Educational Administration: Theory and Practice
2023,29(4), 1960 - 1968
ISSN:2148-2403
https://kuey.net/

Research Article

Vulnerability Analysis Of ML-Based Intrusion Detection Systems Against Evasion Attacks

Sushil Buriya[1*], Neelam Sharma[2]

[1*] Research Scholar, Department of Computer Science, Banasthali Vidyapith
[2] Associate Professor, Department of Computer Science, Banasthali Vidyapith

Citation: Sushil, (2023), Vulnerability Analysis Of ML-Based Intrusion Detection Systems Against Evasion Attacks, *Educational Administration: Theory and Practice*, *29*(4), 1960 - 1968
Doi: 10.53555/kuey.v29i4.6791

ARTICLE INFO

ABSTRACT

Intrusion Detection Systems (IDS) are essential component of cyber security countermeasures to protect networks against cyber-attacks. IDS have become more sophisticated and capable of identifying complex attack patterns with the initiation of machine learning (ML) techniques in IDS detection engine. However, adversarial evasion attacks are a significant threat towards Machine Learning. These attacks involve subtly modifying malicious inputs to evade detection while maintaining their malicious intent. This paper presents a comprehensive comparative analysis of the impact of various adversarial evasion attack techniques on different machine learning models used in IDS implementations. We evaluate the robustness of commonly used models. Logistic Regression, Gradient Boosting Classifier, and Multi-layer Perceptron, are evaluated against FGSM and PGD adversarial attacks. We demonstrate the vulnerabilities of each model and discuss the implications of these findings for the design and deployment of robust IDS. The results highlight the necessity for adversarial defense methods to mitigate the risks posed by adversarial evasion attacks to ensure the reliability and security of ML-based IDS in real-world applications.

1. Introduction

Intrusion Detection Systems (IDS) play a vital role in the cyber security solutions of organizations to protect the networks by monitoring the network traffic. IDS have ability to detect suspicious activities and potential threats in network traffics by searching the signature or patterns of malicious activities in network logs. Traditional IDS highly rely on signature-based or rule-based methods. These methods are only effective against known threats but not perform well in identifying zero-day-attacks [1].

Machine Learning (ML) has been emerged as a powerful tool for enhancing the detection capabilities of IDS. IDS can learn from historical network data to identify patterns and anomalies associated with malicious activities by leveraging the pattern recognition power of ML algorithms. This ability to generalize from past cyber-attacks behaviour enables ML-based IDS to detect previously unseen cyber-threats to make them more robust and versatile compared to traditional methods [2]. Various ML models have been successfully applied in the context of IDS to improve detection accuracy and efficiency. Logistic Regression, Gradient Boosting Classifier, and Multi-layer Perceptron have shown significantly impressive performance for IDS datasets.

Despite all the aforementioned benefits of ML in IDS, these ML models have been proven not to be ideal and are still vulnerable to attacks. One such kind of powerful attack is an adversarial evasion attack, whereby an attacker manipulates input data intentionally in order to fool the ML model to misclassifying a malicious activity as benign [3]. These imperceptible perturbations can result in quite substantial performance degradation in the ML-based IDS. Realization of the implications of adversarial evasion attacks will be key for designing robust and trustworthy IDS.

The primary objective of the present paper is a comparative analysis of the impact of adversarial evasion attacks on the different ML models used in the IDS. In our experiments, we tested three algorithms, Logistic Regression, Gradient Boosting Classifier, and Multi-layer Perceptron against two gradient based adversarial attack techniques: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD. We show in extensive experiments how vulnerable each of these models is and discuss the implications of these on the design and deployment of robust IDS. The contributions of this paper are threefold:

- A fine-grained performance evaluation on performance degradation of various ML models in IDS with adversarial evasion attacks.
- A comparative analysis that illustrates the weaknesses and strengths of the ML models in adversarial manipulations.
- Insight towards potential defence strategies to improve the strength of ML-based IDS in such an attack scenario.
-

The contributions of this research comprise making the IDS more secure and resilient by elaborating on thevulnerabilities of ML models to adversarial evasion attacks through effective countermeasures proposed

2. Related Work

Mourabit *et. al.* conducted on some issues of network security using the Naive Bayes, Random Forest, Support Vector Machine, and K-means ML algorithms that identify attacks of the following four types: DOS, PROBE, U2R, R2L. They have derived the result that the developed RFC is much more effective than existing methods; the hierarchical clustering method can easily improve system performance [4]. In paper [5], the author compares Random Forest, Support Vector Machine, Gaussian Naive Bayes, and Logistic Regression classifiers for network intrusion detection in supervised machine learning. The best-performing algorithm was determined according to metrics like F1-Score, accuracy, precision, and recall. The result indicated that with these parameters, the Random Forest Classifier performed better than other classifiers. Wang *et al.* introduced a intrusion detection model utilizing logistic regression. Evaluation on the NSL-KDD dataset demonstrates that their approach achieves commendable results in accuracy, detection rate, and false alarm rate [6].

Recent Literature has demonstrated the capability of adversarial perturbations, even of small magnitude, to severely affect machine learning model-based detectors, but the solutions are still in their nascent stage [7], [8], [9]. In recent work, Hu, W. *et al.,* presented an algorithm called MalGAN, which performs attacks against a machine learning-based malware detector by running it in a black box setting. A surrogate detector was designed based on a neural network to replace the original detector against malware. An adversarial example generator, trained from a neural network, was utilized to generate adversarial examples that could fool this surrogate detector. Such an approach helps regulate the flow of distributions of adversarial examples and changes in probability distribution quickly disorient the learning process of the malware detector. It showed that adversarial attacks could be potentially effective in a black-box setting without prior knowledge of the machine learning algorithm [10].

Biggio, B. *et al.* proposed a testing evasion attacks method through gradient descent against neural networks (NNs) and Support Vector Machines (SVMs). Their proposed evasion attack has been tested through an attack on a handwritten digit classification task and an attack on a malware detection system for PDF files. The results show how popular classification algorithms, that is, NNs and SVMs, can be easily fooled when the attacker possesses a very low amount of knowledge about the training data. This brings to attention the concerns for using classification algorithms in applications sensitive to security issues [11].

Ensemble models have shown promising performance for Network Intrusion Detection Systems (NIDS) by combining multiple classifiers to enhance detection accuracy [12], [13]. While individual machine learning models like Support Vector Classifier (SVC) and Multi-Layer Perceptron (MLP) have demonstrated effectiveness in NIDS, ensemble methods can further improve detection rates by leveraging the strengths of different classifiers. Additionally, incorporating a feedback mechanism in the NIDS model can enhance learning from past predictions and rectifications, contributing to better overall performance. By utilizing ensemble techniques alongside feature selection methods and feedback mechanisms, NIDS can achieve higher accuracy rates in detecting network intrusions, crucial for maintaining robust network security in the face of evolving cyber threats.

Most of the discussed literature underlines high vulnerabilities for machine learning models, especially neural networks and Support Vector Machines, under evasion attacks. Literatures show that adversarial techniques, including gradient descent and neural network-based adversarial example generation, can compromise effectively popular classification algorithms in the black-box setting with a minimum a priori knowledge. The ML models are evaluated on image datasets in most of the recent literature. There is strong need of assessment of ML models to evaluate the robustness against evasion attack with network datasets due to the different nature of data distribution from image datasets.

3. Methodology

In this study, we adopted the gray-box threat model, where an adversary has partial knowledge of the ML-models incorporated in IDS and have access to the training and testing datasets. It means with these restrictions, the adversary is able to generate adversarial examples based on methods such as Projected Gradient Descent and Fast Gradient Sign Method. Adversaries use these techniques to craft perturbations that mislead the trained model into making wrong decisions that is, classifying the malicious attack traffic as

benign. It exploits partial information and advanced attack methods toward the compromise of the model's integrity and reliability, leading to misbehaviors in distinguishing among legitimate and malicious network traffics. This approach emphasizes the vulnerabilities in the robustness of the model and the need for stronger developments in defenses against sophisticated gray-box attacks. Detailed approach is described in Figure-1.

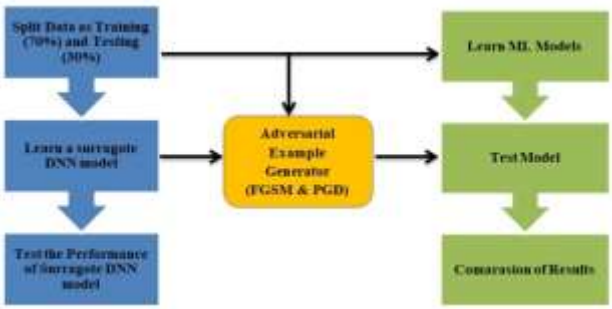

Figure 1: *Block Diagram of Methodology*

3.1 Machine Learning Algorithms

Three machine learning models: Logistic Regression (LR), Gradient Boosting Classifier (GBC), and Multi-layer Perceptron (MLP) are used for the comparison. LR is very simple in architecture and very efficient in terms of computation requirement. GBC is an ensemble learning model and Feed-Forward Neural Network.

3.1.1 Logistic Regression (LR)

Logistic Regression is a linear model for binary classification that estimates the probability of a binary outcome based on one or more predictor variables. Despite its simplicity, it is widely used in IDS due to its interpretability and efficiency. Logistic regression models the probability $P(Y=1|X)$, where Y is the binary dependent variable (target variable) and X represents the independent variables (features). The output of logistic regression is a probability score between 0 and 1 [14].

3.1.2 Gradient Boosting Classifier (GBC)

Gradient Boosting Classifier is an ensemble learning technique that builds a series of decision trees, where each tree corrects the errors of its predecessor. This method is known for its high predictive accuracy and robustness, making it suitable for complex IDS tasks. GBC builds an ensemble of trees sequentially, where each tree corrects errors made by the previous one. It focuses on reducing the errors (residuals) of the model. Unlike traditional boosting algorithms that adjust weights, GBC fits each new tree to the residuals (gradient) of the loss function of the previous model [15]. This gradient descent approach makes it particularly effective in reducing bias and improving predictive accuracy.

3.1.3 Multi-layer Perceptron (MLP)

Multi-layer Perceptron is a class of feed forward artificial neural networks consisting of multiple layers of nodes. Each node, or neuron, in one layer connects with a certain weight to every node in the following layer. MLPs are capable of capturing complex patterns in data, which is advantageous for detecting intricate intrusion patterns [16].

3.2. Adversarial Threat Model

An adversarial example is a training sample with slight, purposeful changes in its features to bring about misclassification in a machine learning (ML) model. The very existence of adversarial examples makes ML models vulnerable to adversarial attacks.

The large capacity of ML architectures often leads to the occurrence of gaps between the distribution of data that the model can withstand and the actual underlying data distribution, resulting in insufficient exploration of the data distribution in training datasets, and hence, a gap existing in the training data manifold. Adversarial attacks exploit such unexplored regions, referred to as adversarial subspaces, by carefully perturbing features in legitimate training samples with synthetic noise.

In this paper, we probe two well-known techniques for adversarial evasion attack against the named ML models above. Evasion attacks produce a form of adversarial example that misleads the model during inference into making the wrong decision.

In this scenario, the adversarial test samples can be tampered in such a way that they can pass by the sensing mechanism and get classified as not malicious. Figure-2 shows the evasion attack adversary while undergoing the DNN model testing procedure. Adversarial attacks can affect both linear and nonlinear classifiers and are

categorized under the white-box attacks because the adversary will have knowledge of the classification model to design adversarial examples.

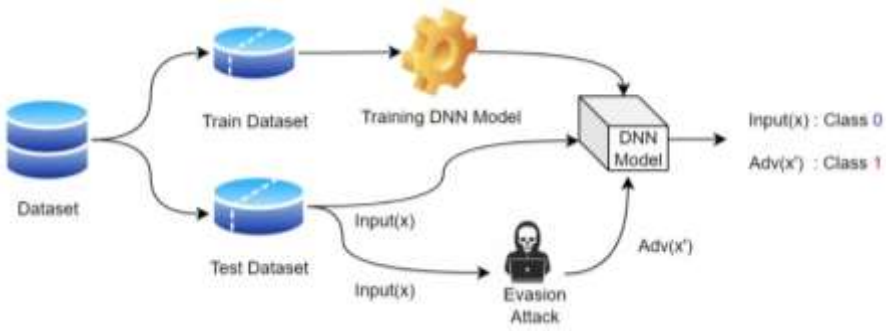

Figure 2*: Adversarial threat model for adversarial examples generation*

3.2.1. Fast Gradient Sign Method (FGSM)

FGSM generates adversarial examples by adjusting the input data in the direction of the gradient of the loss function with respect to the input. This method perturbs the input by a fixed amount in the direction that increases the model's error. FGSM is a single-step adversarial attack that perturbs the input data in the direction of the gradient of the loss function with respect to the input [17]. The formula for generating an adversarial example using FGSM is:

$$x' = x + \varepsilon * sign(\nabla(J(\theta, x, y)) \quad (1)$$

Here x' denotes the adversarial example corresponding to input sample x, ε is the constant parameter, J is the loss function of the classifier, and θ denotes the learning parameter of the model for input sample x with target class label y.

3.2.2. Projected Gradient Descent (PGD)

PGD is an iterative variant of FGSM that applies small perturbations multiple times, projecting the perturbed input back into the feasible input space after each step. This method is more effective than FGSM in finding adversarial examples that are harder to defend against. PGD is an iterative attack that applies FGSM multiple times with smaller step sizes and projects the perturbed input back onto the ε-ball around the original input [18].

Given an input sample x, the goal of the PGD attack is to find an adversarial input x' that maximizes the loss function L(f(x'), y) subject to a constraint on the distance between x and x' under some distance metric d(x, x') ≤ ϵ. Here, f(x') denotes the output of the model for the adversarial input x', and y is the true label of the input x. The PGD attack can be formulated as an iterative algorithm-1.

Algorithm 1 PGD attack for binary classifiers

Input: input vector x, binary classifier f, step size α, number of iteration k, perturbation constant ε
Output: adversarial example x'

1. **Initialize** $x' = x$
2. **for i=1 to k**
3. Compute the gradient of the loss function with respect to input as
$$d = \nabla L(f(x'), y)$$
4. Compute the perturbation
$$\delta = \alpha \cdot sign(d)$$
5. Project the perturbed input
$$\textbf{if}(\delta < -\varepsilon) \quad x' = x' - \varepsilon$$
$$\textbf{else if}(\delta > \varepsilon) \quad x' = x' + \varepsilon$$
$$\textbf{else} \quad x' = x' + \delta$$
6. **end for**
7. **return** x'

For our experiments, we use a well-known benchmark dataset for IDS, such as the NSL-KDD dataset. This dataset includes a wide range of network traffic features and labels indicating normal or various types of attack traffic.

3.3 Experimental Setup

In the experimental setup, we have trained a Deep Neural Network (DNN) model as surrogate model for binary classification of the network traffic instance as benign or attack based on statistical information on network flow. Adversarial examples are generated using the surrogate model with FGSM and PGD methods with gary-box threat model, and the ML models are evaluated on these adversarial examples.

We have used NSL-KDD dataset to evaluate the impact of evasion attacks over ML-models for IDS implementation. NSL-KDD have flow based network traffic. The dataset has been pre-processed to remove missing values and normalization of numerical values. The categorical values are encoded using label encoder technique.

NSL-KDD is the improved version of the KDD 99 dataset that addresses and overcome from various problems of the KDD 99 dataset. This dataset contains 5 different attack classes. In this work, we have used the binary classes (attack and benign) only. Dataset describes 42 features, including class label. It has flow-based features. Each row in the NSL-KDD dataset is labeled as 0 for normal, and 1 for attack records. In our experiments, we used 117478 records for training and 57863 for testing, including 56000 benign samples, and 119341 attack samples [19].

3.3.1 DNN Model Implementation

We have used the DNN architecture for the implementation of IDS as shown in Figure-3 to experiment on both NSL-KDD dataset. The ADAM optimizer has been used and sliced the data into 64 batches repeated over the 20 epochs for training. First, we have trained the model using the training dataset and analyzed the performance using the test dataset. Then we generate the attack samples using this DNN surrogate model with FGSM and PGS adversarial attacks and analyzed the impact of attack on ML models Logistic Regression, Gradinet Boosting Classifier and Mluti-Layer Perceptron under evasion attack and non-adversarial environment for comparison.

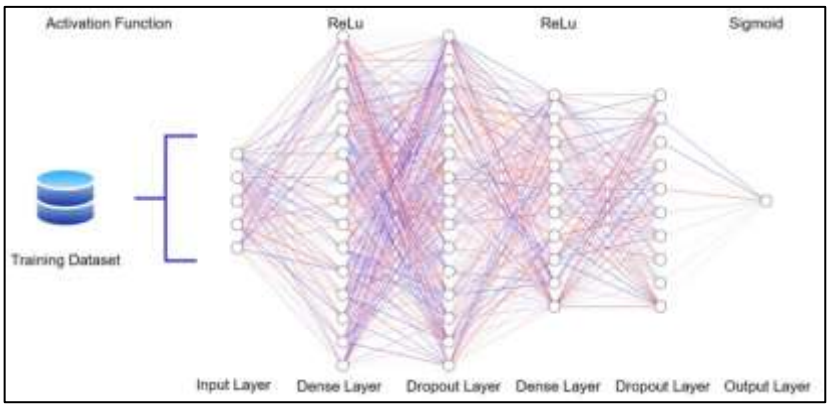

Figure 3: *Architecture of DNN employed as Surragolte model for gray box attacks*

3.3.2 Evaluation Metrics

The assessment metrics are playing important roles in evaluating the performance of a machine learning task. In this study, our task is to classify the network traffic flow into two classes, i.e., binary classification, where positive is an attack traffic flow and negative is a normal traffic flow. To assess the performance of the ML models under adversarial and non-adversarial conditions, we use some parameters like Accuracy, Precision Rate (PR), Recall Rate (RR), and F1-Score. Our model's predictions can be classified into two classes: either correct (True Positive, True Negative) or incorrect (False Positive, False Negative). True Positives (TP) are the number of attack traffic flows correctly labelled as an attack, while True Negatives (TN) represent the number of benign traffic flows correctly labelled as benign. False Positives (FP) are benign samples that get misclassified as attacks, while False Negatives (FN) is actually attack samples that are misclassified as benign. These values have the application of defining the classification metrics. Based on this evaluation metrics are defined:

Accuracy: It describes how often the classifier predicts the correct class for an attack and benign sample. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision Rate: It explains how many of the attacks predicted traffic samples turned out to be positive.

$$Precision\ Rate\ (PR) = \frac{TP}{TP+FP} \quad (3)$$

Recall Rate: It defines how many of the actual attack traffic samples are predicted correctly with our model. The recall rate should be high for a network traffic analyzer.

$$Recall\ Rate\ (RR) = \frac{TP}{TP+FN} \quad (4)$$

F1 Score: The F1 Score is the harmonic mean of precision and recall. It is maximum when Precision is equal to Recall.

1. $F1\ Score\ (F1) = 2 * \frac{PR*RR}{PR+RR}$ (5)

4. Results and Analysis

In this section, we present and analyze the performance of the Logistic Regression (LR), Gradient Boosting Classifier (GBC), and Multi-layer Perceptron (MLP) models under FGSM and PGD adversarial evasion attacks. The models are evaluated using various performance metrics, including Accuracy, Precision, Recall, and F1 Score.

The performance of surrogate DNN model is evaluated first for simulating the peroper adversarial examples. It performed well and achieved detection of attacks with 98.2% accuracy, 96.5% precision rate, 98.9% recall rate, and 97.6% F1 score for the NSL-KDD dataset as presented in Table-1 for non-adversarial settings.

Table 1*: Performance comparison of DNN in the non-adversarial environment and under adversarial attack using the NSL-KDD dataset*

	Accuracy	Precision	Recall	F1 Score
Non-Adversarial Environment	98.2%	96.5%	98.9%	97.6%
Under Adversarial Attack (FGSM)	57.3%	35.6%	19.3%	25.6%
Under Adversarial Attack (PGD)	33.3%	25.6%	11.3%	15.6%

Now, the adversarial examples are generated using FGSM and PGD method by attacking this well trained surrogate model and impact of adversarial examples are evaluated.

4.1 Impact of Adversarial Attacks on ML Models Performance

The results are presented in Table-2, Table-3 and Table-4 for the performance of each model under no attack and two types of adversarial attacks: FGSM, PGD. The Gradient Boosting Classifier consistently demonstrated superior resilience compared to Logistic Regression and Multi-layer Perceptron across various metrics including 99% Accuracy, 98% Precision, 99% Recall, and 98% F1 Score. Specifically, under FGSM and PGD attacks, GBC maintained higher accuracy and precision, reflecting its robustness in distinguishing between benign and malicious network traffic despite adversarial perturbations. Conversely, LR and MLP showed a marked decline in performance under the same conditions, indicating their susceptibility to adversarial manipulation. But, Both the attacks are successful and have degraded the performance of all the classifiers.

Table 2: *Logistic Regression Performance*

	Accuracy	Precision	Recall	F1 Score
Non-Adversarial Environment	96.2%	96.2%	97.1%	96.6%
Under Adversarial Attack (FGSM)	51.1%	45.2%	48.6%	46.6%
Under Adversarial Attack (PGD)	40.2%	38.1%	42.5%	40.3%

Table 3: *Gradient Boosting Classifier Performance*

	Accuracy	Precision	Recall	F1 Score
Non-Adversarial Environment	99.3%	98.2%	99.1%	98.4%
Under Adversarial Attack (FGSM)	70.2%	72.5%	65.6%	67.3%
Under Adversarial Attack (PGD)	56.5%	63.1%	55.3%	59.6%

Table 4*: Multi-layer Perceptron Performance*

	Accuracy	Precision	Recall	F1 Score
Non-Adversarial Environment	98.3%	96.5%	98.8%	97.7%
Under Adversarial Attack (FGSM)	60.2%	68.6%	45.8%	54.3%
Under Adversarial Attack (PGD)	52.3%	55.4%	32.3%	40.6%

PGD generally produces more effective adversarial examples compared to FGSM as refletcetd in Figure-4, Fugure-5 and Figure-6. These findings highlight the importance of evaluating model robustness against adversarial attacks and the need for developing defense mechanisms to mitigate their impact on machine learning models used in Intrusion Detection Systems (IDS). Further research could explore advanced defense strategies and their effectiveness in real-world applications.

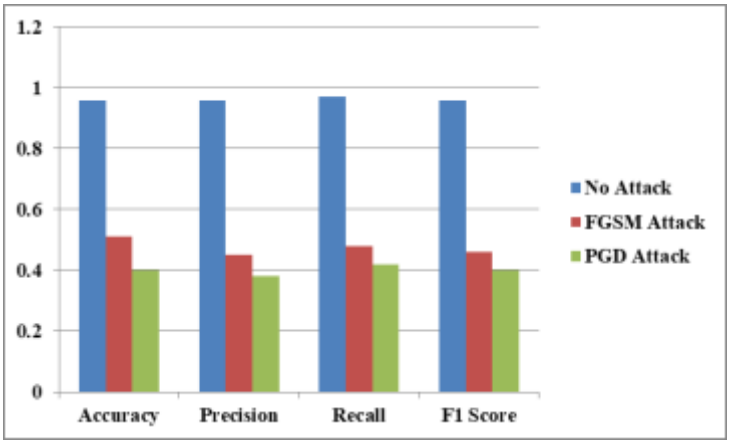

Figure 4: *Performance assessment of Logistic Regression*

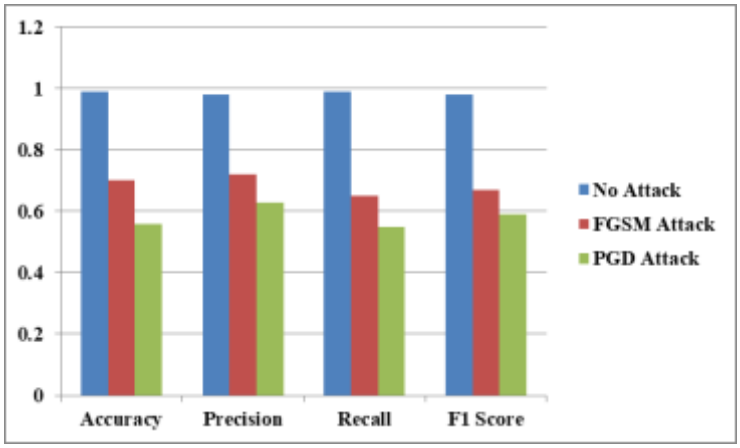

Figure 5: *Performance assessment of Gradient Boosting Classifier*

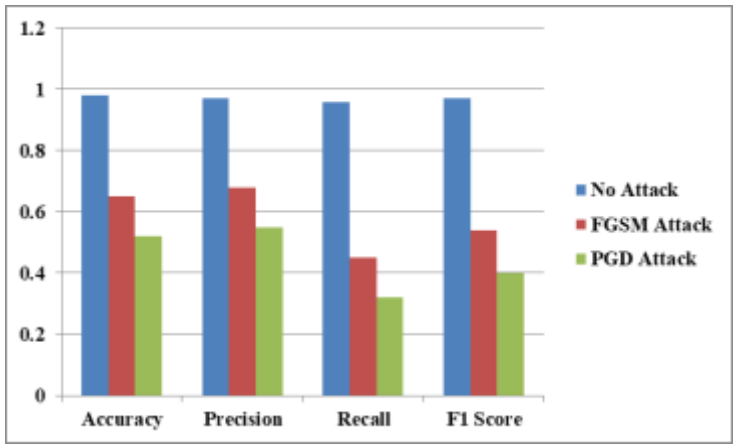

Figure 6: *Performance assessment of Multi-Layer Perceptron*

Logistic Regression shows a significant drop in performance under adversarial attacks, indicating higher vulnerability. Multi-layer Perceptron performs better than Logistic Regression under attacks, but still shows noticeable degradation. Gradient Boosting Classifier exhibits the highest resilience to adversarial attacks, maintaining relatively better performance compared to the other models.

From the results, it is evident that all models experience performance degradation under adversarial attacks, with Logistic Regression being the most vulnerable and Multi-layer Perceptron the most resilient. These findings highlight the need for robust defense mechanisms to mitigate the impact of adversarial attacks on ML-based IDS. Future work should explore advanced defense strategies and evaluate their effectiveness in real-world scenarios.

5. Conclusion

In this study, we conducted a comprehensive comparative analysis of three machine learning models, Logistic Regression (LR), Gradient Boosting Classifier (GBC), and Multi-layer Perceptron (MLP) under adversarial

attack scenarios generated using a gray-box threat model. The adversarial attacks were crafted using two prominent techniques: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Our findings reveal significant insights into the robustness and performance of these models when exposed to adversarial conditions. Under no attack conditions, all models performed excellently, with GBC slightly outperforming the others. However, the introduction of adversarial attacks unveiled critical vulnerabilities.
The Gradient Boosting Classifier consistently demonstrated superior resilience compared to Logistic Regression and Multi-layer Perceptron across various metrics including Accuracy, Precision, Recall, and F1 Score. Specifically, under FGSM and PGD attacks, GBC maintained higher accuracy and precision, reflecting its robustness in distinguishing between benign and malicious network traffic despite adversarial perturbations. Conversely, LR and MLP showed a marked decline in performance under the same conditions, indicating their susceptibility to adversarial manipulation. This comparative analysis underscores the necessity for robust defense mechanisms in Intrusion Detection Systems (IDS) to counter sophisticated adversarial attacks. The superior performance of GBC in adversarial scenarios highlights its potential as a more reliable model for IDS applications. Future research should focus on enhancing the robustness of ML models, particularly LR and MLP, and exploring advanced defense strategies to mitigate the impact of adversarial attacks. This will be crucial in developing resilient and reliable IDS capable of maintaining high performance in adversarial environments.

References

1. P.Sangkatsanee, N. Wattanapongsakorn and C. Charnsripinyo, “Practical Real-Time Intrusion Detection Using Machine Learning Approaches, Computer Communications” , vol. 34, no. 18, pp. 2227–2235, (2011).
2. T. Shon, Y. Kim, C. Lee, and J. Moon, “A machine learning framework for network anomaly detection using SVM and GA,” in Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop, Jun. 2005, pp. 176–183. doi: 10.1109/IAW.2005.1495950.
3. B. Biggio et al., “Evasion Attacks against Machine Learning at Test Time,” in Advanced Information Systems Engineering, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds., in Lecture Notes in Computer Science, vol. 7908. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402. doi: 10.1007/978-3-642-40994-3_25.
4. A. T. Yousef El Mourabit, AnouarBouirden and N. E. Moussaid, “Intrusion Detection Techniques in Wireless Sensor Network Using Data Mining Algorithms: Comparative Evaluation Based on Attacks Detection”, International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 164–172, (2015).
5. J. Manjula C. Belavagi and Balachandra Muniyal, “Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection” Twelfth International Multi-Conference on Information Processing- 2016.
6. Wang, Y.: A multinomial logistic regression modeling approach for anomaly intrusion detection. Comput. Secur. 24(8), 662–674 (2005)
7. 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.
8. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, “SoK: Security and privacy in machine learning,” in Proc. IEEE Eur. Symp. Security Privacy, London, U.K., Apr. 2018, pp. 399–414.
9. J. Gardiner and S. Nagaraja, “On the security of machine learning in malware C&C detection: A survey,” ACM Comput. Surveys, vol. 49, no. 3, p. 59, 2016.
10. Hu, W., & Tan, Y. (2017). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN.
11. Biggio B., I. Corona, D. Maiorca, B. Nelson, N. ˇSrndic, P. Laskov, G. Giacinto, and F. Roli, (2013). Evasion attacks against machine learning at test time. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 387–402.
12. Gaikwad, N., Sangve, S. (2016). A Novel Model for NIDS with Evaluation of Pattern Classifiers and Facility of Rectification. In: Unal, A., Nayak, M., Mishra, D.K., Singh, D., Joshi, A. (eds) Smart Trends in Information Technology and Computer Communications. SmartCom 2016. Communications in Computer and Information Science, vol 628. Springer, Singapore.
13. J. S. Sadioura, S. Singh and A. Das, (2019) "Selection of sub-optimal feature set of network data to implement Machine Learning models to develop an efficient NIDS," International Conference on Data Science and Engineering (ICDSE), Patna, India, 2019, pp. 120-125.
14. O. Almomani, M. A. Almaiah, A. Alsaaidah, S. Smadi, A. H. Mohammad and A. Althunibat, "Machine Learning Classifiers for Network Intrusion Detection System: Comparative Study," 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp. 440-445.
15. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, New York, NY, USA: ACM, 2018, pp. 6638-6648.

16. P. Shettar, A. V. Kachavimath, M. M. Mulla, N. D. G and G. Hanchinmani, "Intrusion Detection System using MLP and Chaotic Neural Networks," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-4.
17. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.
18. M. S. Ayas, S. Ayas and S. M. Djouadi, "Projected Gradient Descent Adversarial Attack and Its Defense on a Fault Diagnosis System," 2022 45th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2022, pp. 36-39.
19. M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
20. Goodfellow I., Bengio Y., and Courville A., "Deep learning", MIT press, 2016.