I.J. Wireless and Microwave Technologies, **2023, 4, 1-10**
Published Online on August 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijwmt.2023.04.01

Adversarial Deep Learning in Anomaly based Intrusion Detection Systems for IoT Environments

Khalid Albulayhi
Computer science department, Technical and Vocational Training Corporation (TVTC), Buraydah, 51452, Saudi Arabia
E-mail: kalbulayhi@tvtc.gov.sa
ORCID iD: https://orcid.org/0000-0001-6084-2283

Qasem Abu Al-Haija*
Department of Cybersecurity, Princess Sumaya University of Technology, Amman 1196, Jordan
Email: qabualhaija@psut.edu.jo
ORCID iD: https://orcid.org/0000-0003-2422-0297
*Corresponding Author

Received: 20 February, 2023; Revised: 15 April, 2023; Accepted: 20 May, 2023; Published: 08 August, 2023

Abstract: Using deep learning networks, anomaly detection systems have seen better performance and precision. However, adversarial examples render deep learning-based anomaly detection systems insecure since attackers can fool them, increasing the attack success rate. Therefore, improving anomaly systems' robustness against adversarial attacks is imperative. This paper tests adversarial examples against three anomaly detection models based on Convolutional Neural Network (CNN), Long Short-term Memory (LSTM), and Deep Belief Network (DBN). It assesses the susceptibility of current datasets (in particular, UNSW-NB15 and Bot-IoT datasets) that represent the contemporary network environment. The result demonstrates the viability of the attacks for both datasets where adversarial samples diminished the overall performance of detection. The result of DL Algorithms gave different results against the adversarial samples in both our datasets. The DBN gave the best performance on the UNSW dataset.

Index Terms: Adversarial, Deep learning, Anomaly detection, Internet of Things.

1. Introduction

The use of deep learning models in the anomaly detection system has increased exponentially [1]. The use of traditional machine learning in anomaly detection systems has been rampant and has provided relatively improved performance. Such traditional machine learning-based anomaly detection systems suffer from major drawbacks, including major drawbacks, and lower detection performance when processing big datasets. This fact renders them inadequate for environments using cutting-edge technologies producing a huge amount of data, particularly smart environments. Compared to traditional machine learning networks, deep learning networks perform better, can represent high-dimensional spatial features, and learn the intrinsic features without handcrafted feature engineering [2-4]. However, deep learning networks are vulnerable to adversarial examples [5]. Adversarial attacks use slightly perturbed inputs, which can mislead the most advanced deep learning networks to produce wrong decisions. Decisions. The adversarial attack usually occurs when DL/ML algorithms get a small change in input. An example of an adversarial attack is when some features are changed to confuse the model to produce a false prediction.

This study's primary contribution and objective is the evaluation of the efficacy of adversarial deep learning attacks on modern datasets that represent the current networking and computing environment. This paper evaluates various intrusion detection datasets, such as UNSW-NB15 and Bot-IoT, to demonstrate their susceptibility to common adversarial attack techniques, such as Jacobian Saliency Map Attack (JSMA), Fast Gradient Method (FGSM), and Carlini Wagner Attack (CW). Comparing and analyzing the impact of the attacks on the two datasets was accomplished by examining several metrics, including precision, accuracy, F1 score, and recall, using various deep learning models.

The remaining sections of the paper are organized as follows: Section 2 provides the necessary background information to establish the research context. Section 3 delves into the existing research on adversarial sample production, specifically in Intrusion Detection System (IDS) models. Section 4 presents the experimental evaluation process employed in the study, encompassing the methodology, dataset selection, and experimental setup. Following this, Section 5 showcases the results and their evaluation, providing quantitative and qualitative analysis of the findings.

Section 6 offers an in-depth discussion of the experiment's findings, examining patterns, trends, and implications. Finally, Section 7 concludes the paper, summarizing the key findings and their significance while highlighting potential avenues for future research.

2. Background Study

This section introduces the background of the research, including Internet of Things security and intrusion detection systems and their weakness against adversarial attacks.

2.1. Internet of Things Security

The IoT is a wide range of heterogeneous devices that seamlessly interconnect physical objects to the information network. This limited heterogeneity advancement toward obtaining internationally agreed-upon standards, particularly for security. Security standards are needed to overcome the major obstacles to implementing IoT devices in private and public sectors. IoT device targeting attacks are in remarkable increase [6]. In the profit-driven business sector, IoT security is often considered an afterthought. For example, attacks on private data generated by IoT devices are considered a high-impact risk. In the healthcare sector, successful attacks may go up to life-threatening. Manipulating data from a body area network device monitoring a user's health may cause alert deviation [6]. In other sectors, since data generated from IoT devices usually sent over the internet may include confidential user data, privacy has become a prime concern. Thus, the domain researchers have intensified work on improving IoT security posture [6]. This has necessitated the use of traditional security mechanisms. The top first line of defense mechanisms proposed to secure sensitive data in IoT environments include Firewalls, authentication schemes, different encryption methods, and antiviruses. IDSs systems are second-line defense mechanisms meant to secure IoT networks from attacks. The IDS systems in IoT are discussed in the next subsection [1, 6].

2.2. Anomaly Detection Systems

IDS system is one of the main approaches used to protect networks. The deployment and detection mechanisms are used for the classification of IDS systems. Based on their deployment, IDS systems can be host-based (HIDS) or network-based (NIDS) [6]. This is based on where the sensor is placed in analyzing information. The HIDS keeps track of a single host's activity, including running processes, system calls, and memory code. The NIDS keeps track of network traffic by inspecting packets for protocol usage, services, and communication partners' IP addresses [6]. NIDS are frequently deployed at the organization's network perimeter at the gateway to inspect ingress and egress traffic [1, 6]. On the other hand, using their detection mechanisms, IDS systems can be classified into three major groups, including signature [7], anomaly [1], and hybrid based [8]. The signature-based IDS systems generate alerts based on a list of pre-determined knowledge, such as hashes, byte sequences, and so on. Anomaly-based IDS systems depend on the principle of the existing normal and anomalous activities [1]. The system triggers alerts when anomalous activity is detected. Anomaly-based IDS systems can detect zero-day attacks that the public has not previously experienced. The hybrid IDS systems employ signature and anomaly-based IDS systems [1, 3, 4]. The primary issue with existing IDSs is the increase in the False Alarm Rate (FAR) for zero-day abnormalities detection [6]. Recently, researchers investigated the possibilities of improving detection accuracy and lowering the FAR for NIDS by utilizing machine learning and deep learning methodologies. Both machine learning and deep learning approaches were effective tools for extracting important patterns from network data and classifying them as anomalous or benign [1]. Due to its architecture, deep learning is deemed efficient compared to traditional machine learning in extracting significant features from raw data without human intervention [1, 3, 9]. The next section discusses more deep learning methods used in IDS systems.

2.3. Deep Learning

Researchers have successfully employed machine learning IDS models to support anomaly-based IDS in reducing the amount of FPRs. Since anomaly-based IDS systems do not rely on static knowledge, they can handle the ever-growing IoT devices and countless zero-day attacks. The increase in the False Alarm Rate (FAR) in detecting zero-day anomalies is the fundamental issue with contemporary IDSs. Researchers have lately investigated the possibilities of improving detection accuracy and lowering the FAR for NIDS by applying machine learning (ML) and deep learning (DL) technologies. ML and DL approaches have been proven to be effective tools for learning valuable patterns from network data and classifying flows as anomalous or benign [1, 9]. ML relies entirely on substantial feature engineering to determine which data attributes will be considered by the model and contribute to its training and classification. The model's performance will vary drastically depending on which features are used. In the context of IoT IDS [9], this reliance on feature engineering and its effects on the model's performance has been established. Unfortunately, considering the intricacy of the subject, feature selection considerations in IoT IDS systems are highly complex [9]. Each feature has its advantages and disadvantages, depending on the selected features. DL outperforms ML in terms of performance. This is because of the large number of parameters that must be calibrated during training before the input data can be classified. Furthermore, due to its design architecture and lack of human interaction, the DL has efficiently learned significant features from raw data, making it important for use in IoT IDS systems. DL is effective in practice since each model is trained for a narrow domain with no requirement for generalization. DL models have a complex structure with hundreds, if not millions, of parameters. Such a huge number of parameters could lead to overfitting. There are several DL networks. To mention the three networks covered in this paper include the Convolutional Neural Network (CNN) [9], Long Short-Term Memory (LSTM) [6], and Deep Belief Network (DBN) [1]. CNN [9] is a DL type meant to interpret unstructured data. Convolutional networks are simple neural networks with at least one layer that uses convolution instead of ordinary matrix multiplication. CNN networks are widely employed in computer vision and natural language processing. It makes use of the mathematical convolution procedure. CNN networks largely consist of input, convolutional, top pooling, and fully connected layers. LSTM [6] is another Recurrent Neural Networks (RNN) variant familiar with voice, handwriting, and signature recognition. LSTM is developed to overcome the issue of gradient disappearance realized in RNN. LSMT is also popular in the intrusion detection and prevention domain. So, LSTM networks consist of a cell, an input gate, a forget gate, and an output gate. LSTM networks use a cell state, a memory cell, to maintain its secular state and hold values for unpredictable periods. On the other hand, the three gates manage the knowledge flow inside and outside the cell.

2.4. Adversarial Attacks

Studies on the peculiar qualities of DL have surged in recent years. The idea that a carefully constructed feature vector can deceive classifiers that perform better than humans on a benchmark dataset has captivated the AI research community. As knowledge of the issue has increased, numerous weak points have been identified. Adversarial examples are samples that appear to be correctly classifiable data but contain a minor, deliberate, worst-case deviation that can cause a variety of DL techniques to fail [10, 11]. The researchers discovered a way for rapidly and reliably producing adversarial examples that cause a variety of DL approaches to misclassify. Initial demonstrations of the approach were conducted on several datasets. It depends on discovering a small adversarial noise vector whose sum matches the sign of the elements of the gradient of the cost function for the assessed sample. The Fast Gradient Sign Method (FGSM) [12] is defined as the linearization of the cost function around the current value Θ, yielding an optimal max-norm constrained perturbation of the gradient.

$$\eta = \epsilon sign\big(\nabla xJ(\Theta, x, y)\big) \tag{1}$$

Where Θ represents the model's parameters, x represents the model's inputs, y represents the corresponding outputs, and $J(\Theta, x, y)$ is the cost used to train the deep learning model. The technique is known as FGSM. Researchers in the domain augmented the FGSM approach by applying it numerous times with a small step size, clipping the values after each transitional pace, and utilizing $\alpha = 1$, corresponding to changing the value of each element (pixel) by 1. In their work, the researchers have chosen a sufficient number of iterations to approach the edge of the ϵ max-norm ball. The following formula is employed:

$$X_0^{adv} = X_1, X_{N+1}^{adv} = Clip_X, \epsilon\left\{X_N^{adv} + \alpha sign\left(\nabla_x j(X_N^{adv}, y_{true})\right)\right\} \tag{2}$$

The method was later called Basic Iterative Method (BIM). Subsequently, Carlini and Wagner (C&W) in [13] provide a solution for developing adversarial examples by phrasing the optimization problem in a manner that existing methods can handle. Formally, the optimization issue is defined as

$$Min\ D(x, x + \delta)\ such\ that\ C(x + \delta) = t,\ x + \delta \in [0, 1]\, n \tag{3}$$

While x remains constant, the objective is to discover δ that minimizes $(x, x + \delta)$. In other words, discovering that modifies the categorization. D is a distance measure; in their work, the authors analyze three distance metrics; nevertheless, this paper $L2$ was chosen. To make the formula solvable, the authors redefine an objective function f in such a way that $C(x + \delta) = t$provides various formulas f. The attack, as defined in [13], was utilized.

Jacobian Based Saliency Map (JSMA) was introduced by [14]. This attack minimizes the $L0$ norm by iteratively constructing a saliency map and then perturbing the feature with the greatest impact. The technique consists of creating the Jacobian matrix where i is the input and j is a class derivative for input [14]:

$$J_F(x) = \frac{\partial F(x)}{\partial(x)} = \left[\frac{\partial j(x)}{\partial x_i}\right]_{ixj} \tag{4}$$

Where F stands for the second-to-last layer, the perturbation is chosen, and the procedure is repeated until misclassification in the target class is accomplished or the maximum number of perturbed features parameter is reached [14]. If it fails, the algorithm adds the next feature to the altered sample.

3. Related Work

Debicha et al. [5] investigated the impact of adversarial attacks on IDSs based on deep learning. The NSL-KDD dataset evaluated three adversarial attacks, FGSM, Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). Their findings demonstrate that adversarial examples can mislead the detector. Khamis and Matrawy [2] studied the efficacy of various evasion attacks and the feasibility of training a resilient deep learning-based intrusion detection system (IDS) utilizing convolutional neural networks (CNN) and recurrent neural networks (RNN). Using the UNSW-NB15 and NSL-KDD datasets, a robust IDS is trained against adversarial samples using the min-max method. Experiments on deep learning algorithms and adopted data sets demonstrate that an adversarial training-based min-max method can provide enhanced protection against adversarial attacks in an intrusion detection system. However, assessing the robustness of adversarial-trained DNN models against various threats remains challenging. Though the analysis gives some light on the model's resistance to a single attack, the model's performance in the face of many adversarial attacks has yet to be studied.

Ding et al. [15] introduced several gradient sign iterative methods for producing adversarial instances in the picture classification domain, particularly in protecting the privacy of photographs on IoT devices. Their findings indicate that the proposed approaches, which need low processing time and minimal alteration in visual effects, can successfully fool neural networks during image classification. However, the study needs a comprehensive examination of the transferability and efficacy of the proposed methodologies in different application domains.

Traditional machine learning and deep learning models were assessed by Papadopoulos et al. [16] to understand adversarial learning better. Using the BoT-IoT dataset, they evaluated the label noise resistance of an SVM model. Additionally, adversarial example creation using FGSM was tested on binary and multiclass ANNs. Their findings indicate that an attacker can influence or circumvent detection with a high likelihood. The BoT-IoT class data are unbalanced, and changing the model's class weighting parameters can enhance its performance. The work should have addressed challenges associated with manipulating labels with a large margin from the SVM hyperplane and how these labels can impact the model's performance.

Han et al. [17] introduced a unique adversarial attack paradigm. They tested machine learning-based network intrusion detection systems (NIDS) robustness. Their findings indicate that their attack can be highly effective, with an evasion rate of over 97%, and that their proposed defense system may protect against such attacks well. In addition, the paper highlights the need to consider detection performance and anti-evasion robustness while developing feature sets. The proposed method, however, is intended to circumvent NIDSs without payload inspection and is, therefore, inapplicable to models that require payload-based detection. Similarly, Apruzzese et al. [18] analyzed the viability of adversarial attacks against ML-based NIDS and identified the conditions and capabilities required to execute such assaults. The authors found numerous research papers proposing adversarial threat models are inapplicable to realistic ML-based NIDS and that a robust model must consider some real difficulties that attackers manage to avoid detection. The primary restriction of this research is its exclusive focus on network intrusion detection issues**.**

4. Experiment

We use deep learning-based anomaly detection systems to carry out this research on two IoT datasets. We then put the deep-learning ADSs to the test against adversarial samples. To demonstrate this, we used the IoT datasets to construct our adversarial samples, which we used to train the deep learning models.

4.1 Dataset Overview

Creating new labeled datasets requires time and effort because it contains critical information about the network, users, and system environment [19]. There has been an effort to create new datasets for the IoT security domain that will aid in developing new research. Two examples of this work are the UNSW-NB15 dataset and the Bot-IoT dataset. However, the NSL-KDD dataset is the most widely used in the IDS domain; it’s underlying network traffic dates to 1998 and has many redundant and missing records, significantly impacting results [20]. This study employs contemporary datasets such as the UNSW-NB15 and the Bot-IoT. The UNSW-NB15 and Bot-IoT datasets [21] reflect current network traffic with various attack scenarios.

The Australian Centre for Cyber Security's (ACCS) Cyber Range Lab created the UNSW-NB15 dataset [20]. It depicts new modern everyday routines that include current attacks. The whole dataset is separated into a training set and a test set using the hierarchical sampling approach, containing 175,341 normal records and 82,332 anomaly records, totaling 257,673 [20]. The dataset has nine forms of current, low-footprint attacks, including Shellcode, backdoor, exploits, worms, reconnaissance, generic, analysis, and analysis DoS and fuzzes [20].

The BoT-IoT dataset [20] was also contributed by the UNSW Canberra Cyber Center's Cyber Range Lab. This dataset, developed in a dedicated IoT environment and comprising an adequate number of records with heterogeneous network profiles, provides a realistic picture of an IoT network. Normal IoT-related and other network traffic and numerous types of attack traffic often employed by botnets are all included in the BoT-IoT dataset. BoT-IoT contains

77511 normal records and 35883009 anomaly records. Keylogging, Data exfiltration, DDoS, DoS, OS, and Service Scan are among the attack categories in the dataset.

The robustness of CNN, LSTM, and DBN-based anomaly detection models against adversarial attacks was tested using targeted attacks. The UNSW-NB15 and Bot-IoT datasets were used to build adversarial examples that were used to evaluate each of the deep learning models.

4.2 Generating the Adversarial Examples

The experiment used three adversarial attack algorithms: FGSM, JSMA, and CW. To implement the adversarial attacks, we used the Cleverhans Library, which contains the three algorithms [14]. Training the deep learning model is conducted using pre-processed training and test data. The same process is followed for the adversarial attack generation; the pre-processed testing set is employed to produce the poisoned samples using the three attack generations FGSM, JSMA, and CW. Following the generation of the poisoned negative test set, it is utilized to evaluate the classifier's ability to make predictions using deep learning networks. On the datasets, targeted attacks on the normal class are conducted. The datasets are divided into two categories: training and test sets. The performance investigation was conducted in a white-box environment, where the attacker has complete knowledge of the model, target, and data.

The Keras model employs a convolutional neural network to generate the adversarial test set [22]. Then, adversarial samples were generated using the FGSM attack algorithm. The final stage was to evaluate the classifier's performance using original and poisoned test sets. A similar procedure was followed in the case of the JSMA and CW attacks. For all of the attacks, default parameters were used. Cleverhans makes it simple to change the values if needed [14]. The difference between the poisoned test set generated by the attacks and the original test set is calculated to determine the features altered by the attacks. To acquire satisfactory results for the deep learning networks, the results were compared to the averages obtained by running the model several times on the UNSW-NB15 and Bot-IoT datasets.

4.3 Evaluation Matrix

The success of an IDS is frequently measured in terms of its accuracy values when examining controlled data of known legitimate and illegitimate behavior. Four metrics comprise the accuracy and are utilized to calculate the accuracy, recall, precision, and F1-score. When an attack is correctly classified, we have a True Positive (TP); when a benign incidence is correctly classified, we get a True Negative (TN). We have a False Positive (FP) when a benign incident is classified as an attack and a False Negative (FN) when an attack is classified as a benign incident [23].

5. Result and Evaluation

A variety of pre-processing techniques have been employed. Both datasets have varying scales and contain some outliers. All feature values were rescaled to a range of 0 to 1 to shorten training time and eliminate outliers. An encoder has been used to transform categorical data into numbers. The process of evaluating the DNN models consists of two stages. The models are trained and tested with clear data to achieve six baseline deep-learning-based ADSs. Each model is trained and tested on the two datasets. For example, CNN is trained and tested first with original UNSW-NB15 and then with original Bot-IoT datasets, resulting in CNN-UNSW and CNN-Bot models. The same is done with the LSTM and DBN models. The same process as the first stage was followed in the second stage. The baseline DNN models were then attacked with the different adversarial samples generated by the three attack generations FGSM, JSMA, and CW.

Table 1. Result of the baseline models

Model	Accuracy	Recall	Precision	F1
CNN-UNSW	96% 97%	95%	98%	99%
CNN-Bot		96%	95%	98%
LSTM-UNSW	97%	95%	95%	98%
LSTM-Bot	95%	97%	96%	99%
DBN-UNSW	98%	98%	98%	98%
DBN-Bot	97%	96%	94%	99%

Table 1 shows the performance metrics baseline anomaly detection models deployed with adversarial-free datasets. That is, each of the deep-learning models was evaluated with the selected two original datasets. For example, the accuracy of the CNN, LSTM, and DBN trained on the two selected datasets achieved 96.5%, 96%, and 97.5% as an average, respectively.

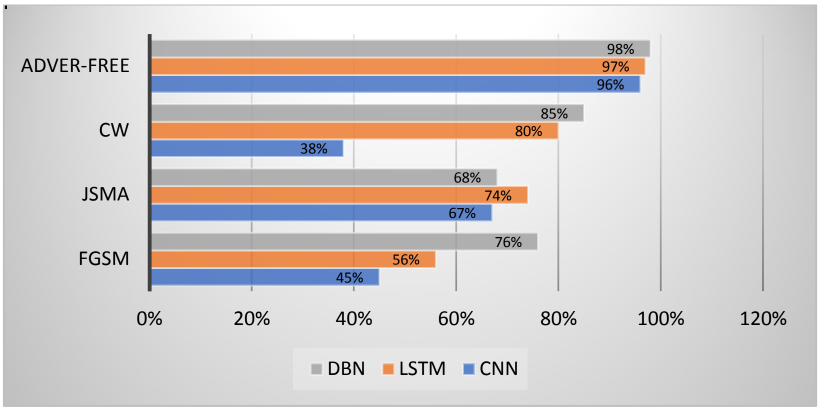

Fig. 1. Accuracy result of the UNSW-NB15 dataset

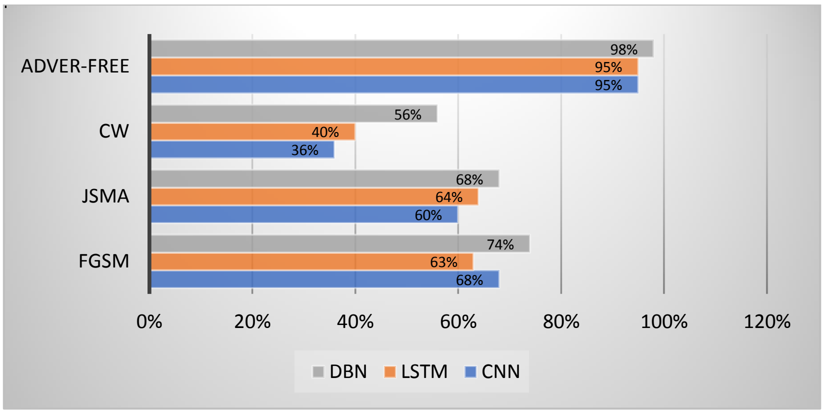

Fig. 2. Recall the result of the UNSW-NB15 dataset.

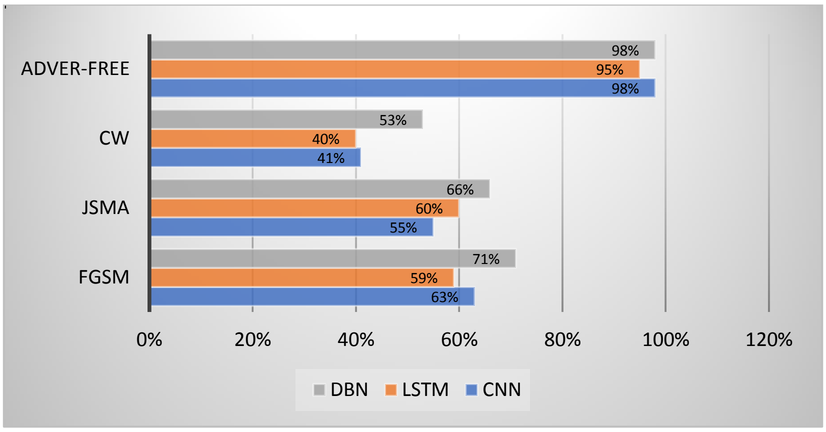

Fig. 3. Precision result of the UNSW-NB15 dataset

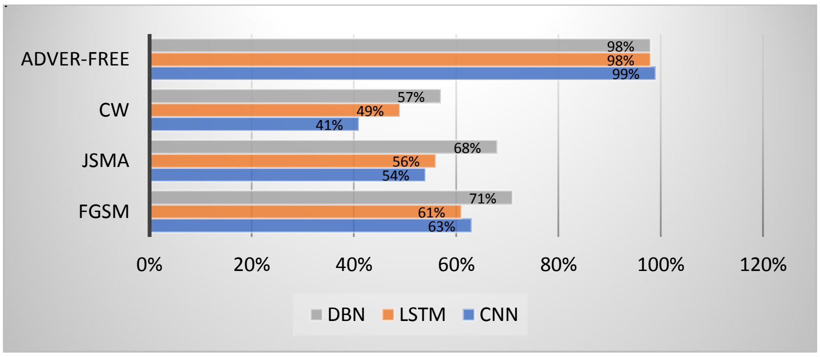

Fig. 4. F1 scores of the UNSW-NB15 dataset

5.1 Using FGSM attacks on UNSW-NB15

Fig. 1-4 display the result of examining the original data after the attack on the UNSW-NB15. As predicted, the baseline models used to produce adversarial samples on the test set are impacted by the FGSM attack for this dataset, with a 38% decrease in accuracy. The UNSW-NB15 demonstrated a drop in accuracy, precision, F1 score, and recall for CNN, LSTM, and DBN. It revealed a decline in the overall performance of the FGSM on the UNSW-NB15.

CNN was the method most affected by the adversarial samples in terms of accuracy, precision, F1 score, and recall (see Fig. 1-4). CNN demonstrated a reduction in accuracy of 51%, precision of 35%, F1 score of 36%, and recall of 27%. In contrast, DBM exhibited a slight change in precision following the attack and only a 20% decrease in accuracy, F1 score, and recall. The attack caused a 41% decrease in LSTM accuracy, 36% precision, 36% F1 score, and 32% Recall.

5.2 Using JSMA attacks on UNSW-NB15

The UNSW-NB15 JSMA attack metrics are reported in Fig. 1 through 4. The baseline models used to generate adversarial samples are impacted by the JSMA attack with a 27% reduction in accuracy, which successfully impaired the model's performance. In addition, the overall accuracy, precision, F1 score, and recall decrease. It demonstrated the UNSW-NB15 dataset's susceptibility to JSMA. Depending on the metrics selected for the study, the performance of deep learning models changes. CNN's precision was reduced by 35%, accuracy by 29%, F1 score by 43%, and recall by 35. In the meantime, LSTM accuracy decreased by 23%, precision by 35%, F1 score by 42%, and recall by 31%. The accuracy of DBM decreased by 30%, the precision by 32%, the F1 score by 31%, and the recall by 30%.

5.3 Using CW attacks on UNSW-NB15

The outcomes of the CW attack on the UNSW-NB15 are displayed in Fig. 1 through 4. As demonstrated in Fig.1, the baseline models used to generate adversarial samples for the normal class are affected by the CW attack for this dataset with an approximately 29% loss in accuracy. The results demonstrate that CW attacks affect accuracy, precision, F1 score, and recall. In terms of accuracy, DBM performs much better than CNN and LSTM. DBM accuracy decreased by 13%, precision by 43%, F1 score by 31%, and recall by 42%. CNN's accuracy decreased by 58%, precision by 57%, F1 score by 58%, and recall by 59%. In contrast, LSTM demonstrated a decline across all parameters, with a loss of 55% for precision, 49% for F1 score, and 55% for recall.

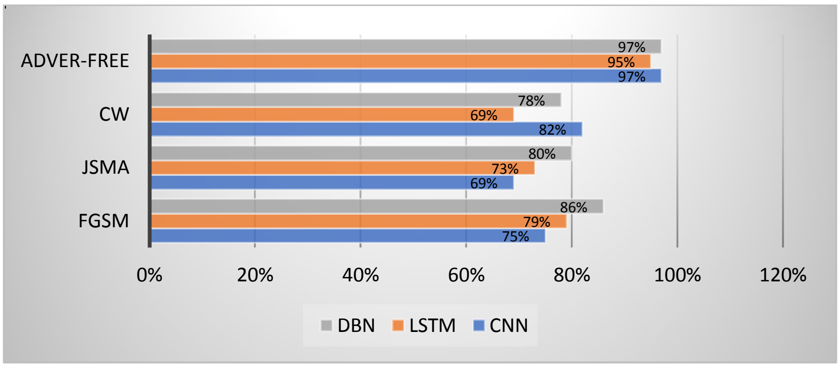

Fig. 5. Accuracy result of the BoT-IoT dataset

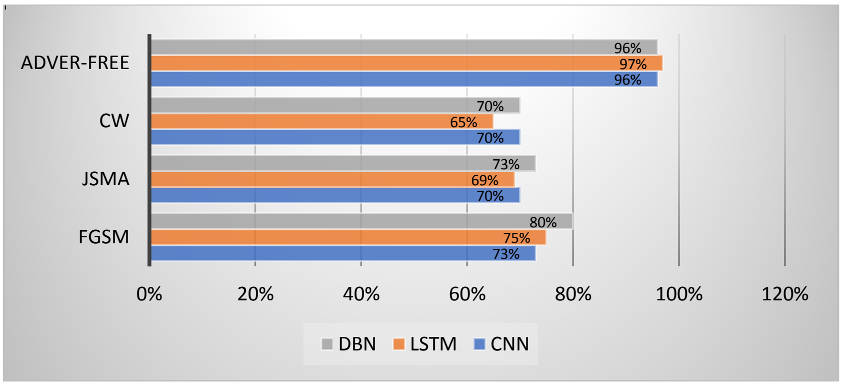

Fig. 6. Recall the result of the BoT-IoT dataset.

Fig. 5 through 8 depict the results of evaluating the original dataset and the BoT-IoT attacks. For this dataset, the FGSM attack reduces the accuracy of the baseline models by an average of 16 percent. After the attack, the BoT-IoT performance decreased across all parameters.

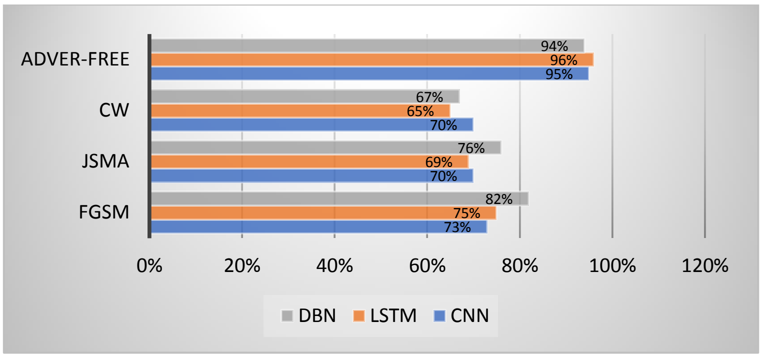

Fig. 7. Precision result of the BoT-IoT dataset

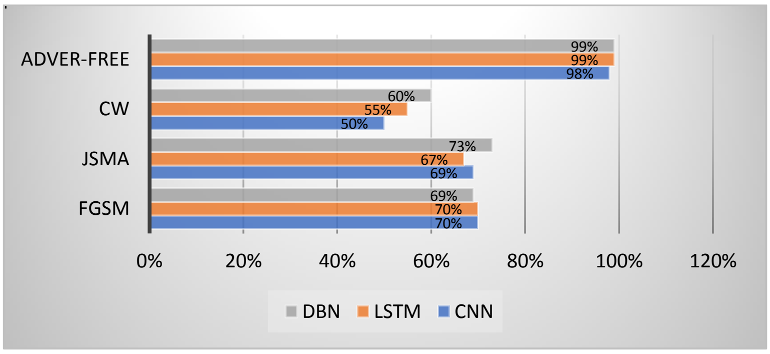

Fig. 8. F1 scores of the BoT-IoT dataset

5.4 Using FGSM attacks on BoT-IoT

CNN demonstrated a 22% decrease in accuracy, a 22% decrease in precision, a 28% decrease in F1 score, and a 23% decrease in recall. CNN was the most affected in terms of precision, recall, and accuracy. LSTM yielded a drop in accuracy of 16%, precision of 21%, F1 score of 29%, and recall of 22%. Accuracy of DBM decreased by 11%, precision by 12%, F1 score by 30%, and recall by 16%.

5.5 Using JSMA attacks on BoT-IoT

Fig. 5-8 describes the findings of the JSMA attack on the Bot-IoT dataset. For this dataset, the JSMA attack reduces the accuracy of the baseline models. After evaluating the adversarial samples for the JSMA attack, all metrics for the CNN, LSTM, and DBM algorithms were reduced. According to all metrics, CNN's accuracy decreased by 28 percentage points, precision by 25 percentage points, F1 score by 29 percentage points, and recall by 26 percentage points. Accuracy of LSTM fell by 22%, precision by 27%, F1 score by 32, and recall by 28%. Accuracy in DBM decreased by 17 percentage points, precision by 18 percentage points, F1 score by 26 percentage points, and recall by 23 percentage points. The results demonstrate that JSMA attacks negatively impacted all metrics on the Bot-IoT dataset.

5.6 Using CW attacks on BoT-IoT

Fig. 5-8 detail the results of the CW attacks on the Bot-IoT dataset. The CW attack on this dataset compromised the accuracy of the baseline models. The results demonstrate that the CW attacks influence parameters in the selected metrics in the Bot-IoT dataset. According to the results of all metrics, the CW attack damaged all deep-learning models. CNN's accuracy declined by 15%, precision by 25%, F1 score by 48, and recall by 26%. Accuracy decreased by 26%, precision by 31%, F1 score by 44%, and recall by 32% for LTSM. The accuracy of DBM decreased by 19%, precision by 27%, F1 score by 39%, and recall by 26%.

6. Discussion

The three selected attack algorithms, FGSM, JSMA, and CW, performed differently on the datasets employed for the evaluation. On the Bot-IoT dataset, the average accuracy of deep learning models declined by 33%, precision by 20%, F1 score by 31%, and recall by 30%. On the other hand, the average accuracy of UNSW-NB15 decreased by 47%, precision by 12%, F1 score by 41%, and recall by 40%. Each dataset is susceptible to JSMA, FGSM, and CW adversarial attacks. In contrast to UNSW-NB15, Bot-IoT saw a smaller drop across all measures.

Additionally, the resilience of the deep-learning models and attacks differs among datasets. CW performed the best in terms of precision and recall on the UNSW-NB15 dataset, with an overall decrease in accuracy of 42% and recall of

39%. In comparison, FGSM was the most effective at reducing the average performance for accuracy by 50%. The precision was reduced by 39% due to the JSMA. On the Bot-IoT dataset, the results indicate that CW is the most effective attack based on all three criteria, decreasing the average precision by 31%, the F1 score by 45%, and the recall by 33%. JSMA was the most effective assault in reducing precision, with an overall average reduction of 30%. JSMA and FGSM were identified as the attacks with the least efficient performance across both datasets. JSMA performed the worst in precision, F1, and recall, whereas FGSM performed the worst in precision. This implies that attack performance varies among datasets and affects metrics differently, which is a crucial consideration for an attacker starting an assault. The amount of time spent on adversarial sample generation is an additional consideration. The CW method takes the longest to make perturbed samples, while the FGSM method is the fastest.

The deep-learning model with the highest performance on the UNSW-NB15 is DBN, with an accuracy overall average decrease of 22%, a precision decrease of 35%, an F1 score decrease of 33%, and a Recall decrease of 32% increase. CNN is the least robust model, with an overall decrease in all parameters, including a 39% fall in accuracy, a 34% decrease in precision, and a 40% decrease in both F1 score and recall. In terms of accuracy, F1, and Recall, DBN is the most robust model on the Bot-IoT dataset by 32%, 33%, and 31%, respectively. With only a 1% decline in precision, LSTM was the most resilient technique. In contrast, CNN and LSTM are the least robust models in terms of precision, with an overall 29% decline. CNN is the least resistant in terms of accuracy and F1, with decreases of 40% and 39%, respectively, while LSTM performed the worst in terms of recall, with a reduction of 32%.

7. Conclusion

The IDS study has grown exponentially due to the involvement of machine learning models. The interest of researchers in this field is increasing day by day. ML-based IDS and DL-based IDS are vulnerable to adversarial attacks, which cause results to be displayed incorrectly and thus give incorrect classification. This paper has explained a critical understanding of adversarial DL Algorithms and how adversarial attacks are a high risk on machine learning-based IDS. This paper examined the impact of common adversarial machine learning attacks JSMA, FGSM, and CW on deep-learning-based anomaly models using contemporary IDS datasets UNSWNB15 and Bot-IoT. This work indicated that the attacks above effectively degraded the overall performance of the DBN, LSTM, and CNN models employed on the two IDS datasets. Adversarial attack threats on DL algorithms were considered a white box because adversarial attacks can control the input data to train the model. DBM was both datasets' most robust deep-learning model, whereas CNN was the least robust. The results of the attacks differed based on the two datasets. Overall, JSMA performed poorly on both datasets. The most effective attack against the two datasets was CW. In a future study, we will incorporate various modern IDS datasets, adversarial attack generation techniques, and deep-learning models. In addition, we will design and evaluate various defensive approaches against adversarial attacks to enhance the resilience of deep learning models utilized by anomaly detection systems.

References

[1] K. Albulayhi and F. T. Sheldon, "An adaptive deep-ensemble anomaly-based intrusion detection system for the Internet of Things," in 2021 IEEE World AI IoT Congress (AIIoT), 2021: IEEE, pp. 0187-0196.

[2] R. Abou Khamis, M. O. Shafiq, and A. Matrawy, "Investigating resistance of deep learning-based ids against adversaries using min-max optimization," in ICC 2020-2020 IEEE International Conference on Communications (ICC), 2020: IEEE, pp. 1-7.

[3] K. Albulayhi, Q. Abu Al-Haija, S. A. Alsuhibany, A. A. Jillepalli, M. Ashrafuzzaman, and F. T. Sheldon, "IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method," Applied Sciences, vol. 12, no. 10, p. 5015, 2022.

[4] K. Albulayhi, A. A. Smadi, F. T. Sheldon, and R. K. Abercrombie, "IoT Intrusion Detection Taxonomy, Reference Architecture, and Analyses," Sensors, vol. 21, no. 19, p. 6432, 2021.

[5] I. Debicha, T. Debatty, J.-M. Dricot, and W. Mees, "Adversarial training for deep learning-based intrusion detection systems," arXiv preprint arXiv:2104.09852, 2021.

[6] Alsulami, A.A.; Abu Al-Haija, Q.; Alqahtani, A.; Alsini, R. Symmetrical Simulation Scheme for Anomaly Detection in Autonomous Vehicles Based on LSTM Model. Symmetry 2022, 14, 1450. https://doi.org/10.3390/sym14071450

[7] P. Ioulianou, V. Vasilakis, I. Moscholios, and M. Logothetis, "A signature-based intrusion detection system for the internet of things," Information and Communication Technology Form, 2018.

[8] R. A. Ramadan and K. Yadav, "A novel hybrid intrusion detection system (IDS) for the detection of internet of things (IoT) network attacks," Annals of Emerging Technologies in Computing (AETiC), Print ISSN, pp. 2516-0281, 2020.

[9] R. V. Mendonça, J. C. Silva, R. L. Rosa, M. Saadi, D. Z. Rodriguez, and A. Farouk, "A lightweight intelligent intrusion detection system for industrial internet of things using deep learning algorithms," Expert Systems, vol. 39, no. 5, p. e12917, 2022.

[10] H. Jmila and M. I. Khedher, "Adversarial machine learning for network intrusion detection: A comparative study," Computer Networks, p. 109073, 2022.

[11] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," Future Generation Computer Systems, vol. 110, pp. 148-154, 2020.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp), 2017: Ieee, pp. 39-57.

[14] 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 European symposium on security and privacy (EuroS&P), 2016: IEEE, pp. 372-387.

[15] X. Ding, S. Zhang, M. Song, X. Ding, and F. Li, "Toward invisible adversarial examples against DNN-based privacy leakage for Internet of Things," IEEE Internet of Things Journal, vol. 8, no. 2, pp. 802-812, 2020.

[16] P. Papadopoulos, O. Thornewill von Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, and W. J. Buchanan, "Launching adversarial attacks against network intrusion detection systems for iot," Journal of Cybersecurity and Privacy, vol. 1, no. 2, pp. 252-273, 2021.

[17] D. Han et al., "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," IEEE Journal on Selected Areas in Communications, vol. 39, no. 8, pp. 2632-2647, 2021.

[18] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, "Modeling realistic adversarial attacks against network intrusion detection systems," Digital Threats: Research and Practice, 2021.

[19] B. A. NG and S. Selvakumar, "Anomaly detection framework for Internet of things traffic using vector convolutional deep learning approach in fog environment," Future Generation Computer Systems, vol. 113, pp. 255-265, 2020.

[20] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Information Security Journal: A Global Perspective, vol. 25, no. 1-3, pp. 18-31, 2016.

[21] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," Future Generation Computer Systems, vol. 100, pp. 779-796, 2019.

[22] C. Keras, "Theano-based deep learning libraryCode: https://github. com/fchollet," Documentation: http://keras. io, 2015.

[23] Abu Al-Haija, Q.; Zein-Sabatto, S. An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks. Electronics 2020, 9, 2152. https://doi.org/10.3390/electronics9122152

Authors' Profiles

Khalid Albulayhi (student Member, IEEE) received the B.S degree in information system from King Saud University, Saudi Arabia, in 2004 and the M.S. degree in computer science from Ball State University, Muncie, USA, in 2012. He received Ph.D. degree in computer science from University of Idaho, USA in 2022. His research interests include cybersecurity, Information System, Intrusion Detection System (IDS), the Internet of Things (IoT), Machin Learning and Artificial Intelligence.

Qasem Abu Al-Haija received his Ph.D. in Computer Engineering from Tennessee State University (TSU), USA, in 2020 and his M.Sc. in Computer Engineering from Jordan University of Science and Technology, Jordan, in 2010. His research interests include Artificial Intelligence (AI), Cybersecurity and Cryptography, Malware Analysis, the Internet of Things (IoT), Cyber-Physical Systems (CPS), Time Series Analysis (TSA), and Computer Arithmetic.

How to cite this paper: Khalid Albulayhi, Qasem Abu Al-Haija, "Adversarial Deep Learning in Anomaly based Intrusion Detection Systems for IoT Environments", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.13, No.4, pp. 1-10, 2023. DOI:10.5815/ijwmt.2023.04.01