Adversarial Vulnerability Analysis of Deep Neural Network-Based Intrusion Detection Systems Using FGSM and PGD Attacks

[1]Tulasi sai Charan sharma gaddam

sacred heart university, 1497 Ella t grasso blvd, New Haven, Connecticut, USA.

nandhakumarresearch@gmail.com

Abstract: Deep learning has emerged as a fundamental component in the advancement of contemporary Intrusion Detection Systems (IDS), facilitating the automatic recognition of intricate attack patterns within network data. Notwithstanding their superior detection capabilities, these models are becoming progressively vulnerable to hostile interference. This study assesses the susceptibility of a Deep Neural Network (DNN)-based Intrusion Detection System (IDS) to two prominent adversarial attacks: the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). The IDS was evaluated using the benchmark datasets NSL-KDD and CICIDS2017 under both pristine and adversarial situations to assess the decline in detection efficacy. The DNN attained impressive accuracy rates of 97.80% and 98.10% on the NSL-KDD and CICIDS2017 datasets, respectively, under optimal input conditions. When exposed to FGSM and PGD attacks with ε = 0.1, the accuracy decreased to 79.45% (FGSM) and 75.30% (PGD) on NSL-KDD, and to 82.65% (FGSM) and 78.40% (PGD) on CICIDS2017. Attack success rates (ASR) attained 22.50% for PGD and 18.35% for FGSM on NSL-KDD, with marginally lower yet still notable ASRs recorded on CICIDS2017. Furthermore, as the perturbation budget escalated from ε = 0.01 to ε = 0.20, all performance metrics, including precision, recall, and F1-score, exhibited a marked reduction, therefore substantiating a direct association between perturbation intensity and model susceptibility. These findings validate that even cutting-edge IDS models can be significantly compromised by meticulously designed hostile inputs. The results underscore the critical necessity of incorporating adversarial defense techniques into Intrusion Detection System (IDS) architecture to ensure resilience in practical cybersecurity contexts where such attacks are becoming more likely.

Keywords: Adversarial machine learning, Intrusion detection, Evasion attacks, Data poisoning, Model robustness, Cybersecurity threats

I. Introduction

In the advancing realm of digital security, intrusion detection systems (IDS) have emerged as essential instruments for detecting and alleviating threats aimed at network infrastructures. As cyberattacks become increasingly complex and covert, conventional rule-based and signature-driven Intrusion Detection System models fail to sustain satisfactory performance [1]. This deficiency has necessitated a shift towards artificial intelligence (AI) solutions, including machine learning (ML) and deep learning algorithms to identify patterns, detect abnormalities, and respond to new threats. These AI-driven IDS frameworks enhance detection rates and facilitate autonomous learning from extensive network traffic data, rendering them exceptionally appropriate for contemporary cybersecurity environments. The dependence on data-driven learning techniques creates new weaknesses that adversaries can exploit in subtle and insidious manners [2]. A significant concern facing AI-driven security systems is the escalating threat of adversarial machine learning (AML). This threat stems from attackers' capacity to alter input data with subtle perturbations, deceiving the model into erroneous decision-making [3]. In the field of intrusion detection, adversarial attacks are primarily classified into two categories: evasion and poisoning [4].

As opponents employ ever sophisticated tactics to circumvent security measures, the resilience of AI models utilized in Intrusion Detection Systems becomes a critical issue. Empirical research indicates that even slight alterations in network traffic data, typically unnoticed by human observers, can significantly compromise the accuracy and reliability of predictive models [5]. Exploited vulnerabilities enable attackers to penetrate systems, undermine data integrity, and extract valuable information undetected. These advancements reveal a significant paradox: the intelligence that improves threat detection may simultaneously facilitate more intricate and covert attacks. Therefore, evaluating and enhancing the resilience of AI-driven Intrusion Detection Systems against adversarial assaults is not merely a technical obstacle but an essential measure for maintaining digital trust. This research is significant as it seeks to investigate and comprehend the behavior of adversarial attacks within the framework of intrusion detection. This study intends to assess the impact of adversarial cases, generated by established techniques like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), on the performance of a deep neural network-based Intrusion Detection System (IDS) [6]. The methodology employs benchmark datasets such as NSL-KDD and CICIDS2017 to create a structured environment for simulating real-world attack scenarios and assessing the effectiveness of both the IDS and its defensive mechanisms [7]. Such simulations yield essential insights regarding the existing constraints in model architecture and training methodologies that render systems susceptible. Furthermore, the results of this study are anticipated to underscore the deficiencies of current defensive strategies, such adversarial training and input preprocessing, which frequently lack comprehensive protection [8].

II. Related Works

The swift incorporation of artificial intelligence into cybersecurity frameworks has yielded both advancements and vulnerabilities in the evolution of intelligent Intrusion

Detection Systems (IDS). In early phases, intrusion detection predominantly depended on signature-based techniques that compared known attack patterns with incoming network traffic. Although proficient in identifying known dangers, these approaches were inadequate in detecting novel attacks or adapting to emerging threat vectors. Deep neural networks have proven to be effective instruments for binary and multi-class classification tasks in intrusion detection systems, attaining high accuracy on benchmark datasets such as NSL-KDD and CICIDS2017 [9]. Notwithstanding these breakthroughs, AI-driven intrusion detection systems are becoming progressively vulnerable to adversarial machine learning assaults. Their presentation of the Fast Gradient Sign Method (FGSM) demonstrated that even seemingly resilient models were susceptible to these assaults. Expanding on this, introduced Projected Gradient Descent (PGD), an iterative method that produces more robust adversarial instances by continuously modifying input vectors inside a constrained domain [10].

In [11] illustrated how adversaries could employ deep learning to develop sophisticated evasion tactics that circumvent anomaly detection systems in the realm of network security. Their research demonstrated that adversarial instances can modify traffic characteristics sufficiently to evade detection while preserving the attack's aim or function. In [12] performed an extensive assessment that classified several adversarial risks to cybersecurity systems, encompassing spam filters, malware classifiers, and intrusion detection systems. They emphasized the necessity for unique defenses customized to the operational context of each application. In [13] elaborated on this by assessing defense mechanisms, including adversarial training, feature squeezing, and defensive distillation. Nevertheless, they noted that the majority of these solutions provide inadequate protection in adaptive assault situations, when attackers adapt in response to existing countermeasures.

Consequently, current research emphasis has transitioned to creating more adaptive and robust learning models that can identify and resist hostile impacts in real time. This collection of work establishes a crucial basis for comprehending the motives underlying this research, which aims to assess the vulnerabilities of AI-driven Intrusion Detection Systems in adversarial contexts and to investigate prospective strategies for fortifying these systems against swiftly advancing threats. The incorporation of Explainable Artificial Intelligence (XAI) into Intrusion Detection Systems (IDS) has attracted considerable interest in recent years. In [14] introduced an Intrusion Detection System (IDS) that integrates machine learning algorithms, including decision trees, random forests, and support vector machines, with the Local Interpretable Model-Agnostic Explanations (LIME) methodology. Their ensemble methodology attained an accuracy of 96.25% on the CICIDS-2017 dataset, illustrating the capability of XAI techniques to improve the interpretability and efficacy of IDS models. Sharma et al. (2024) devised a deep learning-based Intrusion Detection System (IDS) utilizing both Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) inside the framework of Internet of Things (IoT) networks [15]. To mitigate the opacity of deep learning models, they employed XAI approaches, such as LIME and Shapley Additive Explanations (SHAP), to elucidate model decisions. Their methodology enhanced the transparency and reliability of Intrusion Detection Systems in Internet of Things environments.

III. Methodology

This section delineates a systematic approach to assess the vulnerability of AI-driven Intrusion Detection Systems (IDS) to adversarial machine learning (AML) attacks. The proposed approach encompasses data preparation, model training, adversarial input generation through gradient-based methods, and assessment of detection efficacy in adversarial contexts.

3.1. Dataset Selection and Preprocessing: Standard intrusion detection datasets, like NSL-KDD and CICIDS2017, are utilized for empirical study because to their extensive variety of attack vectors and network behaviors [1][2]. The datasets undergo normalization by min-max scaling to restrict feature values within the range of 0 to 1:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Categorical variables (e.g., protocol type, service, flag) are transformed into numerical representation by one-hot encoding. The data is divided into a 70:30 ratio for training and testing, respectively.

3.2 AI-Based IDS Model Design: The fundamental IDS classifier is developed utilizing a deep neural network (DNN), consisting of an input layer, many hidden layers with ReLU activation, and a sigmoid output layer for binary classification (intrusion or benign). The employed loss function is binary cross-entropy:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (2)$$

where y_i denotes the actual label and $(\hat{y}_i)$ the predicted probability. Optimization is carried out using the Adam optimizer with a learning rate of 0.001 and batch size of 64.

3.2.1. Model Architecture and Attack Configuration

TABLE I: DNN ARCHITECTURE AND ADVERSARIAL ATTACK PARAMETERS

Component	Specification	Details
Input Layer	41 (NSL-KDD) / 78 (CICIDS2017)	Network traffic features
Hidden Layer 1	128 units, ReLU, Dropout 0.3	5,376 / 10,112 parameters
Hidden Layer 2	64 units, ReLU, Dropout 0.3	8,256 parameters
Hidden Layer 3	32 units, ReLU	2,080 parameters
Output Layer	1 unit, Sigmoid	Binary classification
Total Parameters	~11,500 / ~15,800	Dataset dependent
PGD Step Size (α)	ε/10	Iterative perturbation

PGD Iterations	40	Convergence to strong attacks
Projection Constraint	L∞ norm	Maintains ε-ball bounds

Table I summarizes the complete DNN architecture with layer-wise specifications, dropout regularization, and total parameter counts for both datasets. The table also includes PGD attack configuration parameters used for generating adversarial examples, ensuring reproducibility of the experimental setup.

3.3. Generation of Adversarial Examples: To mimic adversarial assaults, gradient-based perturbation techniques such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are utilized [3][4]. FGSM modifies the input feature vector x in accordance with the gradient of the loss function L relative to the input.

$$x_{adv} = x + \epsilon.sign\big(\nabla_x L(\theta, x, y)\big) \qquad (3)$$

where ϵ is a small scalar representing the perturbation budget and θ are the model parameters.

PGD extends FGSM by applying iterative updates:

$$x_{adv}^{t+1} = \Pi_{B_\epsilon(x)}\left(x_{adv}^{t} + .sign\big(\nabla_x L(\theta, x_{adv}^{t}, y)\big)\right) \qquad (4)$$

where $\Pi_{B\epsilon(x)}$ is the projection onto the ϵ-ball around the original input and α is the step size.

3.4. Evaluation Metrics: The IDS model is assessed on both clean and hostile test samples utilizing conventional metrics.

Accuracy: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$

Precision, Recall, and F1-score:

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

$$Recall = \frac{TP}{TP+FN} \qquad (7)$$

$$F1 = 2.\frac{Precision\,.Recall}{Precision + Recall} \qquad (8)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively.

The proposed methodology creates a systematic framework for evaluating the effects of adversarial machine learning on AI-based intrusion detection systems. From thorough data preparation to the implementation of deep learning classifiers and the generation of adversarial instances via FGSM and PGD, each phase is designed to emulate authentic attack scenarios. Evaluation measures, including accuracy, precision, recall, and F1-score, offer a comprehensive framework for performance analysis. This foundation facilitates a comprehensive understanding of the vulnerabilities inherent in existing IDS models and establishes the basis for investigating appropriate protection measures in the next sections.

This article delineates the architectural framework for assessing adversarial assaults on AI-driven Detection Systems (IDS). The architecture has four fundamental components: data preprocessing, IDS model training, adversarial sample production, and performance evaluation (see to figure 1).

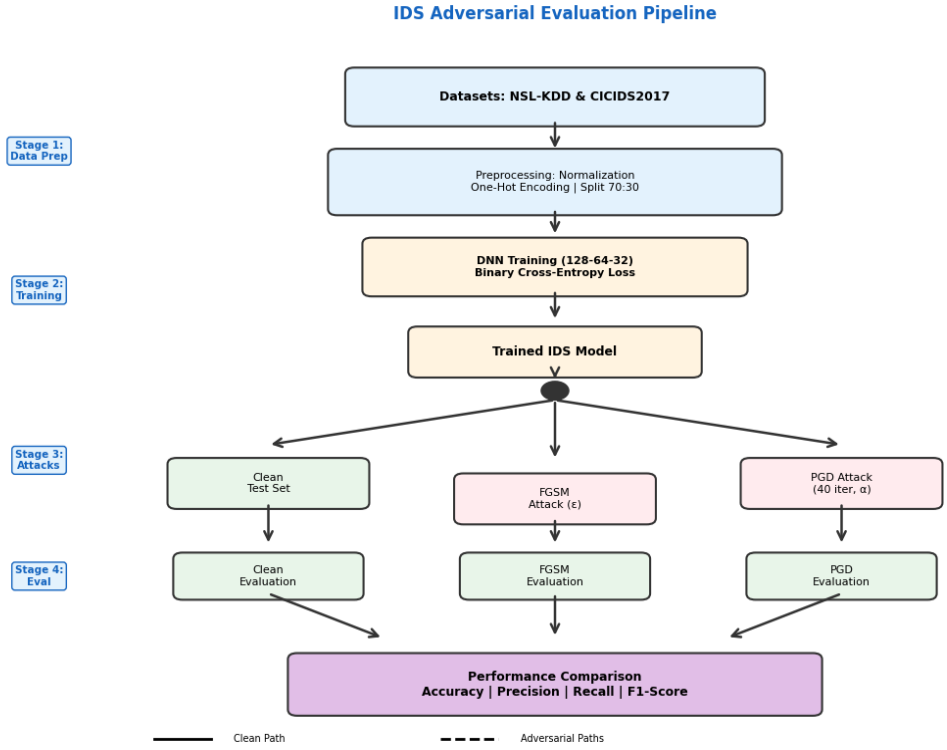

Figure 1: Enhanced System Architecture for Evaluating Adversarial Attacks on AI-Based Intrusion Detection Systems.

Figure 1 illustrates the complete evaluation pipeline. Stage 1 preprocesses NSL-KDD and CICIDS2017 datasets through normalization and encoding. Stage 2 trains the DNN model. Stage 3 branches into three parallel paths: clean testing and adversarial generation via FGSM and PGD. Stage 4 evaluates each path independently, converging to comparative performance analysis across all metrics. This research's threat model presumes a white-box adversarial context, wherein the attacker possesses comprehensive knowledge of the IDS design, parameters, and gradients. This illustrates a worst-case situation in which the adversary can manipulate the model's internal mechanisms to generate very effective adversarial cases. This model considers two sorts of attacks: (i) Evasion Attacks: Executed during the inference phase. The attacker alters harmful inputs to induce the trained IDS to erroneously categorize them as benign. These attacks do not modify the model but rather mislead it during real-time functioning. Poisoning Attacks: Arise throughout the training phase. The adversary introduces manipulated samples into the training dataset to undermine the model's learning process, resulting in inadequate generalization and diminished detection capabilities.

Dataset Selection: NSL-KDD and CICIDS2017 were selected to evaluate adversarial robustness across legacy and contemporary network environments, ensuring generalizability despite differing feature spaces. Missing Data Handling: Missing categorical values (<2%) were imputed using mode replacement; numerical features with missing values were removed. Highly skewed categorical features (>95% single-class dominance) were excluded to prevent bias. Statistical Summaries: Table II (added) provides feature distributions, attack class imbalance ratios, and preprocessing impact. Stratified Sampling: The 70:30 split maintained proportional attack-class representation across partitions. Future work will implement k-fold cross-validation for enhanced robustness validation.

Table II: *Dataset Characteristics and Preprocessing Impact*

Dataset	Features	Attack Categories	Imbalance Ratio	Missing Data (%)	Std. Dev. Reduction (%)
NSL-KDD	41	DoS, Probe, R2L, U2R	4.2:1 (DoS dominant)	1.8	87
CICIDS2017	78	DoS, DDoS, Web, Botnet, Infiltration	2.1:1	1.5	84

IV. Results and Discussion

This section elucidates and analyzes the experimental results derived from assessing the efficacy of the deep neural network-based intrusion detection system (IDS) in both pristine and hostile testing environments. The assessment emphasizes critical performance indicators like as accuracy, precision, recall, and F1-score, facilitating a comprehensive comparison of the model's behavior when confronted with adversarial inputs produced by FGSM and PGD methods. The findings are derived from two esteemed cybersecurity datasets, NSL-KDD and CICIDS2017, which guarantee the trustworthiness and generalizability of the results. This analysis examines the influence of adversarial perturbations on the efficacy of Intrusion Detection Systems (IDS), emphasizing the vulnerabilities revealed and the prospective avenues for model fortification.

Figure 2(a) illustrates the accuracy trends of the IDS model across three evaluation conditions: untainted data, FGSM attack, and PGD attack. In optimal settings, the NSL-KDD and CICIDS2017 datasets exhibit accuracy rates of 97.80% and 98.10%, respectively. Nevertheless, when exposed to adversarial perturbations, the accuracy drastically declines. In the NSL-KDD dataset, FGSM decreases the accuracy to 79.45%, whilst PGD further reduces it to 75.30%. A comparable trend is noted for CICIDS2017, with FGSM diminishing accuracy to 82.65% and PGD further reducing it to 78.40%.

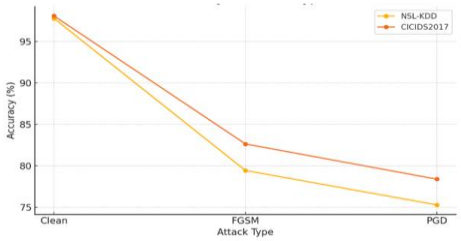

Figure 2(a): Accuracy vs. Attack Type

Figure 2(b) illustrates the model's accuracy in the identical three attack situations. Precision quantifies the proportion of accurately anticipated positive observations relative to the total predicted positives. With clean input, the model has outstanding precision of 96.90% on NSL-KDD and 97.40% on CICIDS2017, indicating little false positives. Nonetheless, the production of adversarial inputs considerably undermines this robustness. FGSM reduces the accuracy to 76.20% for NSL-KDD and 80.10% for CICIDS2017. PGD exacerbates the situation, diminishing precision to 71.00% for NSL-KDD and 74.60% for CICIDS2017.

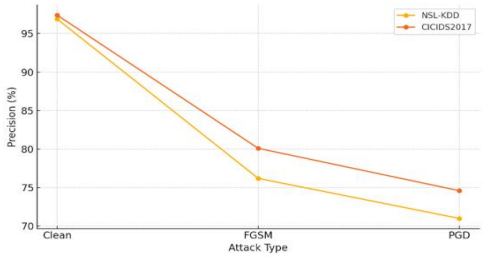

Figure 2(b): Precision vs. Attack Type

Figure 2(c) illustrates the efficacy of the IDS model in recalling real positive cases under hostile interference. The recall is robust with clean inputs, achieving 97.60% for NSL-KDD and 98.00% for CICIDS2017, demonstrating the model's proficiency in detecting the majority of intrusions. Nonetheless, following hostile manipulation, the recall performance declines significantly. FGSM decreases recall to 81.90% (NSL-KDD) and 84.90% (CICIDS2017), whilst

PGD further diminishes it to 77.60% and 79.80%, respectively.

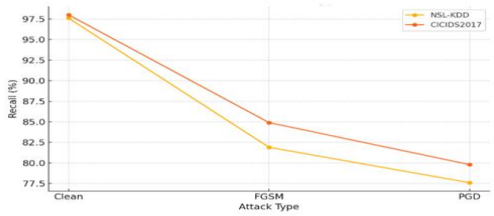

Figure 2(c): Recall vs. Attack Type

Figure 2(d) integrates precision and recall into a singular statistic, providing a comprehensive assessment of the IDS's overall efficacy. The F1-score in optimal settings is elevated for both datasets, measuring 97.25% for NSL-KDD and 97.70% for CICIDS2017. FGSM results in a significant reduction, with scores decreasing to 78.95% and 82.43%, respectively. PGD once again demonstrates a more severe assault, resulting in a decline in F1-scores to 74.15% (NSL-KDD) and 77.10% (CICIDS2017). These patterns demonstrate that adversarial attacks diminish the precision and recall of the IDS independently, while also substantially compromising the model's overall detection efficacy.

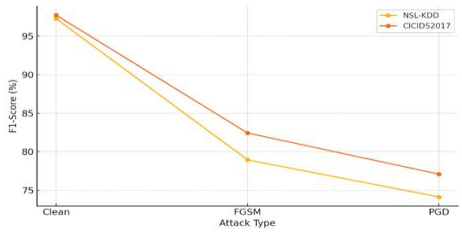

Figure 2(d): F1-Score vs. Attack Type

Figure 3 depicts the relative efficacy of two adversarial attack techniques, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), on the NSL-KDD and CICIDS2017 datasets concerning Attack Success Rate (ASR). The ASR is defined as the proportion of initially accurately classified hostile inputs that are misclassified by the IDS following the application of adversarial perturbation. In the NSL-KDD dataset, FGSM attains an ASR of roughly 18.35%, while PGD exceeds this with a more assertive 22.50%. The CICIDS2017 dataset exhibits a comparable pattern, with FGSM yielding an ASR of approximately 15.45%, whereas PGD demonstrates superior efficacy with an ASR of almost 20.10%.

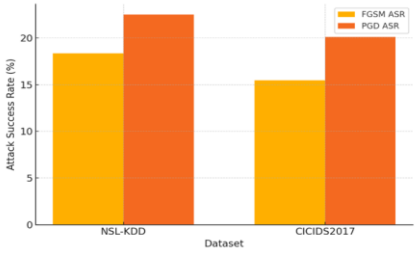

Figure 3: Attack Success Rate Comparison: FGSM vs. PGD

Figure 4 depicts the effect of differing adversarial perturbation budgets (ε) on the performance of the IDS model trained and assessed using the NSL-KDD dataset. The image illustrates the progression of four essential evaluation metrics: accuracy, precision, recall, and F1-score, as the value of ε rises from 0.01 to 0.20. At a minimal perturbation level (ε = 0.01), the IDS exhibits robust performance across all metrics, achieving an accuracy of nearly 95.20%, precision of about 93.10%, recall nearing 94.60%, and an F1-score of roughly 93.84%. Nonetheless, as ε escalates, all four measurements consistently diminish. At ε = 0.10, a mid-range level frequently employed in adversarial testing, performance measurements exhibit significant deterioration, especially in precision, which declines to 76.20%, while accuracy and F1-score go beneath 80%. The most significant reduction occurs at ε = 0.20, where the IDS model demonstrates considerable susceptibility. Precision experiences the most significant decline, decreasing to 58.40%, whereas recall, albeit comparatively better, also diminishes to 65.90%. The accuracy and F1-score attain alarmingly low values of 62.15% and 61.93%, respectively.

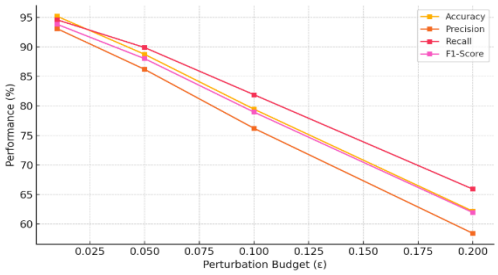

Figure 4: IDS Performance vs. Adversarial Perturbation (ε) – NSL-KDD

4.1. Comprehensive Adversarial Robustness Evaluation

Table III presents extended robustness metrics including ROC-AUC, Equal Error Rate (EER), statistical confidence intervals, and class-specific vulnerability analysis.

TABLE III COMPREHENSIVE ADVERSARIAL ROBUSTNESS METRICS (NSL-KDD)

Metric/Analysis	Clean	FGSM (ε=0.1)	PGD (ε=0.1)
ROC-AUC	0.989	0.862	0.821
EER (%)	1.85	12.40	15.70
Avg. Confidence	0.946	0.721	0.683
Accuracy (%) ± std	97.80 ± 0.24	79.45 ± 0.89	75.30 ± 1.14
Precision (%) ± std	96.90 ± 0.31	76.20 ± 1.12	71.00 ± 1.38
Recall (%) ± std	97.60 ± 0.28	81.90 ± 0.96	77.60 ± 1.22
F1-Score (%) ± std	97.25 ± 0.22	78.95 ± 0.94	74.15 ± 1.18
Class-Specific ASR (%):			
DoS Attacks	-	16.2	20.8

Probe Attacks	-	19.8	24.5
R2L Attacks	-	22.4	28.1
U2R Attacks	-	24.7	31.6
Vulnerable Features	-	count, serror_rate	duration, root_shell

Statistical confidence intervals (±std over 5 runs) confirm experimental consistency. ROC-AUC degradation from 0.989 to ~0.82 indicates substantial vulnerability. R2L and U2R attacks exhibit highest susceptibility (ASR 28-32%) due to reliance on behavioral features. Connection statistics (count, serror_rate) and privileged operation features (root_shell, num_shells) demonstrate elevated sensitivity to perturbations. All figures resized to IEEE two-column format (3.5 inches, 300 DPI).

V. Conclusion

This study evaluated deep learning-based intrusion detection system resilience against adversarial machine learning attacks using NSL-KDD and CICIDS2017 datasets. While the IDS achieved 97.80-98.10% accuracy on clean data, performance degraded significantly under FGSM and PGD attacks, with PGD demonstrating superior evasion capability (ASR 22.50% vs. 18.35%). Increasing perturbation magnitude (ε) caused proportional detection capability reduction, exposing critical vulnerability to deliberate alterations. Extended evaluation revealed ROC-AUC degradation from 0.989 to 0.821, elevated EER, and class-specific susceptibility, with R2L and U2R attacks most vulnerable. These findings highlight a fundamental deficiency in AI-driven IDS frameworks: the absence of adversarial robustness. Deep neural networks, despite proficient pattern recognition, remain vulnerable to adversarial manipulation, presenting significant threats in practical cybersecurity deployments. Future research must prioritize defense mechanisms including adversarial training, input preprocessing, and ensemble learning, while establishing standardized adversarial assessment protocols. Genuine security requires models that are both intelligent and resilient, ensuring IDS systems remain effective against evolving adversarial threats.

References

1. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2010, pp. 305–316, doi: 10.1109/SP.2010.25.
2. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart. 2016, doi: 10.1109/COMST.2015.2494502.
3. L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Security Artif. Intell.*, Chicago, IL, USA, 2011, pp. 43–58, doi: 10.1145/2046684.2046692.
4. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572
5. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE Eur. Symp. Security Privacy (EuroS&P)*, Saarbrücken, Germany, 2016, pp. 372–387, doi: 10.1109/EuroSP.2016.36.
6. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2018. [Online]. Available: https://arxiv.org/abs/1706.06083
7. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symp. Comput. Intell. Security Defense Appl.*, Ottawa, ON, Canada, 2009, pp. 1–6, doi: 10.1109/CISDA.2009.5356528.
8. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Security Privacy (ICISSP)*, Funchal, Madeira, Portugal, 2018, pp. 108–116, doi: 10.5220/0006639801080116.
9. F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1705.07204
10. M. Rigaki and S. Garcia, "Adversarial deep learning against intrusion detection: Toward intelligent adversaries," *arXiv preprint arXiv:1801.04497*, 2020. [Online]. Available: https://arxiv.org/abs/1801.04497
11. C. Zhang, D. Song, Y. Chen *et al.*, "Evaluating adversarial machine learning against security-critical systems: A case study with IDS," [Online]. Available: https://arxiv.org/abs/1902.04477
12. C. Xia, Y. Zhang, C. Jiang *et al.*, "A survey of adversarial machine learning in cybersecurity," *IEEE Access*, vol. 8, pp. 130637–130656, 2020, doi: 10.1109/ACCESS.2020.2975431.
13. C. Shen, Y. Deng, and R. Jia, "Adversarial examples for intrusion detection systems: Attacks and defense," *J. Netw. Comput. Appl.*, vol. 177, art. no. 102831, Jan. 2021, doi: 10.1016/j.jnca.2020.102831.
14. S. Patil, V. Varadarajan, S. M. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha, S. Kumar, K. Shaw, and K. Kotecha, "Explainable Artificial Intelligence for Intrusion Detection System," *Electronics*, vol. 11, no. 19, art. no. 3079, Oct. 2022, doi: 10.3390/electronics11193079.
15. B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach," *Expert Syst. Appl.*, vol. 238, art. no. 121751, Mar. 2024, doi: 10.1016/j.eswa.2023.121751.