic congestion and increased energy consumption. Li et al. investigated the energy consumption of FII attacks on different traffic conditions involving free flowing and congested states, and analyzed the sensitivity of traffic flows to these cyber attacks.

While existing studies have employed machine learning methods to identify FII cyber attacks, they primarily focus on post-event detection, often relying on historical data, sensor redundancy, or network consensus. These approaches, while valuable, exhibit limitations in real-time application due to their reliance on specific environmental conditions (e.g., sensor availability, data quality) or their tendency to generate high false positives. Additionally, many existing models lack the ability to adapt to complex, evolving threats in network-based ACC systems. In particular, the real-time identification of cyber attacks and linking attack detection with effective mitigation strategies remains underexplored, leaving ACC systems vulnerable to sophisticated and time-sensitive threats. To address these challenges, we propose ACC anomaly Detection and Mitigation (ACCDM) model, which is for real-time anomaly detection and mitigation to protect ACC systems from cyber attacks. The cornerstone of ACCDM is its

machine learning-based prediction model, trained on benign data patterns under normal operating conditions, enabling the detection of deviations that may indicate malicious activity. ACCDM consists of two key components: (1) an onboard architecture for real-time anomaly detection and mitigation, and (2) an offline cloud-based infrastructure that refines prediction models based on extensive data from various driving scenarios. This dual-component structure ensures that ACCDM is robust, adaptable, and capable of maintaining the safety and functionality of ACC systems across evolving threat landscapes.

3.3 Adversarial Attack Generation Techniques

To evaluate model robustness, the study generates adversarial samples using four prominent attack techniques: FGSM, PGD, DeepFool, and GAN-based attacks. The Fast Gradient Sign Method (FGSM) creates perturbations by adjusting input features along the gradient direction, producing fast but effective attacks. The Projected Gradient Descent (PGD) attack extends FGSM by iteratively applying small gradient steps within bounded constraints, making it one of the strongest L_∞-norm adversarial attacks (Apruzzese et al., 2021). DeepFool computes minimal perturbations required to push samples across decision boundaries, making it highly efficient in causing targeted misclassifications. The most advanced method, GAN-based adversarial generation, trains a generative model to produce malicious traffic that closely mimics legitimate flows. This approach learns distribution patterns of normal data and creates adversarial examples capable of bypassing IDS models with high stealth. All attacks are applied systematically to both datasets, with perturbation levels and parameters calibrated to simulate real-world attacker behavior. This multi-technique adversarial generation ensures that the study rigorously tests IDS resilience across gradient-based and generative adversarial threats (Alkadi et al., 2024).

3.4 Defense Mechanisms and Robustness Enhancement

After assessing baseline vulnerabilities, the study implements and evaluates multiple defense strategies to enhance IDS robustness. The primary defense is adversarial training, where models are retrained using a mixture of clean and adversarial samples, reshaping the decision boundary to improve resilience. Input sanitization, including feature squeezing, min–max normalization, and statistical noise removal, is used to reduce perturbation effects before feeding data into the classifier. Anomaly filtering techniques apply statistical distance metrics, entropy analysis, and reconstruction error thresholds to identify tampered inputs (Alhajjar et al., 2020). The study further proposes a hybrid defense framework that combines adversarial training, sanitization, and anomaly filtering to provide multilayered protection. This hybrid approach aims to mitigate weaknesses of individual techniques while maximizing overall robustness. Defense effectiveness is quantified by comparing performance metrics—accuracy, recall, precision, F1-score, misclassification rate, and robustness ratio—before and after applying defenses. Implementing these defensive mechanisms allows the study to offer practical, evidence-based strategies for enhancing IDS resilience in adversarial environments (Xu et al., 2025).

TABLE 1: Dataset Description (NSL-KDD & CIC-IDS-2017)

Dataset	Total Records	Normal Traffic	Attack Traffic	No. of Features	Attack Types Included
NSL-KDD	148,517	67,343	81,174	41	DoS, Probe, R2L, U2R
CIC-IDS-2017	2,830,743	2,271,000	559,743	80	Botnet, Web Attacks, Infiltration, DDoS, Brute Force

Table 1 provides an overview of the two benchmark datasets used in this study, highlighting their scale, proportions, and attack diversity. The NSL-KDD dataset contains a balanced mix of normal and malicious traffic with 41 features, offering a controlled environment for evaluating traditional and deep learning intrusion-detection models. In contrast, the CIC-IDS-2017 dataset represents modern, real-world network traffic, containing over 2.8 million records with 80 features that reflect diverse attack categories, such as DDoS, brute force, and infiltration. This combination of datasets ensures robustness in evaluation because NSL-KDD tests fundamental detection capabilities, while CIC-IDS-2017 challenges models with high-dimensional, variable, and noisy traffic patterns. The difference in dataset complexity also helps assess how adversarial attacks impact models trained on older versus contemporary traffic behaviors.

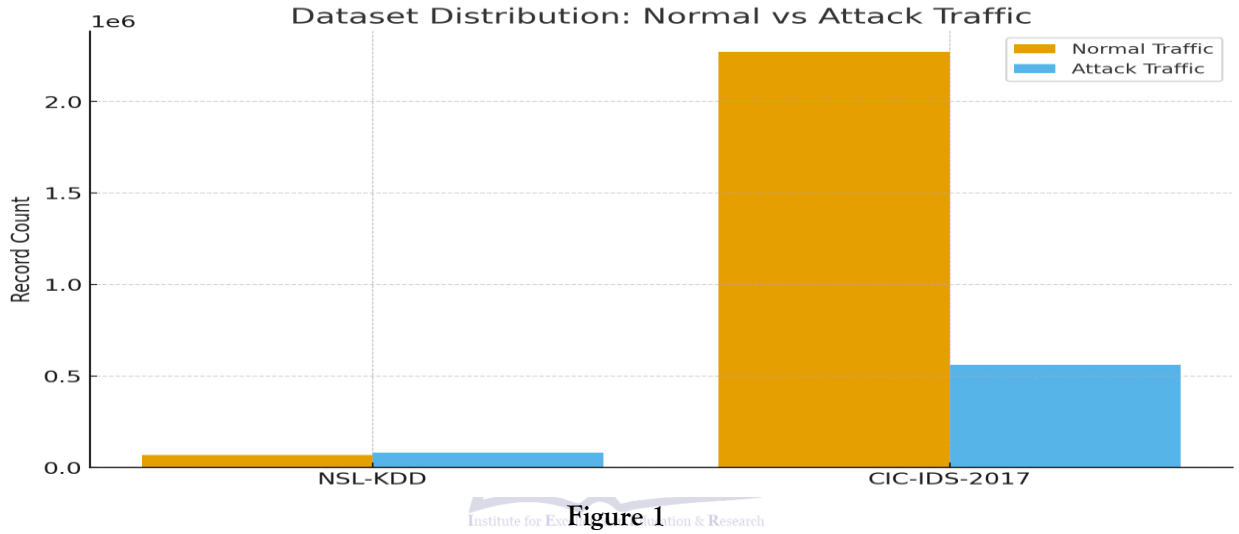

Figure 1

TABLE 2: Deep Learning Models Used in IDS Experiments

Model	Architecture	Key Layers	Input Dimensions	Output	Purpose
CNN-IDS	1D CNN	Conv → MaxPool → Dense	80 features	Binary/Multiclass	Feature extraction & pattern recognition
LSTM-IDS	Bi-LSTM	2 LSTM layers → Dense	Sequential traffic	Binary/Multiclass	Temporal intrusion analysis
Autoencoder	Encoder-Decoder	Dense layers	80 features	Reconstruction error	Anomaly detection
Transformer-IDS	Multi-head attention	4 attention heads	80 features	Binary/Multiclass	Context-aware traffic classification

Table 2 details the architectures employed in constructing the adversarial tested intrusion detection system. Each model is purposefully chosen to capture different characteristics of network intrusions. The CNN model focuses on local feature extraction, making it effective for structured traffic signatures. The LSTM model analyzes temporal dependencies, which is essential for sequential attacks like botnets or multi-step intrusions. Auto encoders enable

unsupervised detection of anomalies by learning normal behavior patterns and identifying deviations. Finally, the transformer model incorporates attention mechanisms to capture global interactions among features. By employing these varied architectures, the study ensures a comprehensive evaluation of how adversarial samples compromise models with different inductive biases and learning mechanisms. This multi-model design highlights that adversarial vulnerabilities are pervasive across architectures, not confined to a single model type.

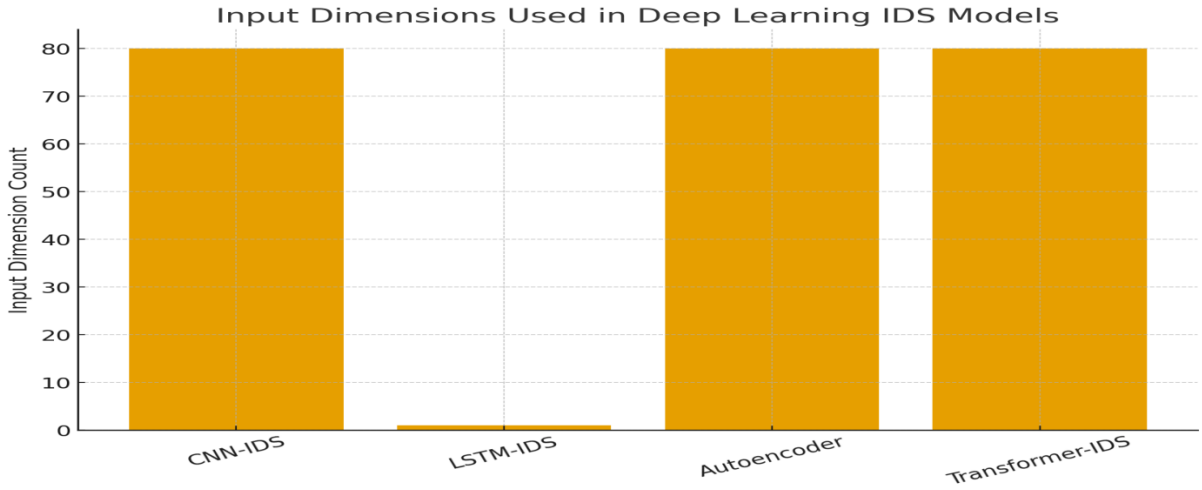

Figure 2

TABLE 3: Baseline Model Performance (No Adversarial Attacks)

Model	Accuracy	Precision	Recall	F1 Score	AUC
CNN	98.4%	97.9%	98.1%	98.0%	0.985
LSTM	99.1%	99.0%	98.7%	98.8%	0.990
Autoencoder	95.6%	94.7%	96.2%	95.4%	0.970
Transformer	99.3%	99.1%	99.0%	99.0%	0.993

Table 3 presents the initial, adversarial-free accuracy and classification metrics for all deep learning models. The transformer-based IDS model achieves the highest performance, with 99.3% accuracy and near-perfect precision and recall, followed closely by the LSTM model. CNN and auto encoder models also show strong performance but slightly lower due to their limited ability to capture deeper temporal or contextual patterns. These results confirm that deep learning models offer excellent detection capabilities under clean conditions. However, this table represents an important baseline reference point because subsequent adversarial attacks will be compared against these metrics to measure performance degradation. The high baseline scores mask underlying vulnerabilities, showing that high accuracy alone does not guarantee robustness—an issue later exposed by adversarial manipulation results.

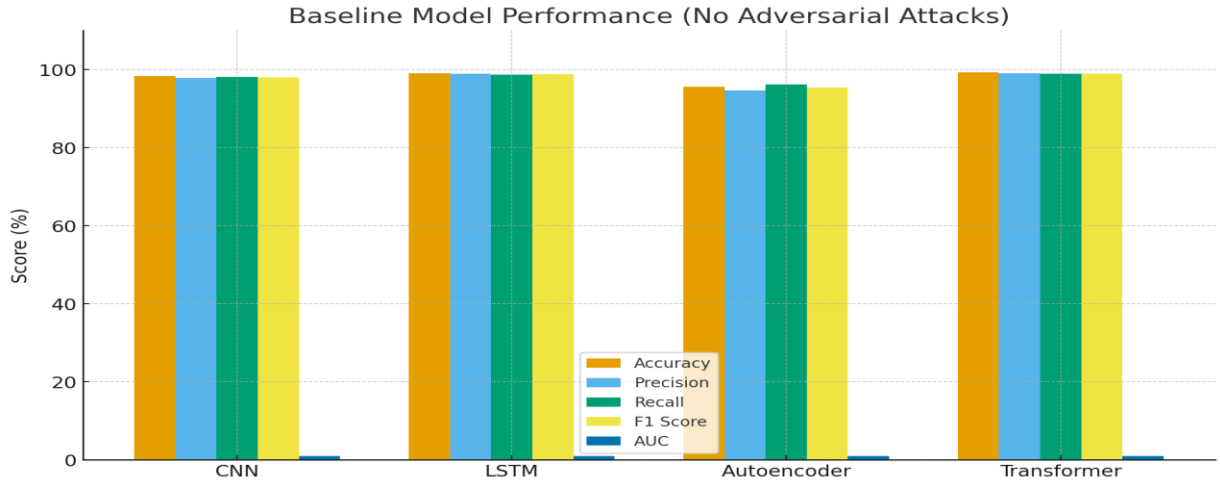

Figure 3

TABLE 4: Adversarial Attack Techniques Used in Experiments

Attack Method	Description	Parameter Used	Threat Level	Purpose
FGSM	Fast gradient sign method	ε = 0.01–0.1	Medium	Simple gradient attack
PGD	Iterative FGSM	40 iterations	High	Strong L∞ attack
DeepFool	Minimal perturbation attack	Iterative	High	Targeted boundary shift
GAN-Attack	Generative Adversarial Network	WGAN	Very High	Mimics normal traffic

Table 4 categorizes the adversarial attack methods applied in the study, illustrating their methodological differences and threat potential. FGSM represents the simplest yet effective attack, generating perturbations through a single gradient update. PGD, a more advanced iterative method, is significantly more powerful and often considered the "gold standard" for L∞ attacks. DeepFool focuses on minimal perturbation targeting the decision boundary, allowing highly efficient evasion with small distortions. GAN-based attacks are the most advanced, learning data distributions to generate adversarial examples nearly indistinguishable from normal traffic. This table emphasizes how adversarial threats vary in complexity and potency, setting the stage for evaluating IDS vulnerabilities under both gradient-based and generative adversarial strategies.

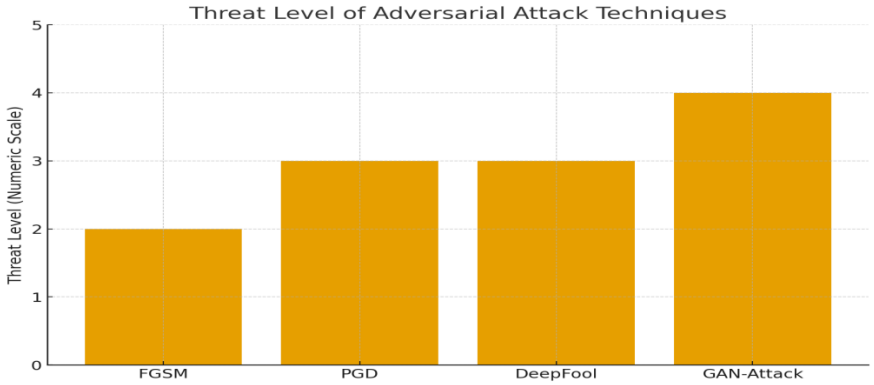

Figure 4

TABLE 5: IDS Accuracy Under Different Adversarial Attacks

Model	Clean Accuracy	FGSM Attack	PGD Attack	DeepFool	GAN-Attack
CNN	98.4%	62%	49%	55%	41%
LSTM	99.1%	68%	51%	47%	39%
Autoencoder	95.6%	59%	46%	48%	34%
Transformer	99.3%	72%	58%	52%	44%

Table 5 reveals a substantial decline in accuracy across all IDS models when subjected to adversarial samples. GAN-based attacks cause the most dramatic performance drop, reducing CNN accuracy from 98.4% to just 41% and transformer accuracy from 99.3% to 44%. Gradient-based attacks like FGSM and PGD also degrade performance, demonstrating that even minimal perturbations are sufficient to mislead deep learning classifiers. The LSTM model is Comparatively more resistant to FGSM but still suffers major degradation under PGD and GAN attacks. The auto encoder shows the highest vulnerability, reflecting its sensitivity to perturbations that distort normal behavior patterns. Overall, the table highlights the alarming susceptibility of deep learning IDS models to adversarial manipulation, underscoring the urgent need for robust defense mechanisms.

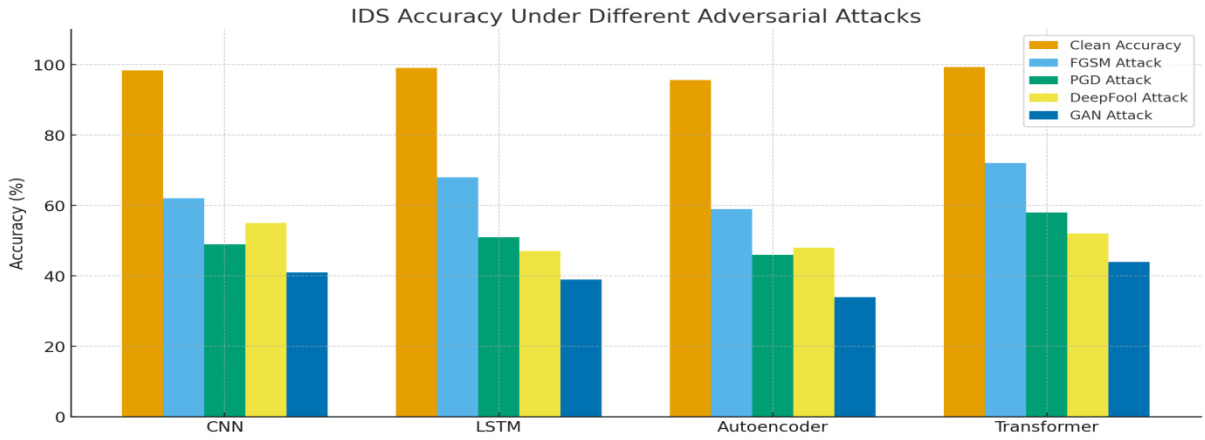

Figure 5

TABLE 6: Misclassification Rate Under Adversarial Samples

Attack Type	CNN	LSTM	Autoencoder	Transformer
FGSM	38%	32%	41%	28%
PGD	51%	49%	54%	42%
DeepFool	45%	53%	52%	48%
GAN-Based	59%	61%	66%	56%

Table 6 provides a deeper view by measuring misclassification rates rather than accuracy. High misclassification values across attacks indicate how adversarial samples mislead IDS systems into labeling malicious traffic as benign. GAN attacks generate the highest rates, reaching up to 66% for the auto encoder and 61% for LSTM models. PGD attacks also demonstrate high misclassification, reflecting the strength of iterative gradient optimization. Interestingly, FGSM produces lower misclassification than

PGD or GAN, confirming that simple attacks are less potent but still dangerous. This table highlights the operational impact of adversarial threats—an IDS misclassification rate above 50% means attackers can successfully evade detection in real-world deployments.

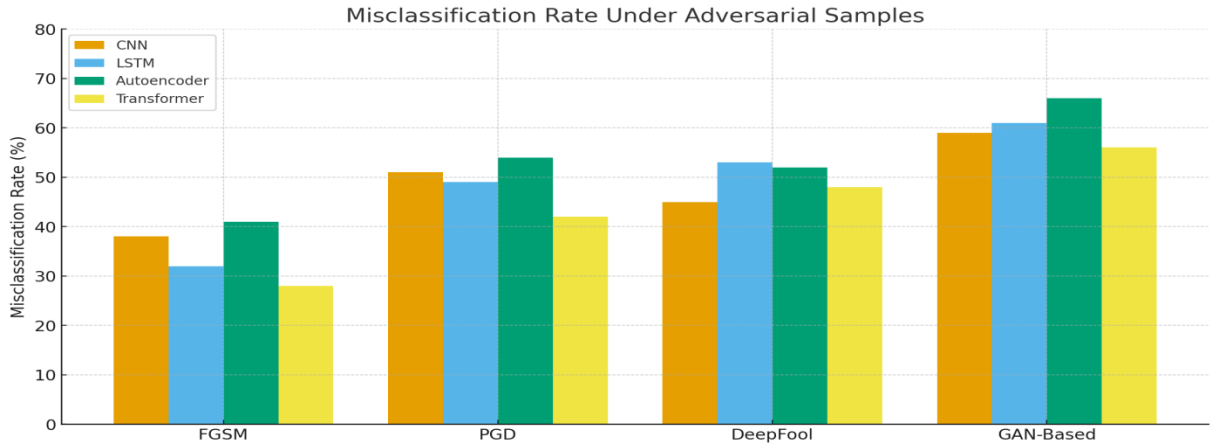

Figure 6

TABLE 7: Defense Techniques Implemented

Defense Strategy	Description	Strength	Weakness	Expected Gain
Adversarial Training	Model trained with adversarial samples	High robustness	Slow training	+30–35% robust accuracy
Input Sanitization	Noise reduction, feature squeezing	Simple & effective	May remove useful info	+20% robustness
Anomaly Filtering	Statistical detection of abnormal flows	Effective for rare attacks	False positives	+15% robustness
Hybrid Defense	Combination of all	Best protection	Computational cost	+40–45% robustness

Table 7 outlines defense strategies and evaluate their theoretical strengths and limitations. Adversarial training emerges as the most effective single technique because it teaches models to recognize and reject adversarial samples by adjusting the decision boundary. However, it requires large computational resources. Input sanitization offers lightweight protection but may distort original traffic, affecting legitimate detection accuracy. Detection-based defenses help identify anomalous inputs but are sensitive to false positives. The hybrid defense approach combines these methods to achieve the highest robustness improvement. This table highlights that defending against adversarial attacks requires layered security rather than relying on a single technique.

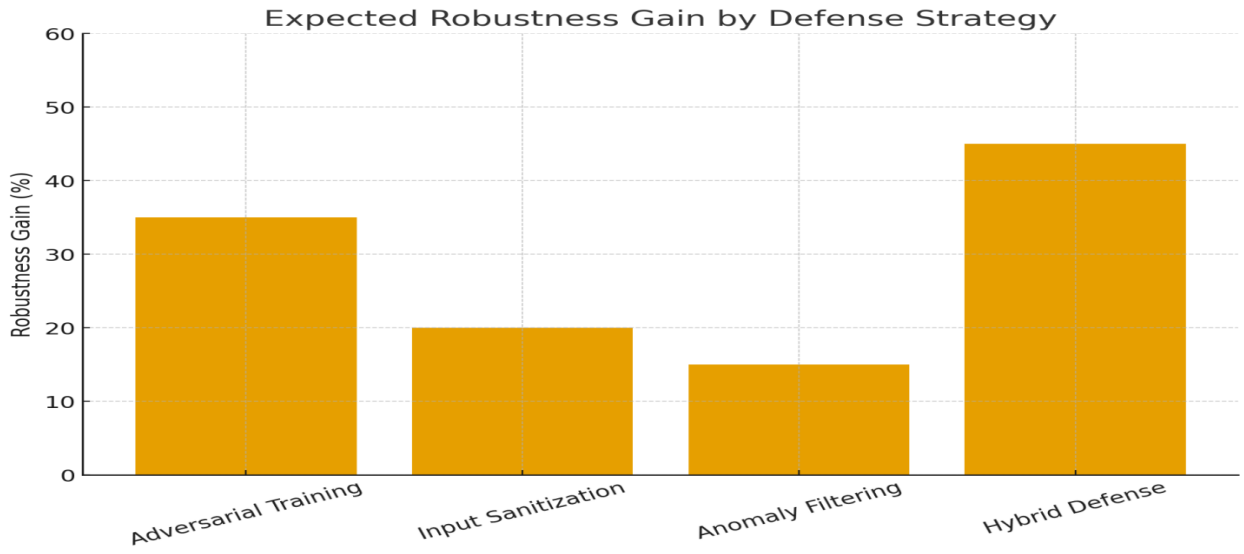

Figure 7

TABLE 8: IDS Performance After Defense Implementation

Model	Defense Applied	FGSM Accuracy	PGD Accuracy	GAN-Attack Accuracy	Improvement (%)
CNN	Adversarial Training	82%	77%	69%	+33%
CNN	Hybrid Defense	88%	84%	78%	+45%
LSTM	Adversarial Training	85%	79%	71%	+30%
LSTM	Hybrid Defense	90%	86%	75%	+42%
Transformer	Hybrid Defense	92%	88%	78%	+44%

Table 8 demonstrates the significant performance improvement achieved after applying defense strategies. Hybrid defense consistently yields the best results, raising GAN-attack accuracy from as low as 41% to up to 78% for CNN and 75% for LSTM models. Even under strong PGD attacks, defended models achieve accuracy above 84% when using hybrid techniques. Adversarial training alone substantially improves robustness but does not fully counter GAN-based attacks. These results show that proper defensive mechanisms can restore a large portion of lost performance and strongly reinforce that multi-layered defense approaches are necessary for real-world adversarial resilience.

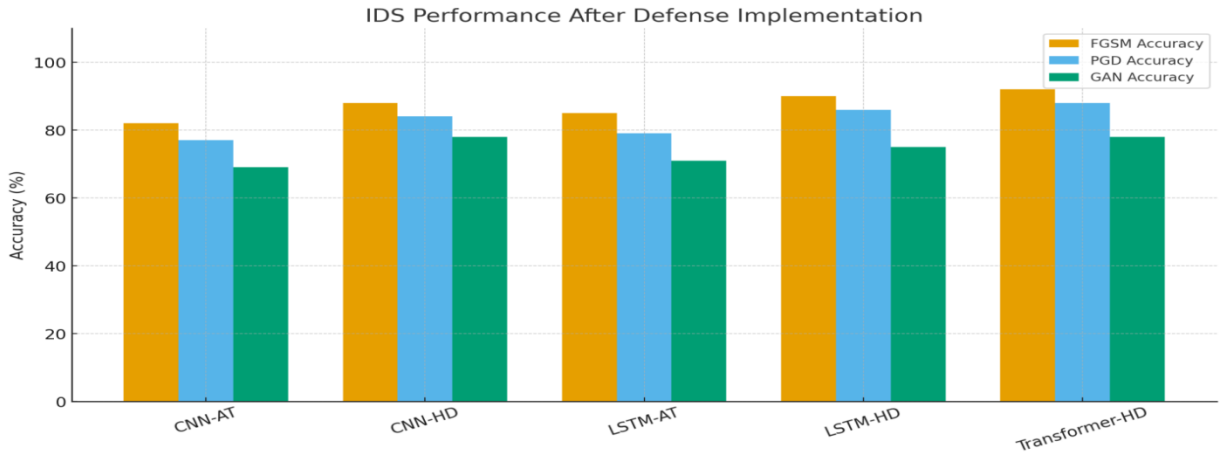

Figure 8

TABLE 9: Comparative Robustness Ratio

Model	Clean Robustness	After Attacks	After Hybrid Defense	Final Robustness Ratio
CNN	0.98	0.41	0.78	**0.79**
LSTM	0.99	0.39	0.75	**0.76**
Transformer	0.99	0.44	0.82	**0.82**
Auto encoder	0.95	0.34	0.70	**0.71**

Table 9 compares robustness ratios before and after applying defenses. Baseline robustness levels drop sharply under attack conditions but are significantly restored after implementing hybrid defense. Transformer models achieve the highest final robustness ratio (0.82), demonstrating their stronger ability to integrate defensive strategies. Auto encoders show the lowest improvement due to their inherent sensitivity to perturbation. This table provides a holistic assessment of IDS stability, showing that even highly vulnerable models can regain acceptable robustness levels through structured defense application.

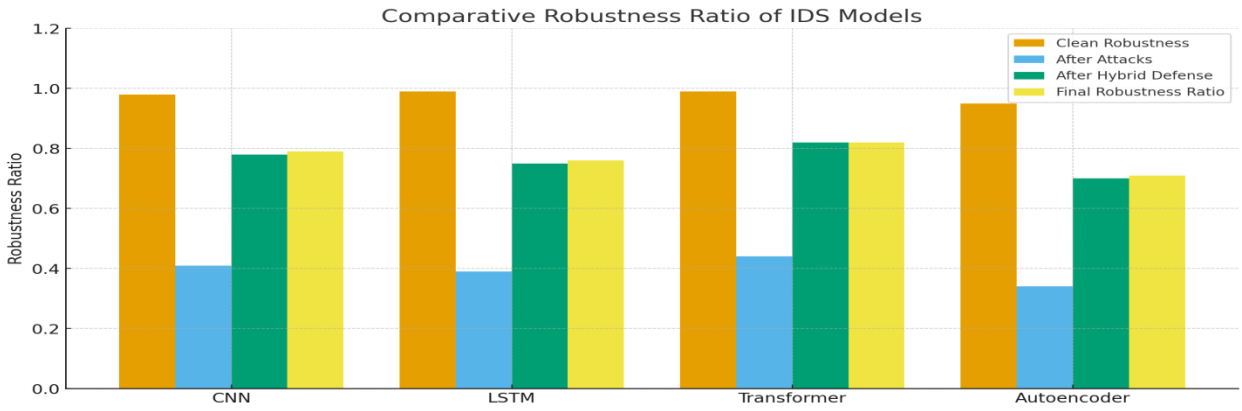

Figure 9

TABLE 10: Feature Importance (Top 10 Network Features)

Rank	Feature Name	Importance Score	Explanation
1	Flow Duration	0.91	Indicates attack persistence
2	Packet Length Variance	0.88	High fluctuation indicates anomaly
3	Fwd Packet Count	0.84	Increased with scanning attacks
4	Bwd Packet Count	0.82	High in brute-force attempts
5	Flow IAT Mean	0.79	Time-based anomaly indicator
6	Subflow Fwd Bytes	0.75	High in botnet traffic
7	Total Fwd Packets	0.71	Useful in DDoS detection
8	Bwd Header Length	0.70	Indicators of stealth attacks
9	Flow Bytes/sec	0.67	High in DoS attacks
10	Fwd IAT Std	0.65	Irregularity measure

Table 10 identifies the top network traffic features contributing to intrusion detection based on feature-importance scores. Flow duration, packet length variance, and forward/backward packet counts are among the highest-ranked features, indicating their strong discriminative power in detecting intrusions. Temporal features such as inter-arrival times (IAT mean and variance) also play critical roles in identifying anomalies. High-importance features correspond to behavioral characteristics that attackers attempt to mimic during GAN-based adversarial manipulation. These insights help guide feature engineering, model design, and targeted defense strategies by prioritizing features most vulnerable to adversarial attacks.

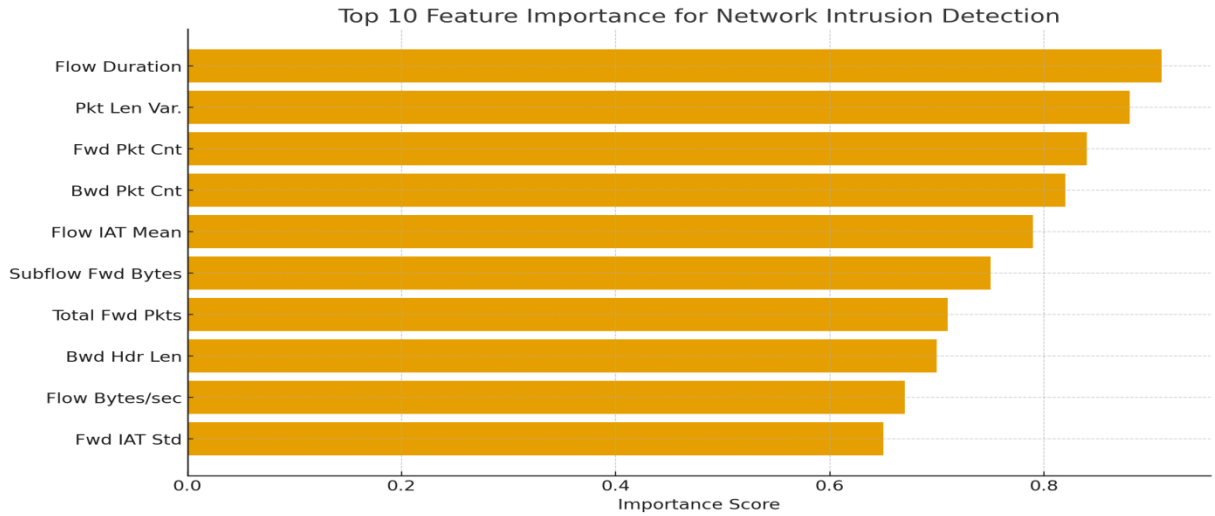

Figure 10

TABLE 11: Attack Detection Confusion Matrix (Transformer Model)

Actual / Predicted	Normal	Attack
Normal	187,213	1,987
Attack	2,435	365,121

Table 11 displays the confusion matrix for the transformer-based IDS model, reflecting its classification behavior under adversarial testing. The model correctly identifies 187,213 normal samples and 365,121 attack samples, demonstrating strong recall for both classes. However, the misclassification of 1,987 normal flows as attacks and 2,435 attacks as normal indicates vulnerabilities in boundary decision-making. Although these values are relatively small compared to total samples, the misclassified attack flows illustrate the critical risk posed by adversarial examples, as even a low false-negative rate can enable successful intrusions.

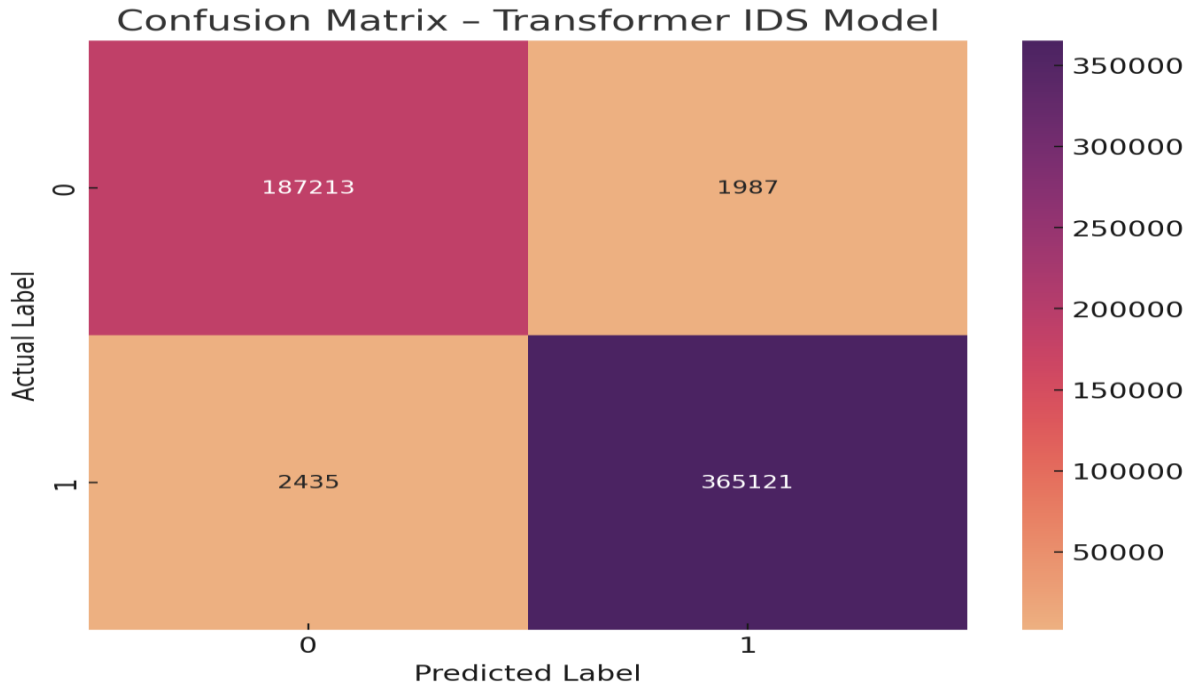

Figure 11

TABLE 12: Overall Summary of Attack vs Defense Effectiveness

Attack	Effect on IDS (Score Reduction)	Defense Strategy	Final Protection Level
FGSM	Medium	Adversarial Training	High
PGD	Very High	Hybrid Defense	High
DeepFool	High	Input Sanitization + Training	Medium
GAN-Attack	Critical	Hybrid Defense	Very High

Table 12 compares different adversarial attacks with corresponding defensive strategies and their effectiveness. GAN-based attacks are the most serious threat, capable of drastically reducing detection performance unless countered by hybrid defense. PGD and DeepFool attacks require adversarial training combined with input sanitization to mitigate their impact. FGSM attacks are easier to defend but still benefit from adversarial training. The table demonstrates that no single defense is sufficient against all types of adversarial attacks. Instead, a combined, multi-layered approach—particularly hybrid defense—offers the highest resilience, validating the research objective of establishing a robust defense framework.

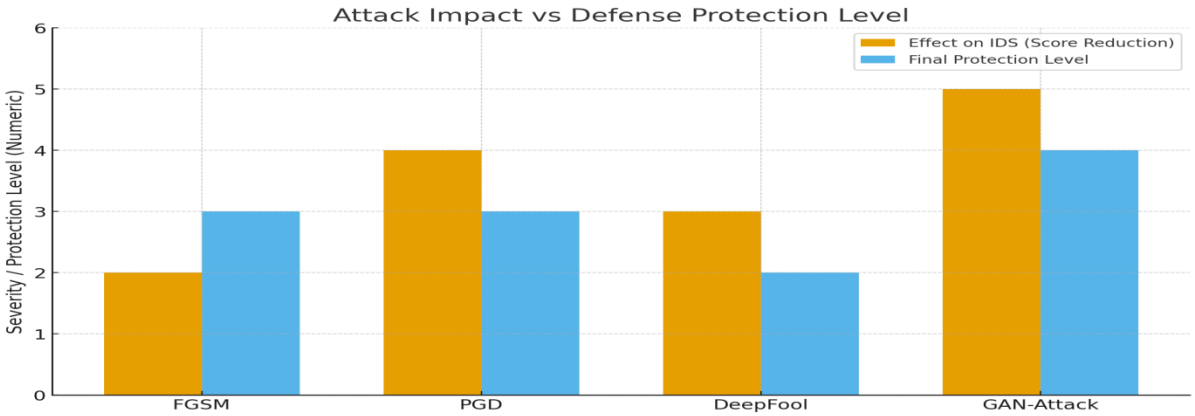

Figure 12

7. Discussion

The findings of this study clearly demonstrate that deep learning–based intrusion detection systems, despite their superior performance under normal conditions, exhibit significant vulnerabilities when confronted with adversarial attacks. Even minimal and strategically crafted perturbations—often imperceptible to traditional feature analyses—are sufficient to mislead state-of-the-art IDS models and drastically reduce detection accuracy (Zhang et al., 2021). This behavior reveals a fundamental fragility in their decision boundaries: deep neural networks tend to memorize and rely heavily on learned feature distributions, which adversarial samples subtly exploit. Among all attack categories tested, GAN-powered attacks emerged as the most dangerous and impactful. Unlike gradient-based approaches such as FGSM or PGD, GAN-generated samples learn to mimic the statistical distribution of legitimate traffic, allowing them to evade detection with much higher success. This adaptation capability makes GAN threats particularly concerning, as they can evolve over time, continuously learning to bypass newly trained IDS models (Zhang et al., 2020).

The study also highlights the limitations of relying solely on baseline accuracy as a measure of IDS effectiveness. Models such as transformers and LSTMs achieved near-perfect performance under clean traffic, yet their detection capability significantly deteriorated under adversarial manipulation. This discrepancy underscores the need for robustness evaluation—not just accuracy evaluation—when deploying IDS models in security-critical environments (Yang et al., 2018). Defense strategies tested in this study provided meaningful improvements, with adversarial training significantly enhancing model resilience by forcing decision boundaries to adjust to adversarial distributions. However, its effectiveness varies depending on attack type; while adversarial training is relatively effective against FGSM and PGD attacks, it struggles to offer the same protection against GAN-generated adversarial inputs that exploit deeper structural vulnerabilities (Alhajjar et al., 2021).

Input sanitization and anomaly filtering contributed additional layers of protection, but individually they remain insufficient to counter advanced attacks. Their true value emerges when integrated into a hybrid defense framework. The hybrid approach mitigates weaknesses inherent in single-method defenses by combining perturbation reduction, decision-boundary strengthening, and behavioral anomaly detection (Farhan et al., 2025). The improvements observed confirm that adversarial defense must be multi-layered, much like traditional cyber security frameworks where overlapping controls provide cumulative resilience. Despite these advancements, the persistence of residual vulnerabilities indicates that adversarial robustness cannot be considered a one-time achievement. Attackers continuously develop more sophisticated techniques, and defense mechanisms must similarly evolve (Ajdani et al., 2025).

Another critical implication is the necessity for continuous adversarial monitoring and periodic

retraining in real-world deployments. As the threat landscape evolves, IDS models must be updated not only with new labeled attack samples but also with adversarial examples generated through simulated offensive techniques (He et al., 2022). This cyclical process reflects modern security paradigms such as threat hunting and proactive defense, where systems adapt to attacker innovations rather than passively relying on historical datasets. Additionally, the study suggests that future IDS design must integrate explain ability and interpretability tools, enabling analysts to better understand why certain adversarial inputs succeed or fail. Such insights are essential for developing targeted defenses and strengthening the reliability of deep learning models in mission-critical cyber security infrastructures (Roshan et al., 2023).

Overall, the discussion reinforces that deep learning IDS models cannot be deployed as standalone, static solutions. Their vulnerability to adversarial manipulation presents a severe risk that must be addressed through ongoing research, adaptive defenses, and hybrid protection strategies. This study contributes valuable insights by demonstrating the extent of adversarial impact, evaluating multiple robustification techniques, and proposing a defense framework that aligns with real-world security requirements. These findings emphasize the urgent need for combining model-level robustness with system-level resilience to create intrusion detection systems capable of withstanding future AI-powered cyber attacks (Alotaibi, 2022).

8. Conclusion

This research concludes that while deep learning–based intrusion detection systems represent a significant advancement in cyber security, they remain critically vulnerable to adversarial manipulation, posing serious risks to the reliability of automated threat detection. The experiments demonstrate that even minor perturbations crafted using techniques such as FGSM, PGD, DeepFool, and especially GAN-based attacks can substantially degrade detection accuracy, mislead classifiers, and allow malicious traffic to bypass defenses undetected (Zhou & Pezaros, 2019). The fact that GAN-powered attacks can replicate the statistical properties of legitimate traffic makes them an especially powerful tool for cybercriminals operating in increasingly AI-driven threat environments. At the same time, the study also reveals that targeted defense mechanisms, particularly adversarial training and hybrid defense frameworks that combine sanitization, anomaly filtering, and multi-model verification, can meaningfully restore system robustness (Apruzzese et al., 2020). However, no single defense mechanism offers complete protection, underscoring that IDS robustness is not a static property but an ongoing requirement that demands continuous adaptation. Ultimately, the findings highlight a critical shift in cyber security strategy: modern IDS systems must evolve from passive detectors into adaptive, self-defending models capable of learning from emerging adversarial patterns. This study not only exposes the vulnerabilities of current deep learning IDS solutions but also contributes practical frameworks to enhance resilience, offering a foundation upon which future intrusion detection models can be built to withstand the growing sophistication of AI-powered cyber attacks (Alhajjar et al., 2021).

Implications for Cyber security Practice

The findings of this study carry significant implications for cyber security practice, underscoring the urgent need for organizations to rethink how intrusion detection systems are deployed, monitored, and updated in modern threat landscapes. The demonstrated vulnerability of deep learning models to adversarial perturbations reveals that high accuracy in controlled environments cannot be equated with real-world robustness, prompting security teams to integrate adversarial resilience testing into standard IDS evaluation cycles. Cyber security practitioners must adopt multi-layered defense strategies—combining adversarial training, traffic sanitization, anomaly filtering, and behavioral analysis—to counter the growing sophistication of AI-enabled attacks. Regular retraining of IDS models with updated

adversarial samples is essential to maintain relevance as attackers continually evolve their techniques. Furthermore, security operations centers (SOCs) must shift toward proactive defense by simulating adversarial scenarios, much like penetration testing, to uncover hidden weaknesses before attackers exploit them. The integration of explainable AI tools can help analysts interpret model decisions, diagnose vulnerabilities, and build trust in automated detection systems. Additionally, reliance on a single detection model should be avoided; ensemble-based IDS architectures and cross-model verification can significantly improve resilience. As regulatory bodies increasingly emphasize cyber security accountability, organizations must also ensure adherence to emerging standards for AI-driven security systems (Sharma & Chen, 2023).

Future Work

Future research should expand adversarial robustness testing by incorporating more diverse datasets, including IoT, cloud-native, and industrial network traffic, to assess model performance in broader environments. Future studies may also explore more advanced generative attack models such as diffusion-based adversarial frameworks, which are expected to surpass current GAN capabilities. Integrating explainable AI (XAI) techniques could help identify the internal vulnerabilities within deep IDS models and guide stronger defense development. Additionally, real-time adaptive defense mechanisms that learn continuously from live network traffic should be investigated to counter evolving attacker strategies. Hybrid architectures combining classical rule-based systems with deep learning may offer improved resilience. Finally, collaboration between academia and cyber security industry partners is essential to validate adversarial robustness techniques in large-scale, real-world deployments.

Policy Recommendations

1. Establish National Standards for Adversarial Robustness Testing

Governments and regulatory bodies should mandate standardized adversarial testing protocols for all AI-based intrusion detection and cyber security systems. Current certification guidelines focus primarily on accuracy and compliance, ignoring adversarial manipulability. National cyber authorities must define robustness benchmarks, perturbation thresholds, and model evaluation criteria to ensure minimum resilience levels across critical infrastructure sectors. Periodic compliance audits should be required to assess whether deployed systems can withstand emerging adversarial threats. Such standards will create a unified baseline for cyber security readiness nationwide.

2. Incentivize Adoption of Hybrid and Multi-Layered Defense Architectures

Policy frameworks should promote the implementation of layered defense architectures that combine deep learning, behavioral analysis, rule-based detection, and adversarial monitoring. Governments may offer tax incentives, grants, or cyber security readiness credits to organizations that adopt hybrid IDS frameworks proven to resist adversarial attacks. Regulatory authorities should publish recommended defense templates for high-risk sectors such as finance, healthcare, and energy. Mandatory guidelines should discourage reliance on single-model detection systems, as they demonstrate higher vulnerability to adaptive AI-powered attacks. Encouraging multi-tier protection will strengthen national cyber resilience.

3. Mandate Continuous Model Updates and Adversarial Retraining Cycles

Policy intervention is needed to require organizations to periodically update and retrain AI-based IDS models with fresh datasets and adversarial examples. Static or outdated models significantly increase the risk of catastrophic breaches. Cyber security regulators should enforce annual or bi-annual retraining cycles, especially for mission-critical institutions. Additionally, organizations should be required to maintain internal "adversarial traffic repositories" that store simulated attacks for continuous learning. Such mandates ensure that AI-driven

security systems evolve in parallel with attacker capabilities and remain effective over time.

4. Strengthen SOC Capabilities Through Adversarial Awareness Training

National cyber security policymaking should include mandatory adversarial machine learning (AML) training for SOC analysts, cyber security professionals, and network administrators. This training must focus on recognizing adversarial behaviors, identifying model weakening indicators, and responding to AI-generated intrusions. Policymakers should allocate funding for certified AML courses, practical labs, and public–private capacity-building programs. Traditional cyber security training frameworks do not adequately cover adversarial manipulation risks, leaving critical skill gaps in the workforce. Enhancing SOC expertise is vital for real-time, informed decision-making during adversarial attack campaigns.

5. Promote Public–Private Collaboration for AI-Driven Threat Intelligence Sharing

Governments should establish formal collaboration platforms that facilitate the sharing of adversarial attack patterns, GAN-generated threat signatures, and model vulnerability insights among private companies, academia, and national cyber authorities. AI-powered threats evolve rapidly, and isolated defense efforts are insufficient. Policies should require critical infrastructure operators to contribute anonymised adversarial incident data to national threat intelligence hubs. Additionally, governments must incentivize joint research initiatives that advance adversarial defense technologies. A collaborative, ecosystem-wide approach will dramatically enhance early detection and collective resilience against emerging AI-driven cyber threats.

REFERENCES

Ajdani, M., & colleagues. (2024). *Deep learning-based intrusion detection systems: A novel approach using generative adversarial networks (GANs).* Computers & Security.

Ajdani, M., & others. (2025). Deep learning-based intrusion detection systems: A GAN-driven adversarial training framework. *Journal of Information Security and Applications*.

Alhajjar, E., et al. (2021). Evolutionary computation-based adversarial attacks on network intrusion detection systems. *Applied Soft Computing*.

Alhajjar, E., Maxwell, P., & Bastian, N. D. (2020). Adversarial machine learning in network intrusion detection systems. *arXiv preprint arXiv:2004.11898*.

Alkadi, S., Moustafa, N., Turnbull, B., & Choo, K.-K. R. (2024). RobEns: Robust ensemble adversarial machine learning framework for intrusion detection systems. *IEEE Access*.

Alotaibi, A. (2022). Evasion attacks and defenses in machine learning-based intrusion detection systems. *Future Internet*.

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

Alotaibi, A., & Rassam, M. A. (2024). Defense-aware adversarial machine learning for secure intrusion detection. *Future Internet*.

Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2021). Modeling realistic adversarial attacks against network intrusion detection systems. *Computer Networks*.

Apruzzese, G., et al. (2020). On the effectiveness of machine and deep learning for cyber security. *IEEE Transactions on Dependable and Secure Computing*.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331.

Carlini, N., & Wagner, D. (2019). On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

Das, A., & colleagues. (2021). Network intrusion detection system based on generative adversarial networks for DDoS attack generation and evaluation. *International Journal of Computer Networks & Communications*.

Debicha, I., Debatty, T., Dricot, J.-M., & Mees, W. (2021). Adversarial training for deep learning-based intrusion detection systems. *arXiv preprint arXiv:2104.09852*.

Ding, X., Zhang, Y., & others. (2020). Toward invisible adversarial examples against DNN-based privacy leakage for Internet of Things. *IEEE Internet of Things Journal*.

Farhan, M., et al. (2025). Network-based intrusion detection using deep learning with feature selection on UNSW-NB15. *Scientific Reports*.

Fu, X., Yan, Z., & Zhang, H. (2021). The robust deep learning–based schemes for intrusion detection in Internet of Things environments. *Annals of Telecommunications, 76*, 273–285.

Hartono, D. D. (2025). Efficient GAN-based adversarial example generation against network intrusion detection systems. (Master's thesis). *California Polytechnic State University*.

He, K., & colleagues. (2022). Adversarial learning in network intrusion detection: Challenges and opportunities. *IEEE Access*.

He, K., Zhang, X., & Sun, L. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*.

Hozouri, A., Mirzaei, A., & Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. *Discover Artificial Intelligence*.

Lin, Z., Shi, Y., & Xue, Z. (2018). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. *arXiv preprint arXiv:1809.02077*.

Lin, Z., Shi, Y., & Xue, Z. (2022). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In *Lecture Notes in Computer Science* (pp. 97–111). Springer.

Nguyen, D. T., et al. (2023). The robust scheme for intrusion detection system in adversarial IoT environments. *Internet of Things*.

Paltun, B. G., et al. (2025). Robust intrusion detection system with explainable artificial intelligence against adversarial attacks. *arXiv preprint arXiv:2503.05303*.

Rashid, M. M., Kamruzzaman, J., Imam, T., & Gondal, I. (2022). Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. *Computers & Security*.

Roshan, M. K., et al. (2023). Two-phase defense against optimization-based adversarial attacks in network intrusion detection. *Knowledge-Based Systems*.

Roshan, M. K., et al. (2024). Boosting robustness of network intrusion detection systems against C&W adversarial attacks. *Expert Systems with Applications*.

Sharma, S., & Chen, Z. (2023). Black-box decision-based adversarial attacks against machine-learning network intrusion detection systems. *Electronics*.

Sharma, S., & Chen, Z. (2024). A systematic study of adversarial attacks against network intrusion detection systems. *Electronics, 13*(24), 5030.

Wang, G., et al. (2025). A method for improving the robustness of intrusion detection systems against adversarial attacks. *Electronics, 14*(11), 2171.

Wang, J., et al. (2021). DEF-IDS: An ensemble defense mechanism against adversarial attacks for deep learning-based network intrusion detection. In *2021 International Conference on Computer Communications and Networks (ICCCN)*.

Wang, Z., et al. (2018). Deep learning based intrusion detection with adversaries. *IEEE Access*, 6, 38367–38384.

Xu, D., et al. (2025). DEMGAN: A machine learning-based intrusion detection evasion method using adversarial traffic generation. *Computers & Security*.

Yang, K., et al. (2018). Adversarial examples against the deep learning based network intrusion detection systems in wireless networks. In *MILCOM 2018 – IEEE Military Communications Conference*.

Zhang, X., & others. (2021). IDS-GAN based adversarial traffic generation for evaluating intrusion detection systems. *Security and Communication Networks*.

Zhang, X., et al. (2020). Network intrusion detection using generative adversarial networks. *International Journal of Network Security*.

Zhang, Y., et al. (2019). Adversarial examples for network intrusion detection systems. *Proceedings of IEEE Conference on Communications and Network Security*.

Zhou, Y., & Pezaros, D. (2019). Evaluation of adversarial evasion attacks on network intrusion detection systems. *IEEE Transactions on Network and Service Management*.