Towards Evaluating the Robustness of Deep Intrusion Detection Models in Adversarial Environment

Sriram S[1], Simran K[1], Vinayakumar R[2,1], Akarsh S[1], and Soman KP[1]

[1] Center for Computational Engineering and Networking, Amrita School Of Engineering, Amrita vishwa vidyapeetham, Coimbatore, India.
sri27395ram@gmail.com, simiketha19@gmail.com

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States.
Vinayakumar.Ravi@cchmc.org, vinayakumarr77@gmail.com

Abstract. Network Intrusion Detection System (NIDS) is a method that is utilized to categorize network traffic as malicious or normal. Anomaly-based method and signature-based method are the traditional approaches used for network intrusion detection. The signature-based approach can only detect familiar attacks whereas the anomaly-based approach shows promising results in detecting new unknown attacks. Machine Learning (ML) based approaches have been studied in the past for anomaly-based NIDS. In recent years, the Deep Learning (DL) algorithms have been widely utilized for intrusion detection due to its capability to obtain optimal feature representation automatically. Even though DL based approaches improves the accuracy of the detection tremendously, they are prone to adversarial attacks. The attackers can trick the model to wrongly classify the adversarial samples into a particular target class. In this paper, the performance analysis of several ML and DL models are carried out for intrusion detection in both adversarial and non-adversarial environment. The models are trained on the NSLKDD dataset which contains a total of 148,517 data points. The robustness of several models against adversarial samples is studied.

Keywords: Intrusion detection · Deep learning · Machine learning · Cyber Security · Adversarial attacks

1 Introduction

In today's world, cyber-attacks and threats on Information and Communication Technologies (ICT) systems are growing rapidly. Various new attacks are invented daily by attackers to bypass the current security systems and steal crucial information. To detect and prevent these attacks on ICT systems, we need flexible and reliable integrated network security solutions. Various security structures and methods are used to deal with these malicious attacks namely firewalls, Intrusion Detection System (IDS), software updates, encryption and

decryption methods, etc. In that, IDS plays a big role in defending the network from all kinds of intrusion and malicious acts, both from outside and inside the network. IDS has been actively studied area from the 1980s, a seminal work by [1] on the computer security threat monitoring and surveillance. IDS is mainly categorized into two types. One is Network IDS (NIDS): It is utilized to monitor and analyze network traffic records to safeguard a system from network-based attacks. The next type is Host-based IDS (HIDS): it monitors the system in which it is installed to detect both internal and external intrusion and misuse and it responds by recording the activities and alerts the authority. NIDS monitors the network traffic and classifies the network records between normal ones and malicious ones. Since this is a classification problem, various Machine Learning (ML) and Deep Learning (DL) models are widely used in these detection systems and have achieved good results. However, ML and DL models are prone to adversarial attacks. Attackers can fool the detection system by using adversarial samples and make the classifier misclassify those sample data [2]. Therefore, it is necessary to check the robustness of those models that are used in NIDS against adversarial samples. In this paper, Several DL and ML models are trained on the openly available NSLKDD dataset for IDS. The robustness of those models against adversarial samples is studied. The main contributions of this work are the following:

- We have trained several DL and ML models using NSLKDD dataset in a non-adversarial environment and reported their performance using standard metrics.
- We have also studied the robustness of the trained models in the adversarial environment using the samples generated by two different adversarial attack techniques.

The rest of the paper is arranged as follows. Section 2 presents the related works. Section 3 includes the background information. Section 4 and 5 presents description of the dataset and statistical measures respectively. Section 6 and 7 covers the experimental results and conclusion.

2 Related Work

Many ML and DL based approaches have been applied for various problems in the field of cyber security including IDS [3–7]. The authors Tsai et al. utilized Support Vector Machine (SVM), Self-organizing maps, Artificial Neural Networks (ANN), Naive Bayes (NB), K-Nearest Neighbor (KNN), Genetic algorithms, Decision Tree (DT), Fuzzy logic, etc for detecting the intrusion [8]. Buczak and Guven have done a comprehensive survey [9] on ML-based NIDS where many ML classifiers such as DT, ensemble learning, SVM, clustering, Hidden Markov Models (HMM), NB, etc. Since ML techniques require manual features, DL based approaches are proposed. DL architectures can obtain salient features from the input data automatically. In [10], the authors have proposed multiple Deep Neural Network (DNN) models for both network and host-based

intrusion detection. They have trained models using several benchmark datasets and compared its performance with ML-based approaches. Similar to [10], [11] proposes a DNN based IDS for Software Defined Networking (SDN) environment. The proposed model only takes 6 basic features from 41 features of the NSLKDD dataset. [12] studies the effectiveness of DL networks such as DNN, Convolutional Neural Network (CNN), and Hybrid CNN for binary and multi-class classification. [13] compares the performance of many shallow and deep neural networks in detecting intrusion and [14] proposes a recurrent neural network and its variants for intrusion detection.

ML and DL models are prone to adversarial attacks. This vulnerability, which was discovered in recent years, limits the application of ML and DL models in various security-critical areas like IDS, autonomous vehicles, health care, etc. The authors Szegedy et al experimented on AlexNet with some adversarial sample images [15]. AlexNet [16] is the name of a convolutional neural network, designed by Alex Krizhevsky. They showed that by making very small variations in the input image, they could make the model misclassify it. Since then, the profound implications of this vulnerability sparked several researchers to develop various adversarial attacks and defenses. Some of the most commonly known attacks are Jacobian based Saliency Map Attack (JSMA) [17] and Fast Gradient Sign Method (FGSM) [18]. In this paper, the effects of adversarial samples generated by [18] and [17] on various ML and DL models are studied.

3 Background

3.1 Adversarial attacks

Fast Gradient Sign Method (FGSM): It is a straightforward method of creating adversarial samples, which was proposed by Goodfellow et al. In FSGM, a small deviation is calculated in the direction of the gradient and it is defined as follows.

$$p = \epsilon sign(\bigtriangledown_x L(\theta, x, y)) \tag{1}$$

where p is the perturbation, ϵ is a small constant, $\bigtriangledown x L(\theta, x, y)$ is the gradient of loss function L which is used for training the model, θ denotes the model, x denotes the input and y denotes the class of input x. This perturbation p is added to the input data to generate adversarial samples:

$$x^{adversarial} = x + p \tag{2}$$

FGSM is computationally more efficient when compared to JSMA. But it has a lower rate of success.

Jacobian-based Saliency Map Attack (JSMA) It uses the concept of saliency maps to generate adversarial samples. A saliency map gives insights about the features of the input data that are most likely to create a change of

targeted class. In other words, saliency maps rate each feature of how influential it is for causing the model to predict a target class. JSMA causes the model to misclassify the resulting adversarial sample to a specific erroneous target class by modifying the high-saliency features. The formulation of the saliency map is given as:

$$A^{+}(x_{(i)}, y) = \begin{cases} 0 \; if \frac{\partial f(x)_{(y)}}{\partial x_{(i)}} < 0 \; or \sum_{y' \neq y} \frac{\partial f(x)_{(y')}}{\partial x_{(i)}} > 0 \\ -\frac{\partial f(x)_{(y)}}{\partial x_{(i)}} \cdot \sum_{y' \neq y} \frac{\partial f(x)_{(y')}}{\partial x_{(i)}} \; otherwise \end{cases} \tag{3}$$

Where $x_{(i)}$ is input feature, y is a class, and $A^{+}(.)$ is the measure of positive correlation of $x_{(i)}$ with class y and negative correlation of $x_{(i)}$ with all other classes. If both cases in the formulation fail, then the saliency is zero. JSMA can create adversarial samples with less degree of distortion and has a better success rate while compared to FGSM.

3.2 Intrusion Detection System (IDS)

IDS is a tool that deals with unauthorized access and threats to systems and information by any type of user or software. Intrusion can be external or internal. External intrusion is when an intruder tries to gain access to a protected internal network. Internal intrusion is when an insider with a motive tries to misuse, attack or steal information. This is also called an insider threat. Two major categories of IDS are HIDS and NIDS. HIDS is a tool that monitors the system in which it is installed to detect both external and internal intrusion, misuse and responds by recording activities and alerts the authority. NIDS is utilized to monitor and analyze network traffic to safeguard a system from network-based attacks. Figure 1 shows a model of Intrusion detection system. Signature-based NIDS uses signatures that are extracted from previously known attacks. Signatures are manually generated and stored in the database whenever a new attack is identified. New attacks will not be detected by this system. Anomaly-based NIDS models the normal behavior of the network and raises alarm whenever it detects an anomalous behavior. Hybrid NIDS uses the combination of the above two approaches.

3.3 Deep Learning (DL) Models

The DL models are used for solving various research problems in a wide range of fields like biomedical, speech processing, natural language processing, etc since DL models have the capability of extracting salient features automatically with very less or no human intervention. The Deep Neural Network (DNN) model used in work has 5 hidden layers and overall it has a total of 1,399,557 trainable parameters. These five layers have 1024, 768, 512, 256, 128 neurons respectively. The dropout regularization technique is also employed to avoid overfitting.

The Convolutional Neural Network (CNN) model is widely used in the area of computer vision as it is capable of extracting location invariant features automatically. The CNN model, which is used in this work, has four convolution

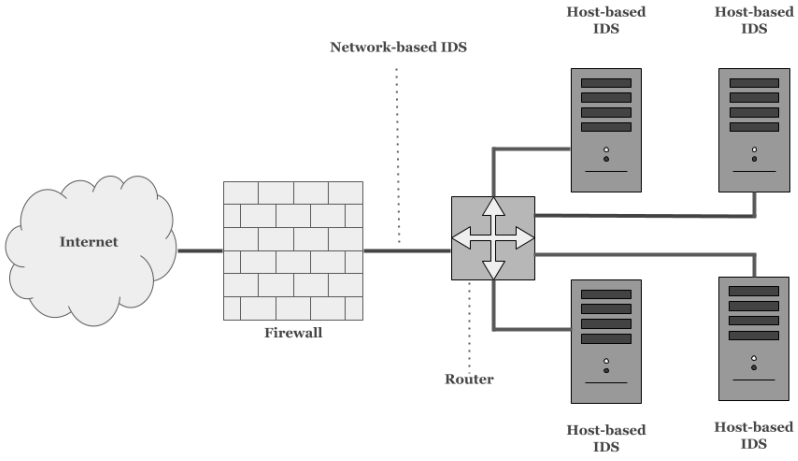

Fig. 1. Intrusion detection system model.

layers followed by a fully connected layer of 128 neurons. The CNN model has a total of 251,205 trainable parameters whereas the Long Short-Term Memory (LSTM) model, which is also used in this work has 1,26,533 trainable parameters.

4 Description of Dataset

One of the most used datasets is KDDCUP 99 which was obtained from the DARPA98 dataset. The KDDCUP 99 dataset has several issues that are resolved by a newly refined version called NSL-KDD [19]. In this dataset, the invalid and redundant connection records are omitted from the entire train and test data. Table 1 represents the statistics of the NSLKDD dataset. This dataset has various attacks that belong to four major families such as User to Root (U2R), Probing attacks, Denial of Service (DoS) and Remote to Local (R2L). The purpose of the DoS attack is to work against resource availability. U2R attacks represent attempts for privilege escalation. R2L attacks attempt to exploit a vulnerability and gain remote access to a machine. Probe attacks are mainly information gathering attempts by scanning parts of the networks. The dataset contains a total of 41 features.

5 Statistical Measures

The performance evaluation of the models against adversarial attacks is conducted based on some of the popular performance metrics such as precision, accuracy, f1-score, and recall. Accuracy gives an oversight of the performance of the classifier. F1-score gives the harmonic mean between recall and precision. In

Table 1. Statistics of NSLKDD data set.

Attack Types	Description	NSLKDD (10% of Data)	
		Train	Test
Normal	Normal connection records	67,343	9,710
DoS	Attacker aims at making network resources down	45,927	7,458
Probe	Obtaining detailed statistics of system and network configuration details	11,656	2,422
R2L	Illegal access originated from remote computer	995	2,887
U2R	Obtaining the root or superuser access on a particular computer	52	67
Total		125,973	22,544

a binary classification setting, true labels versus the predicted labels are represented by confusion matrix and the matrix contains four terms. The first one is True Positive (TP). It denotes the amount of malicious traffic records that are correctly predicted as malicious. The second one is False Positive (FP). It denotes the amount of normal traffic records that are incorrectly predicted as malicious. The next one is True Negative (TN) and it denotes the amount of normal traffic records that are correctly predicted as normal. The final one is False Negative (FN) and it denotes the amount of malicious traffic records that are incorrectly predicted as normal. Based on these four terms, we can define several metrics:

- **Accuracy:** It denotes the total amount of correct predictions (TP and TN) over the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

- **Precision:** It denotes the amount of correct positive results over the amount of positive results predicted by the model.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

- **Recall:** It denotes the total amount of correct positive results over the amount of all relevant samples.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

- **F1 score:** F1 score denotes the harmonic mean between recall and precision.

$$F1score = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

The adversarial attacks reduce the overall performance of the model by tricking it to perform misclassification. Therefore, the above-mentioned metrics which show the performance of the system can be used to measure the robustness of the model in the adversarial environment.

6 Experimental results

The adversarial attacks such as FGSM and JSMA are implemented using Adversarial Robustness Toolbox v0.10.0 [8] and the ML and DL models are implemented using Scikit-Learn and Keras python libraries respectively. The models implemented Table 2 represents the performance of models such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) and Deep Neural Network (DNN), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Adaboost (AB) in non-adversarial environment. The performance of the trained models is compared with the performance of the Soft-Max Regression (SMR) classifier [20].

Table 2. Performance of baseline models for test set.

ML Model	Accuracy	Precision	Recall	F1-score
DNN	77.39	78.36	77.39	75.80
CNN	75.37	80.61	75.37	71.88
LSTM	74.65	71.73	74.65	70.01
SMR [20]	75.23	86.71	62.30	72.14
LR	63.32	55.88	63.32	57.07
NB	44.41	63.22	44.41	48.29
KNN	73.50	74.13	73.50	70.02
DT	74.78	74.58	74.78	71.95
AB	43.12	51.08	43.12	45.84
RF	73.84	81.28	73.84	69.33
Linear-SVM	66.51	68.20	66.51	61.59
RBF-SVM	64.71	60.13	64.71	59.08

It can be observed from Table 2 that the DNN performed better than all the other models that are trained in this work. Based on the accuracy metric, the DNN, CNN, and DT are the top three models that are trained in this work and their accuracies are 77.39%, 75.37%, and 74.78%. Adaboost classifier gives the least performance in terms of accuracy. In terms of F1-score, both SMR and DT models performed better than CNN and LSTM models. All the models that are trained in this work are also tested on adversarial samples generated by FGSM and JSMA to evaluate how robust they are under an adversarial environment. The Table 3 and Table 4 represents the performance of all the models tested on adversarial samples generated by FGSM and JSMA methods

respectively. It can be observed from both the tables that the adversarial attacks tremendously reduced the performance of the baseline models that are trained in a non-adversarial environment.

Table 3. Performance of models for the adversarial sample generated by FGSM.

ML Model	Accuracy	Precision	Recall	F1-score
DNN	16.74	30.62	16.74	16.02
CNN	37.83	35.82	37.83	35.82
LSTM	24.51	32.47	24.51	25.36
LR	62.27	54.81	62.27	55.43
NB	33.76	22.64	33.76	25.31
KNN	66.35	61.79	66.35	61.47
DT	17.65	23.85	17.65	17.49
AB	17.28	19.71	17.28	14.74
RF	39.97	29.88	39.97	30.81
Linear-SVM	63.25	56.98	63.25	57.32
RBF-SVM	63.05	56.79	63.05	56.18

Table 4. Performance of models for the adversarial sample generated by JSMA.

ML Model	Accuracy	Precision	Recall	F1-score
DNN	10.87	3.93	10.87	3.05
CNN	10.06	24.24	10.06	7.44
LSTM	46.49	45.44	46.49	0.33
LR	14.38	25.84	14.38	1
NB	43.61	29.39	43.61	30.58
KNN	49.27	47.95	49.27	38.50
DT	12.21	38.41	12.21	6.98
AB	3.88	20.12	3.88	6.05
RF	14.18	47.63	14.18	10.95
Linear-SVM	46.60	38.54	46.60	34.04
RBF-SVM	62.01	54.59	62.01	54.11

The performance of the models is affected tremendously by both FGSM and JSMA techniques. The top three most affected models by FGSM in terms of accuracy are DNN, LSTM, and DT. The FGSM attack reduced the performance of DNN from 77.39 to 16.74 (78% reduction), LSTM from 74.65 to 24.51 (76% reduction), and DT from 74.78 to 17.65 (67% reduction). The least affected models by FGSM attack is RBF-SVM (2% reduction), LR (2% reduction), and LSVM (4% reduction). The top three most affected models by JSMA in terms of accuracy are CNN, DNN, and DT. The JSMA attack reduced the performance of CNN from 74.65 to 10.06 (87% reduction), DNN from 77.39 to 10.87 (86% reduction), and DT from 74.78 to 12.21 (83% reduction). The least affected models by JSMA attack are NB (2% reduction), RBF-SVM (4% reduction), and LSVM (30% reduction).

It can be observed from both the tables that, FGSM worked well in the case of LSTM and NB and JSMA worked better than FGSM in all other cases. RBF-SVM, LSVM, KNN, and NB are the models which show more robustness against both adversarial attacks when compared to the rest of the models. The adversarial samples that are created using the DNN model generalize well over other DL and ML models as well. In other words, the attack samples, which are created by both FGSM and JSMA for the DNN model as the target, also affect the performance of other ML and DL models.

7 Conclusion

In this paper, we have observed that the adversarial samples can lower the accuracy of many DL and ML classifiers with varying degrees of success. This shows that it is necessary to test the robustness of any DL or ML model against adversarial samples especially when they are used in security-critical applications. In this paper, the models that are trained did not perform well when compared to other state-of-the-art approaches, but its robustness towards adversarial attacks are studied. In the future, we will further focus on the defense techniques that avoid such attacks.

Acknowledgement

This research was supported in part by Paramount Computer Systems and Lakhshya Cyber Security Labs. Also, the authors would like to express gratitude to NVIDIA India for supporting the research by providing the GPU hardware. They would also like to express gratitude to Computational Engineering and Networking (CEN) department for encouraging the research.

References

1. Anderson, J. P. (1980). Computer security threat monitoring and surveillance, James P.Anderson Co., Fort Washington, PA.

2. Rigaki, M. (2017). Adversarial deep learning against intrusion detection classifiers.
3. Vinayakumar, R., Alazab, M., Srinivasan, S., Pham, Q. V., Padannayil, S. K., & Simran, K. (2020). A Visualized Botnet Detection System based Deep Learning for the Internet of Things Networks of Smart Cities. IEEE Transactions on Industry Applications.
4. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019). Robust intelligent malware detection using deep learning. IEEE Access, 7, 46717-46738.
5. Vinayakumar, R., & Soman, K. P. (2020). Siamese neural network architecture for homoglyph attacks detection. ICT Express, 6(1), 16-19.
6. Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2019). A Comparative Analysis of Deep Learning Approaches for Network Intrusion Detection Systems (N-IDSs): Deep Learning for N-IDSs. International Journal of Digital Crime and Forensics (IJDCF), 11(3), 65-89.
7. Vinayakumar, R., Alazab, M., Jolfaei, A., Soman, K. P., & Poornachandran, P. (2019, May). Ransomware triage using deep learning: twitter as a case study. In 2019 Cybersecurity and Cyberforensics Conference (CCC) (pp. 67-73). IEEE.
8. Chih-Fong Tsai et al. Intrusion detection by machine learning: A review. In:Expert Systems with Applications 36 (Dec.2009), pp. 1199412000.DOI: 10 . 1016 / j .eswa.2009.05.029.
9. A. L. Buczak and E. Guven. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. In:IEEE Communications Surveys Tutorials 18.2 (Second quarter 2016),pp. 11531176.ISSN: 1553-877X.DOI: 10 .1109/COMST.2015.2494502.
10. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. IEEE Access, 7, 41525-41550.
11. Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2016, October). Deep learning approach for network intrusion detection in software defined networking. In 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 258-263). IEEE.
12. Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017, September). Applying convolutional neural network for network intrusion detection. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1222-1228). IEEE.
13. Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017, September). Evaluating effectiveness of shallow and deep networks to intrusion detection system. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1282-1289). IEEE.
14. Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017). Evaluation of recurrent neural network and its variants for intrusion detection system (IDS). International Journal of Information System Modeling and Design (IJISMD), 8(3), 43-63.
15. Christian Szegedy et al. Intriguing properties of neural networks. In:arXiv preprint arXiv:1312.6199 (2013).
16. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In:Advances in neural information processing systems. 2012, pp. 10971105.
17. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. and Swami, A. (2016), The limitations of deep learning in adversarial settings, in Security and Privacy (EuroS & P), 2016 IEEE European Symposium on, IEEE, pp. 372.387.

18. Goodfellow, I. J., Shlens, J. and Szegedy, C. (2014), Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 .
19. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1-6). IEEE.
20. Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016, May). A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) (pp. 21-26). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).