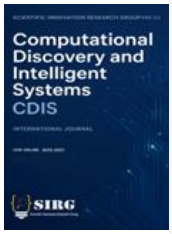

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/
Journal Homepage
https://pub.scientificirg.com/index.php/CDIS/index

A Cross-Dataset Empirical Evaluation of Adversarial Evasion Attacks and Defenses in Machine Learning-Based Intrusion Detection Systems

Salsabil Tarek[a,1], Muthmainnah Muthmainnah[b], Ahmed J. Obaid[c]

[a] Computer Science Department, Faculty of Computer Science, Nahda University Beni-Suef, Beni Suef, Egypt. Email: salsabil.tarek@nub.edu.eg.com
[b] Department of Informatics Engineering, Faculty of Computer Science, Al Asyariah Mandar University, Indonesia. Email: muthmainnah@unasman.ac.id
[c] Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq. Email: ahmedj.aljanabi@uokufa.edu.iq

ABSTRACT

The study aims to assess the adversarial robustness of two intrusion detection systems (IDS), namely XGBoost and Multilayer Perceptron (MLP), using three datasets: NSL-KDD, CIC-IDS2017, and UNSW-NB15. The attacks were carried out using two methods: FGSM and PGD, and transfer-based attacks for non-differentiable models. The results indicate that adversarial attacks significantly affect intrusion detection systems' performance, with attacks reducing the MLP IDS detection rate to 33.19% on the CIC-IDS2017 dataset, to 4.43% with FGSM attacks, and to 0.00% with transfer-based PGD attacks, on the XGBoost IDS. The study also indicates that adversarial training improves the robustness of intrusion detection systems' performance, with the MLP IDS maintaining a 96%+ detection rate even after undergoing adversarial attacks on the CIC-IDS2017 dataset and 68% on the UNSW-NB15 dataset.

PAPER INFORMATION

HISTORY
Received: 5 January 2026
Revised: 4 March 2026
Accepted: 19 April 2026
Online: 25 April 2026

MSC
68T07; 68R10; 94A60; 68M15

KEYWORDS
Intrusion Detection Systems (IDS);
Adversarial Machine Learning;
Network Security;
FGSM;
Adversarial Training.

[1]Corresponding author at Computer Science Dept., Faculty of Computer Science, Nahda University Beni-Suef, Beni Suef, Egypt, E-mail: salsabil.tarek@nub.edu.eg.com.

1 INTRODUCTION

Cyberattacks, both in terms of their volume and complexity, have increased manifold owing to the rapid advancement of computer networks, cloud computing, and Internet of Things technology. The detection capabilities of traditional intrusion detection systems, which rely on signature-based detection, are found to be very low for handling unexpected attacks. Machine learning-based intrusion detection systems have become increasingly popular for handling such attacks, as they can identify unknown attacks with high accuracy [1].

DOI: https://doi.org/10.66279/3k5mqs50

To provide proactive security, it is necessary for the machine learning algorithm to classify both malicious and benign activities on its own, based on the information extracted from network flows. Invasion detection, scanning, and denial of service attacks can cause data loss, which can be very expensive for organizations. Therefore, it is necessary for intrusion detection systems to identify intrusions even in hostile environments. A severe security threat can arise if the intrusion detection system is unable to identify intrusions in real time. Therefore, it is necessary for intrusion detection systems to be robust.

The use of machine learning algorithms is of critical importance in modern IDS systems. These algorithms include neural network algorithms, ensemble learning algorithms, and statistical-based classifiers. These algorithms can identify suspicious behaviors based on the volume of network traffic. Deep learning algorithms can improve the accuracy of detecting non-linear relationships that exist among the variables of network traffic. These algorithms have the advantage of achieving higher accuracy compared to other algorithms. The use of these algorithms in the IDS systems can improve the adaptation of the dynamic patterns of attacks.[37]

However, the use of machine learning algorithms in IDS systems has faced challenges that have impacted the adoption of algorithms. These challenges include the accuracy of the algorithms. The accuracy of the algorithms can be impacted by the presence of concept drift, class imbalance, and data quality. In addition, the algorithms can perform better on the training data but can fail when deployed on new or different data. Other challenges include the use of computational power, interpretability, and the integration of the algorithms with the existing systems.

The most critical challenges include adversarial attacks. Research suggests that ML models are prone to adversarial attacks, where small changes in the input data are made with the objective of misleading the ML model's output. In the context of IDS systems, adversarial evasion attacks have been identified as changes made to malicious network data that retain the attack potency but are misclassified as non-malicious data by the ML model [4]. Other attacks include gradient-based attacks, such as Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM), which have been identified as critical challenges that reduce the IDS's performance. These attacks have raised critical questions regarding the ML-based IDS systems' reliability.

Several defense techniques have been proposed for addressing adversarial attacks. These include adversarial training, where the ML model is trained with adversarial data. Adversarial training has been identified as an efficient technique for improving the ML model's robustness. By training the ML model with adversarial data, the model's decision boundaries are strengthened, ensuring that the model is not vulnerable to evasion attacks. However, the ML model's performance on clean data is not sufficient for evaluating the IDS. Other attack scenarios, including white-box and black-box attacks, need to be considered.

In addition, since IDS often handles sensitive network traffic, there is a need to prioritize data privacy and ensure that the data is processed in a secure manner. This is important in meeting the requirements of cybersecurity standards and protecting sensitive data, hence providing trust and confidence in security monitoring systems [6].

Soon, research in intrusion detection systems is likely to be centered on creating systems that can learn from attacks and adapt to them, hence providing near-optimal performance and incorporating technologies such as edge computing and artificial intelligence [7].

Despite the existing knowledge in intrusion detection systems, there are still significant gaps in the research that need to be filled. Most of the existing research on intrusion detection systems has been conducted by evaluating metrics on benign network traffic without considering manipulated network traffic by attackers. The evaluation of the model's adversarial robustness has been conducted on a limited type of dataset, and there is a lack of evaluation of defense strategies, such as adversarial training, on multiple models and benchmark datasets with security-related metrics.

The research questions that this study aims to answer are:

- How sensitive are machine learning-based IDS to evasion attacks using different datasets and complexities of network traffic?
- How effectively can adversarial training improve IDS robustness while maintaining clean data performance?
- Are there any variations in robustness to adversarial attacks and transfer learning between different architectures

The main contributions of this paper are summarized as follows:

- **Comprehensive evaluation of adversarial vulnerability:**
  The research examines the robustness of machine learning-based intrusion detection systems in response to adversarial evasion attacks.
- **Cross-dataset evaluation:**
  The robustness of IDS models was evaluated on various benchmark datasets, including NSL-KDD, CIC-IDS2017, and UNSW-NB15, to assess the universality of adversarial vulnerability.
- **Comparative study of various architectures:**
  The research evaluated and compared the robustness of two prominent models, XGBoost and Multilayer Perceptron (MLP), in response to adversarial evasion attacks.
- **Evaluation of various attack scenarios:**

The robustness of IDS models was evaluated in response to various attack scenarios, including white-box and black-box attacks, using gradient-based and transfer-based attacks, respectively.

- **Evaluation of adversarial training:**
  The effectiveness of adversarial training in enhancing the robustness of IDS models was evaluated in this study.
- **Evaluation metric for robust security:**
- The research emphasized the importance of malicious detection performance in response to adversarial attacks.

2 Literature Review

When it comes to recognizing malicious activity on a computer network, intrusion detection systems play a critical role. Recently, to improve intrusion detection systems' performance and detect more sophisticated attacks, machine learning and deep learning methods have been applied to intrusion detection systems [8], [9]. However, it has been demonstrated that intrusion detection systems relying on machine learning methods are vulnerable to adversarial attacks, wherein an attacker uses machine learning methods to evade intrusion detection systems or compromise their training process [10]. Recently, there has been a surge in studies on adversarial attacks on intrusion detection systems and their countermeasures since 2020. This literature review attempts to gather published articles on adversarial attacks on intrusion detection systems relying on machine learning methods published between 2020 and 2025 and compare these studies on various aspects.

Adversarial attacks on IDS can be categorized according to knowledge and timing. Based on timing:[37]

- **Evasion attacks (testing-time attacks):** These attacks occur during deployment. Attackers insert malicious data to trick the IDS into classifying it as legitimate traffic. This is the most common type of attack and assumes that the attacker can manipulate traffic during testing but not the training phase [11]. Typically, attackers introduce small perturbations to the features of network traffic to confuse the IDS.
- **Poisoning attacks (training-time attacks):** Here, attackers inject malicious samples into the training dataset or interfere with the training process to compromise the final model. By using deliberately crafted samples labeled as benign, the attacker can teach the IDS an incorrect decision boundary [11]. Poisoning attacks are particularly concerning in collaborative or online learning IDS, although they have received less attention than evasion attacks. For instance, Alshahrani et al. [11] simulated poisoning attacks and found that adding adversarial samples during training significantly disrupted the IDS's learning process, with logistic regression IDS being more vulnerable than decision tree IDS.

It is important to note that the goal of all adversarial attacks is the same: to disrupt IDS operations to achieve the attacker's objectives. Attacks can also be categorized based on the attacker's knowledge of the IDS model [12]. In white-box attacks, the attacker knows the model's architecture and parameters, enabling gradient-based attacks. In black-box attacks, the attacker has limited or no knowledge of the model and may use queries or transfer attacks from a surrogate model. Furthermore, attacks may be untargeted, simply causing misclassification into any incorrect class, or targeted, forcing misclassification into a specific class. In IDS evasion, untargeted attacks are often sufficient to bypass alarms [13].

Previous research has demonstrated the effectiveness of evasion attacks in compromising even the most accurate IDS models. Ayub et al. [8] utilized the Jacobian-based Saliency Map Attack, which was previously proposed in computer vision, to conduct evasion attacks on an MLP-based neural network IDS, which was trained on CICIDS2017 and TRAbID, an extension of the NSL-KDD dataset. The IDS model was able to attain a high accuracy of over 99%, but after conducting JSMA attacks, its accuracy was reduced to 22-30%. This highlights how easily evasion attacks can be performed on neural network-based IDS models. In this context, other attacks, such as FGSM, BIM, and PGD, have been proposed in computer vision and have been utilized to conduct evasion attacks on IDS models. Zhang et al. [14] utilized the FGSM and iterative gradient attacks on a DNN-based IDS, which was trained on the NSL-KDD dataset, and demonstrated how even slight perturbations in the attack can drastically reduce the accuracy of the IDS model. Piplai et al. [10] demonstrated how the FGSM attack can even compromise IDS models that have been trained using adversarial training.

Generative techniques such as Generative Adversarial Networks (GANs) have also been used to create adversarial network traffic. Chauhan & Shahid [16] demonstrated a polymorphic DDoS attack using GANs on CIC-IDS2017, significantly reducing IDS detection rates. Zhao et al. [17] employed AttackGAN to bypass black-box IDS, achieving over 87% attack success across models, including SVM, Decision Tree, Random Forest, Naïve Bayes, and DNN [18]. Alhajjar et al. [19] used GAN, genetic algorithms (GA), and particle swarm optimization (PSO) to generate adversarial perturbations, evaluating multiple classifiers on NSL-KDD and UNSW-NB15. They found that MLP classifiers were more resilient (maintaining 83.3% accuracy under attack) than other models, indicating that model choice impacts adversarial robustness. Sharma & Chen [12] evaluated nine common IDS models (including logistic regression, MLP,

random forest, k-NN, and SVM) and found that instance-based learners like k-NN and label spreading were generally more resistant to attacks, while logistic regression and MLP were more vulnerable.

Adversarial attacks have also been studied in specialized environments such as ICS and IoT networks. Anthi et al. [20] applied JSMA attacks on two ML-based IDS (Random Forest and J48) using an ICS dataset, showing significant performance degradation and highlighting the need for generalized defenses. Overall, evasion attacks, including white-box and black-box methods, gradient-based attacks, GAN-generated samples, and heuristic attacks, can reduce ML/DL-based IDS accuracy to near zero in some cases [21].

Poisoning attacks, which target the training phase, have received less attention but remain a growing concern. Alshahrani et al. [11] studied both poisoning and evasion attacks on CICIDS2017. Their results showed that poisoning affected logistic regression more than decision tree models, highlighting the influence of model architecture on vulnerability. Successful poisoning requires access to the data stream or model updates, which may occur in collaborative IDS or if an insider compromises the training process.

Various IDS defense strategies have been proposed. Adversarial training of IDS models on a mix of clean and adversarial samples has shown significant effectiveness. Heydari & Nyarko [9] proposed ADV_NN, a curriculum-based adversarially trained neural network on UNSW-NB15. Their approach maintained over 80% detection accuracy under strong PGD and FGSM attacks, compared to below 50% for standard transformer-based IDS. ADV_NN achieved ~85% accuracy even against black-box transfer attacks, demonstrating the potential of adversarial training despite a slight trade-off in clean-data performance.

Other defense mechanisms include ensemble strategies (combining adversarial training, Gaussian noise, and label smoothing) and auxiliary detection systems that pre-process inputs to identify adversarial samples [13, 23–25]. Techniques such as dropout-based detection exploit abnormal network behavior under random neuron removal to detect adversarial inputs [15]. Def-IDS [26], combining GANs and adversarially retrained classifiers, improved robustness against multiple attacks (FGSM, BIM, DeepFool, JSMA) compared to single-model baselines.

It is important to note that defenses may be circumvented by adaptive attackers; an effective strategy against one attack may fail against a slightly modified or aware adversary. Only a few IDS solutions currently account for this, and FGSM-based adaptive attacks have demonstrated that even adversarially trained models can be defeated [10]. The IDS community may adopt advanced strategies, including game-theoretic approaches, to address this ongoing arms race.

Table 1. Summarizes key studies on adversarial attacks against intrusion detection systems, highlighting their objectives, datasets, models, and attack methods to position the current work within existing research.

Table 1. A comparative overview of important research on adversarial attacks against IDS

Reference	Task	Dataset(s)	Model(s)	Attack Method(s)
Ayub et al. (2020) [8]	Use adversarial scenarios to demonstrate how to evade an MLP-based IDS (white-box).	CICIDS2017; TRAbID (NSL-KDD variant)	Multilayer Perceptron (ANN)	JSMA (gradient-based saliency attack)
Piplai et al. (2020) [10]	Evaluate the impact of FGSM and feature significance while avoiding a GAN-based IDS classifier.	BigData Cup 2019 (Suspicious Network Events)	GAN-based classifier (neural network)	FGSM (Fast Gradient Sign Method)
Chauhan & Heydari (2020) [16]	Create polymorphic DDoS attack traffic using a black box generating technique to get around IDS.	CICIDS2017	Multiple: DT, RF, LR, NB (classical IDS models)	GAN-generated adversarial DDoS samples
Zhang et al. (2020) [14]	Analyze how resilient DNN is to gradient-based evasion attempts.	NSL-KDD	Deep Neural Network (fully connected)	FGSM and basic iterative (white box)
Alhajjar et al. (2021) [19]	Determine vulnerabilities by analyzing different attack generation strategies on various IDS models.	NSL-KDD; UNSW-NB15	Multiple: MLP, SVM, NB, RF, etc.	PSO, GA, and GAN (heuristic and generative attacks)
Zhao et al. (2021) [17]	Apply a generative adversarial method to attack black-box intrusion detection systems (AttackGAN).	NSL-KDD (features)	Multiple: SVM, DT, RF, NB, DNN	GAN (generating adversarial network flows)

Duy et al. (2021) [27]	DIGFuPAS: Create adversarial samples that are functionally maintained to avoid IDS in SDN networks (black box).	NSL-KDD; CICIDS2018	Extensive: SVM, NB, MLP, LR, DT, RF, KNN, CNN, RNN	GAN (with function-preserving constraints)
Anthi et al. (2021) [20]	Evaluate adversarial attacks on Industrial Control Systems' (ICS) ML-based IDS.	Power system network data (ICS dataset)	Random Forest; J48 (decision tree)	JSMA (and similar evasion attacks)
Lin et al. (2022) [21]	IDSGAN: To create covert attack traffic against different IDS models, use a GAN.	NSL-KDD	Multiple: LR, DT, RF, KNN, SVM, NB, MLP	GAN (generative evasion of mixed attacks)
Alshahrani et al. (2022) [11]	Evaluate the differences between poisoning and evasion attacks on IDS and assess the effects on various methods.	CICIDS2017	Decision Tree; Logistic Regression	GAN-based Evasion; also Poisoning (training injection)
Sharma & Chen (2024) [12]	systematic assessment of several adversarial attack methods on various NIDS models (black-box & white-box).	NSL-KDD	9 models (LR, MLP, SVM, RF, KNN, etc.)	PGD (gradients); ZOO (score-based); Boundary & HopSkipJump (decision-based)
Heydari & Nyarko (2025) [9]	improving effectiveness, suggest and assess an adversarial trained IDS model (ADV_NN).	UNSW-NB15	Neural Network (ADV_NN); also, RF and Transformer (for baseline comparison)	FGSM and PGD (white-box); Transfer black-box attacks

3 Proposed Methodology

3.1 System Overview

The purpose of this suggested IDS system is to evaluate the strength of the adversarial evasion attack method and the machine learning-based IDS system while maintaining a high level of accuracy in the IDS system's ability to determine the difference between attack and regular traffic. To differentiate between normal and attack traffic, the system processes the network flow characteristics using benchmark datasets. To provide a realistic system environment, adversarial attack techniques are included while the system's machine learning and deep learning algorithms study both regular and attacking traffic. The flow illustrated in **Figure 1** provides the basis for system design.

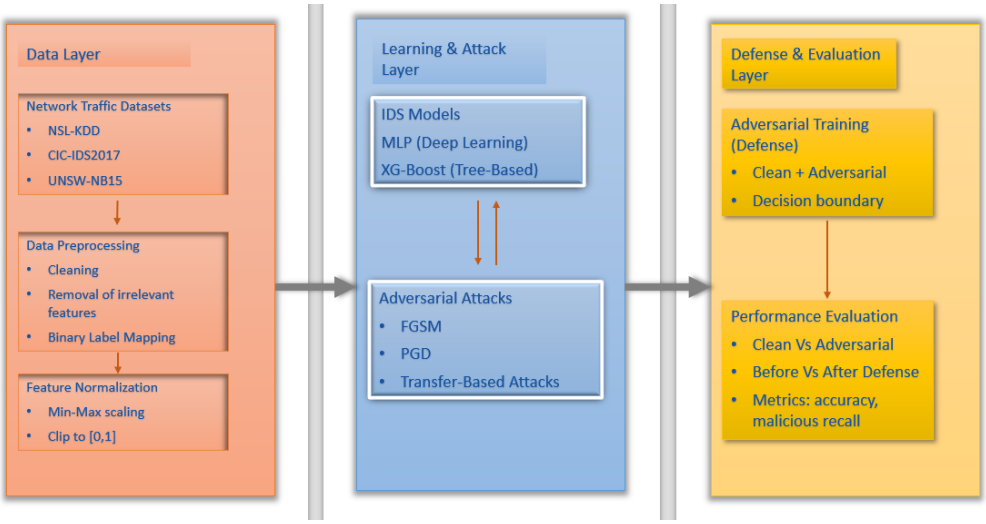

Figure 1. Proposed Architecture for Intrusion Detection System

3.2 Dataset Collection Description

Three public intrusion prevention datasets, NSL-KDD[28], CIC-IDS2017[29], and UNSW-NB15[30] are employed in this approach. These datasets offer various levels of attack variety and traffic pattern complexity. Each dataset's experiments include classification tasks in which the traffic flow records are divided into attack and normal records. In addition to the integration of various datasets, the experiments will not yield biased conclusions about the efficiency of prevention against attack sensitivity. Network traffic instances were categorized as either benign or malicious in all datasets, which were designed as binary classification tasks. For intrusion detection, the NSL-KDD dataset is an improved version of the widely used KDD'99 dataset. The performance metrics were also deceptively impacted by several problems in the previous dataset, including the existence of duplicate data and unequal difficulty levels. The NSL-KDD dataset contains records of network connections linked to 41 attributes, including content-based features, statistical characteristics based on traffic, and fundamental TCP/IP connection characteristics. Additionally, the records are suitably categorized as either attack or regular connections.

An improved intrusion detection data set called CIC-IDS2017 captures actual network traces in an experimental setup. This data collection includes both authentic traces and several recent attacks, such as denial-of-service, brute-force, and system-breach attempts. The CICFlowMeter generates around 80 numerical characteristics that describe packet activity, flow length, and statistical data to depict network traces. The CIC-IDS2017 data set will essentially serve as the standard for conducting adversarial evasion attacks, transfer attacks, and defenses for different learning methods throughout this research due to its realism and complexity.

To address the unrealistic character of earlier intrusion detection datasets, the UNSW-NB15 dataset was developed. The UNSW-NB15 dataset is extremely diversified because it blends artificially inserted attack behavior with real-world network traffic. 49 attributes that represent flow-level statistics, content, and protocol-level behavior are used to represent each network traffic record. To preserve its independence, the dataset's pre-splits training and test sets are left unaltered in this work. Analyzing adversarial susceptibility and protection in increasingly sophisticated network traffic will be made easier with the help of the UNSW-NB15 dataset.[36]

Three benchmark intrusion detection datasets, each with varying degrees of complexity and realism for thorough assessments, were used for experiments. In this work, several of these datasets serve a variety of purposes, from accurate generalization to initial validation, as presented in **Table 2**.

Table 2. Summary of datasets

Dataset	Traffic Type	Number of Features	Task	Function
NSL-KDD	Simulated	41	Binary IDS	Validation
CIC-IDS2017	Realistic Flows	80	Binary IDS	Benchmark
UNSW-NB15	Modern Mixed Traffic	49	Binary IDS	Generalization Analysis

3.3 Data Preprocessing and Normalization

To guarantee the consistency and dependability of model training, data preparation is an essential stage. Extreme values, duplicate properties, and missing values are common in raw network traffic features, which could influence the learning process. In this work, missing and infinite values are substituted with correct numerical representations, and non-informative characteristics and identifiers are eliminated. To convert all features into a single range [0,1], feature scaling is carried out via min–max normalization. This is crucial for gradient-based adversarial attacks. The following is an expression for the normalization process [31] shown in **Equation 1**:

$$\dot{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where x' refers to the normalized value and x denotes the original feature value and it rescales a feature value $x$to a range between **0 and 1** based on the minimum and maximum values in the dataset.

3.4 Feature Representation and Input Preparation

Every flow in the networks is represented by a feature vector that includes several aspects of the flow, such as size, duration, and protocol specifics. The learning model, which aids in differentiating between benign and malicious flows, uses the feature vector as input. Normalization, which aids the model in identifying patterns from the data distribution, is the sole human intervention involved in the development of features.

3.5 Models Used for Intrusion Detection

The suggested system's detection is mostly based on supervised machine learning models. The suggested method makes use of a Multilayer Perceptron (MLP) neural network due to its capacity to represent complicated non-linear correlations between features. Furthermore, because the XGBoost method can assess the resilience of models based on several machine learning paradigms, it is chosen for the suggested system. Each model's result will show the likelihood that the specified traffic flow is malicious.

3.6 Adversarial Attack Generation

Adversarial evasion attacks are developed utilizing gradient-based methods to assess the strength of the model. Specifically, as stated in **Equation 2**, the Fast Gradient Sign Method (FGSM) [32] modifies input samples by a little perturbation in the direction of the loss gradient:

$$adv\_x = x + \epsilon * sign(\nabla_x J(\theta, x, y)) \qquad (2)$$

It generates an adversarial example by adding a small, carefully calculated perturbation to the original input x.

- ϵ: controls the strength of the perturbation
- $\nabla_x J(\theta, x, y)$: gradient of the loss with respect to the input
- sign: keeps only the direction of the gradient

where x is the initial input, ϵ is the perturbation amount, and $J(\cdot)$ is the loss function.

To create better adversarial samples, an iterative version of FGSM called Projected Gradient Descent (PGD) [33] is also employed. Because XGBoost employs adversarial samples generated by a surrogate neural network, transfer-based or black-box attacks are employed for non-differentiable models. Adversarial perturbations, which reflect normal attacker behavior, only affect malicious samples.

Adversarial training is incorporated into the neural network models' learning process to mitigate adversarial risk. The clean set of training samples is mixed with adversarial data acquired by FGSM attacks during the learning process. As a result, the model can create robust decision surfaces. To gauge the increase in robustness, the final robust model is then put to the test using PGD and FGSM attacks.

The proposed architecture is predicated on a threat model where attackers seek to avoid detection by minimally altering harmful communications while maintaining attack capabilities. Realistic adversarial capabilities are expected to be reflected in both white-box and black-box attack scenarios. The technique offers a thorough evaluation of system security that delves beyond conventional accuracy-based evaluation by assessing IDS performance in various situations.

3.7 System Evaluation and Performance Metrics

Accuracy [34] as calculated in **Equation 3**, malicious recall [35] as formulated in **Equation 4**, and the extent of resilience loss under attack by FGSM/PGD techniques are some of the metrics used to assess the efficacy of the suggested framework for IDS systems. Both with and without adversarial training, the system's performance on clean examples is contrasted

with that of FGSM/PGD attack cases. In this experiment, the value of malware recall is essential since it indicates how well the system works to reduce attacks.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified instances (both positive and negative) out of all predictions.

$$Malicious\ recall = \frac{TP}{TP+FN} \quad (4)$$

Malicious Recall measures the model's ability to correctly detect malicious instances, indicating how many actual attacks are successfully identified.

4 RESULTS AND DISCUSSION

The experimental findings of the suggested framework on clean data, evasion attacks, and following the use of defense measures are shown in this section.

The baseline performance of the models tested in a clean environment free from adversarial attacks is shown in **Table 3.** Every model performs almost perfectly on the NSL-KDD and CIC-IDS2017 datasets. However, because the UNSW-NB15 dataset is more complicated, the models are challenged with a more realistic adversarial setting, which results in less performance, as shown in **Figure 2**.

Table 3. Performance on Clean Data

Dataset	Model	Accuracy (%)	Malicious Recall (%)
NSL-KDD	ML-Based IDS	99	99
CIC-IDS2017	MLP	99.93	99.95
CIC-IDS2017	XGBoost	99.99	99.99
UNSW-NB15	MLP	92.49	93.57

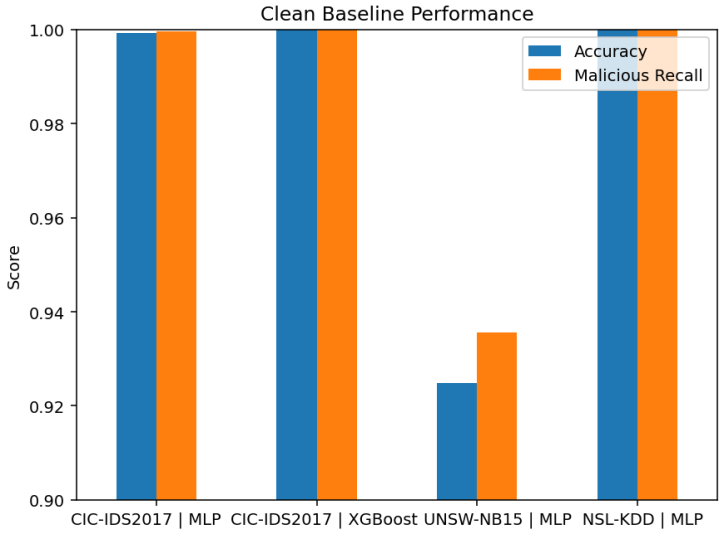

Figure 2. Clean Baseline performance

Table 4. shows the effect of the adversarial evasion attack on CIC-IDS2017 when no defense plan is used. The adversarial attack has significantly impacted both the decision tree model and the neural network model's performance, as illustrated in **Figure 3.**

Table 4. Adversarial attack on CIC-IDS2017 (Before Defense)

Model	Attack	ε	Accuracy (%)	Malicious Recall (%)
MLP	FGSM	0.01	62.6	33.19
MLP	PGD	0.01	77.24	59.94
XGBoost	Transfer-FGSM	0.01	71.25	49.31
XGBoost	Transfer-PGD	0.01	45.79	4.43

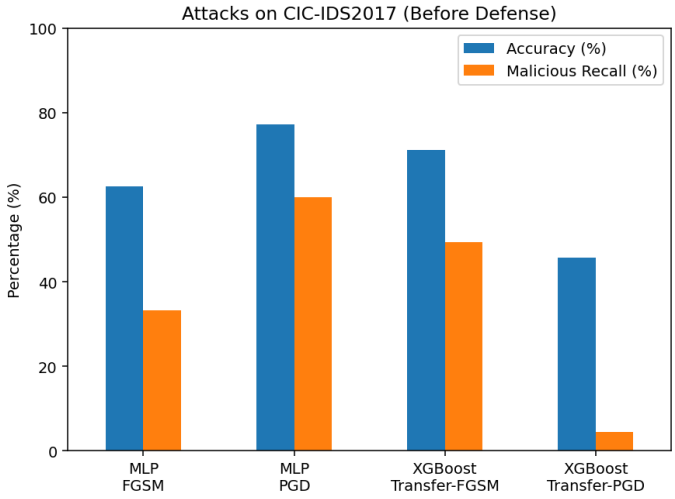

Figure 3. Attacks on CIC-IDS2017 (Before Defense)

The impact of adversarial training on CIC-IDS2017 is presented in **Table 5.** While there is little increase in XGBoost, the defended MLP model has proven to be resilient against FGSM and PGD attack techniques, as shown in **Figure 4.**

Table 5. Adversarial Training on CIC-IDS2017 Impact

Model	State	Attack	ε	Accuracy (%)	Malicious Recall (%)
MLP	After Defense	FGSM	0.01	97.75	97.21
MLP	After Defense	PGD	0.01	97.19	96.21
XGBoost	After Defense	Transfer-FGSM	0.01	71.25	49.31
XGBoost	After Defense	Transfer-PGD	0.01	55.27	21.14

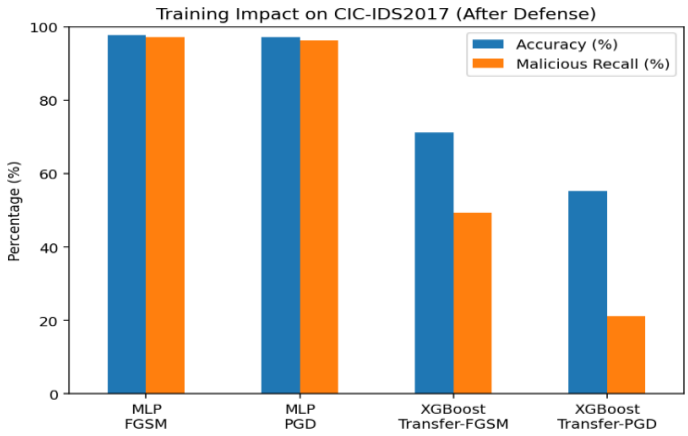

Figure 4. Training on CIC-IDS2017 effect

Table 6. lists the performance analysis for the UNSW-NB15 dataset before defense. The MLP model is highly sensitive to attacks, while also exhibiting exceptional clean precision, as shown in **Figure 5**.

Table 6. Adversarial attack on UNSW-NB15 (Before Defense)

Attack	ε	Accuracy (%)	Malicious Recall (%)	Recall Drop (%)
Clean	0.00	92.49	93.57	-
FGSM	0.02	60.40	46.42	47.15
PGD	0.02	59.61	45.26	48.30

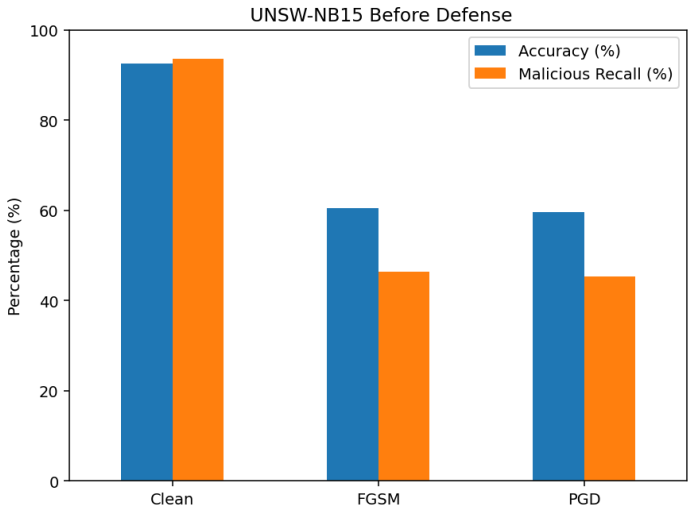

Figure 5. Attacks on UNSW-NB15 (Before Defense)

The MLP's outcome on the UNSW-NB15 following adversarial training is shown in **Table 7.** and **Figure 6.** Adversarial attacks have less impact, and the clean data suffers low to medium penalties.

Table 7. Adversarial Training on UNSW-NB15 impact

State	Attack	ε	Accuracy (%)	Malicious Recall (%)	Recall Drop (%)
Before	FGSM	0.02	60.40	46.24	47.15
Before	PGD	0.02	59.61	45.26	48.30
After	FGSM	0.02	78.26	70.43	18.81
After	PGD	0.02	79.63	68.48	20.77

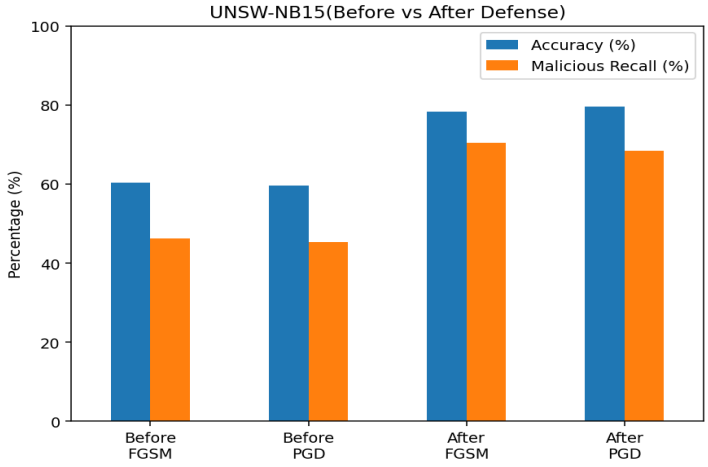

Figure 6. UNSW-NB15 (Before and After Defense)

The key strength indicators before and after attacks are highlighted in **Table 8.** below for easier identification of the strength patterns in the datasets, as illustrated in **Figure 7.**

Table 8. Cross-Dataset Summary

Dataset	Model	Clean Recall (%)	Worst Recall Under Attack (%)	Recall After Defense (%)
NSL-KDD	MLP / Classical	99	55	-
CIC-IDS2017	MLP	99.95	33.19	96.21
CIC-IDS2017	XGBoost	99.99	4.43	21.14
UNSW-NB15	MLP	93.57	45.26	68.48

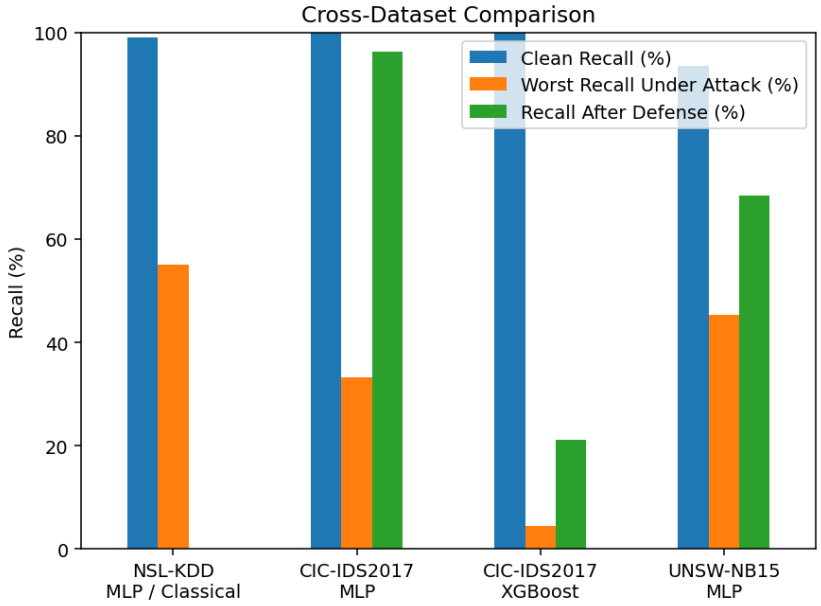

Figure 7. Cross-Dataset Recall Summary

The datasets exemplify that even though tree-based models like XGBoost show sensitivity in transfer-based adversarial environments, and even though adversarial training improves the robustness of neural network models, the sensitivity of XGBoost and the sensitivity and the robustness of neural network models shows no relationship to the networks resilience to adversarial evasion attacks and/or the networks accuracy to non-adversarial data sets.

The results of the study show that to the contrary, high accuracy detection and response of the model to the non-adversarial data sets does demonstrate an impenetrable defensive response to the evasive attacks and does instill confidence in the impenetrability of the deployed system, and as a result, the deployed system will not demonstrate effective defensive responses to evasive attacks, adversarial guess, or adversarial perturbation attacks. The machine learning models that were evaluated in the study failed to demonstrate defensive measures to adversarial perturbation attacks, and unfortunately, the lack of defensive measures to guess attacks, perturbation attacks, or adversarial attacks can be explained by a lack of defensive measures deployed in the ML model.

Finally, the most important conclusion that can be made is that regardless of the datasets, the threats and vulnerabilities always exist. The vulnerabilities and threats did not only happen with the NSL-KDD ridiculously high accuracy but also with the more advanced datasets like CIC-IDS2017 and UNSW-NB. The adversarial attack is more a weakness of the machine learning engine and not of the datasets that were used in training.

Further analysis of the models' architecture divulges the subtleties of the differences in the degree of robustness. Gradient-based attacks, particularly the FGSM and the PGD, cause dramatic decreases in the recall of the malicious class before any sort of defense has been implemented. This is particularly true for older models like the MLPs, which suffer extreme drops in defense recall. However, the addition of adversarial training has shown to increase the robustness of

gradient-based attacks without an increase in the recall of the true positive class. This defense mechanism recall stability is a positive secondary effect of the trading stability of the neural machine models' decision-making borders.

On the other hand, one of the most promising results in the empirical analysis has been achieved for the non-neural model based on decision trees, which is the XGBoost model. This model was particularly resilient to transfer-based black-box adversarial attacks for PGD, which is a notable achievement. Having demonstrated almost perfect accuracy on the clean data, the slight positive shift in performance, which was observed with the adversarial data augmentation, is a strong signal of non-differentiable approaches. This goes against the misconception that tree-based and ensemble models are more robust than neural networks, highlighting the importance of assessing the robustness of intrusion detection systems against black-box attacks.

Another important finding relates to the evaluation of the metrics employed. The malicious recall noted a steep decline, suggesting a sizable proportion of attacks remained undetected. This occurred despite the frontline detection accuracy being relatively high across several competing scenarios. This reiterates the argument that accuracy cannot be the sole measure of the effectiveness of an IDS. Practically, greater false negatives imply that more attacks are undetected, which is a more significant security problem than has previously been indicated.

This UNSW-NB15 dataset further illustrates the trade-off between robustness and accuracy on clean data. The adversarial training yielded a significant improvement in the robustness of attacks, demonstrating that this method is appropriate for security-sensitive scenarios, even though there is a slight drop in clean data accuracy and recall. Due to the small variations observed in the high-accuracy numbers, more work on adaptive or multi-step defense strategies is warranted.

In conclusion, these results present defense against adversarial attacks as a central objective in developing machine learning-based intrusion detection systems, as opposed to just an auxiliary consideration. Furthermore, results indicate that the development of adequate threat analysis and robust defensive strategies is imperative for both the neural and the non-neural IDS systems.

5 Conclusion

This study used the NSL-KDD, CIC-IDS2017, and UNSW-NB15 datasets to assess the efficiency of machine learning-based intrusion detection systems under evasion attacks. While the accuracy of both Neural Network-based systems and decision-tree-based systems significantly decreased under FGSM attacks and PGD attacks, respectively, the obtained findings clearly show that achieving high accuracy values for typical cases is insufficient for achieving evasion resistance. It was found that adversarial training was a successful strategy for protecting the models, resulting in a significant increase in malicious recall and a slight decrease in clean performance for the neural models. However, transfer-based attacks remained a threat to tree-based architectures such as XGBoost.

In conclusion, the findings highlight the integral significance of adversarial strength in intrusion detection system design and analysis. The challenge of guaranteeing the dependability of intrusion detection systems in a hostile environment exceeds the significance of traditional cleanliness-based accuracy.

Acknowledgments

As a result, the authors would especially like to express gratitude to the academic supervisors and faculty members for their wise counsel, helpful critiques, and support during this project. To make this research possible, the authors additionally acknowledge the publicly accessible benchmark datasets on NSL-KDD, CIC-IDS2017, and UNSW-NB15.

Funding

The authors state that no outside funding was received for this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Y. Al-Nashif, A. A. Kumar, S. Hariri, G. Qu, Y. Luo, and F. Szidarovsky, "Multi-level intrusion detection system (ML-IDS)," in Proc. 5th Int. Conf. Autonomic Computing (ICAC), 2008, pp. 131–140, doi: 10.1109/ICAC.2008.25.

[2] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance analysis of machine learning algorithms in intrusion detection system: A review," Procedia Comput. Sci., vol. 171, pp. 1251–1260, 2020, doi: 10.1016/j.procs.2020.04.133.
[3] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Trans. Emerg. Telecommun. Technol., vol. 32, no. 1, p. e4150, 2021, doi: 10.1002/ett.4150.
[4] O. H. Abdulganiyu, T. A. Tchakoucht, and Y. K. Saheed, "A systematic literature review for network intrusion detection system (IDS)," Int. J. Inf. Secur., vol. 22, no. 5, pp. 1125–1162, 2023, doi: 10.1007/s10207-023-00682-2.
[5] L. Yang and A. Shami, "IDS-ML: An open-source code for intrusion detection system development using machine learning," Software Impacts, vol. 14, p. 100446, 2022, doi: 10.1016/j.simpa.2022.100446.
[6] A. H. Azizan et al., "A machine learning approach for improving the performance of network intrusion detection systems," Ann. Emerg. Technol. Comput., vol. 5, no. 5, pp. 201–208, 2021, doi: 10.33166/AETiC.2021.05.025.
[7] G. Kocher and G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: Recent developments and challenges," Soft Comput., vol. 25, no. 15, pp. 9731–9763, 2021, doi: 10.1007/s00500-021-05893-0.
[8] Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS), Princeton, NJ, USA, Mar. 2020.
[9] V. Heydari and K. Nyarko, "Enhancing adversarial robustness in network intrusion detection," Electronics, vol. 14, no. 16, p. 3249, 2025, doi: 10.3390/electronics14163249.
[10] A. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando, "Functionality-preserving black-box optimization of adversarial malware," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 3469–3484, 2021, doi: 10.1109/TIFS.2021.3082330.
[11] E. Alshahrani, D. Alghazzawi, R. Alotaibi, and O. Rabie, "Adversarial attacks against supervised machine learning based network intrusion detection systems," PLoS One, vol. 17, no. 10, p. e0275971, 2022, doi: 10.1371/journal.pone.0275971.
[12] S. Sharma and Z. Chen, "A systematic study of adversarial attacks against network intrusion detection systems," Electronics, vol. 13, no. 24, p. 5030, 2024, doi: 10.3390/electronics13245030.
[13] Z. Awad, M. Zakaria, and R. Hassan, "An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems," Sci. Rep., vol. 15, no. 1, p. 14177, 2025, doi: 10.1038/s41598-025-94023-z.
[14] X. Zhang, X. Zheng, and D. D. Wu, "Attacking DNN-based intrusion detection models," IFAC-PapersOnLine, vol. 53, no. 2, pp. 415–419, 2020, doi: 10.1016/j.ifacol.2020.12.115.
[15] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey," Future Internet, vol. 15, no. 2, p. 62, 2023, doi: 10.3390/fi15020062.
[16] R. Chauhan and S. S. Heydari, "Polymorphic adversarial DDoS attack using GAN," in Proc. Int. Symp. Netw. Comput. Commun. (ISNCC), 2020, pp. 1–6, doi: 10.1109/ISNCC49221.2020.9297240.
[17] S. Zhao et al., "AttackGAN: Adversarial attack against intrusion detection system," Procedia Comput. Sci., vol. 187, pp. 128–135, 2021, doi: 10.1016/j.procs.2021.04.043.
[18] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," Appl. Sci., vol. 9, no. 20, p. 4396, 2019, doi: 10.3390/app9204396.
[19] E. Alhajjar, P. Maxwell, and N. Bastian, "Adversarial machine learning in intrusion detection systems," Expert Syst. Appl., vol. 186, p. 115715, 2021, doi: 10.1016/j.eswa.2021.115715.
[20] E. Anthi et al., "Adversarial attacks on machine learning cybersecurity defenses in industrial control systems," J. Inf. Secur. Appl., vol. 58, p. 102717, 2021, doi: 10.1016/j.jisa.2020.102717.
[21] Z. Lin and X. Shi, "IDSGAN: Generative adversarial networks for attack generation against intrusion detection," in Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining (PAKDD), 2022, pp. 79–91, doi: 10.1007/978-3-031-05933-9_61.
[22] A. Oprea, A. Singhal, and A. Vassilev, "Poisoning attacks against machine learning," Computer, vol. 55, no. 11, pp. 94–99, 2022, doi: 10.1109/MC.2022.3197662.
[23] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, and A. Madry, "On evaluating adversarial robustness," arXiv preprint arXiv:1902.06705, 2019. (Note: This methodological guideline is primarily available and cited as an arXiv preprint).
[24] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 1310–1320.
[25] Y. Peng et al., "Detecting adversarial examples for intrusion detection system with generative adversarial networks," in Proc. 11th IEEE Int. Conf. Softw. Eng. Serv. Sci. (ICSESS), 2020, pp. 31–34, doi: 10.1109/ICSESS52187.2020.9322676.
[26] J. Wang et al., "Def-IDS: An ensemble defense mechanism against adversarial attacks for intrusion detection system," in Proc. 30th Int. Conf. Comput. Commun. Netw. (ICCCN), 2021, pp. 1–9, doi: 10.1109/ICCCN52240.2021.9522227.
[27] P. T. Duy et al., "DIGFuPAS: Deception-based IDS with GAN for defending against adversarial samples," Comput. Secur., vol. 109, p. 102371, 2021, doi: 10.1016/j.cose.2021.102371.

[28] R. R. Devi and M. Abualkibash, “Intrusion detection system classification using machine learning algorithms,” Int. J. Comput. Sci. Inf. Secur., vol. 17, no. 3, pp. 239–246, 2019.
[29] A. Yulianto, P. Sukarno, and N. A. Suwastika, “Improving AdaBoost-based intrusion detection system (IDS) performance on CICIDS2017 dataset,” J. Phys. Conf. Ser., vol. 1192, no. 1, p. 012018, 2019, doi: 10.1088/1742-6596/1192/1/012018.
[30] N. Moustafa and J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” in Proc. Military Commun. Inf. Syst. Conf. (MilCIS), 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.
[31] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.
[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[34] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[35] A. A. Salih and A. M. Abdulazeez, “Evaluation of classification algorithms for intrusion detection system: A review,” J. Soft Comput. Data Min., vol. 2, no. 1, pp. 31–40, 2021, doi: 10.30880/jscdm.2021.02.01.004.
[36] Canadian Institute for Cybersecurity, “NSL-KDD Dataset,” 2009. [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html (Accessed: Mar. 24, 2026).
[37] Y. Fouad, A. N. Ghareeb, and E. Selem, “Routing protocols in wireless sensor networks: A review and classification,” Comput. Discov. Intell. Syst., vol. 2, no. 2, pp. 45–60, 2026.