computers

MDPI

Article

Mitigation of Black-Box Attacks on Intrusion Detection Systems-Based ML

Shahad Alahmed [1], Qutaiba Alasad [2,*], Maytham M. Hammood [1], Jiann-Shiun Yuan [3] and Mohammed Alawad [4]

[1] Department of Computer Science, Tikrit University, Al Qadisiyah P.O. Box 42, Iraq; shahad.m.mustafa35522@st.tu.edu.iq (S.A.); maythamhammood@tu.edu.iq (M.M.H.)
[2] Department of Petroleum Processing Engineering, Tikrit University, Al Qadisiyah P.O. Box 42, Iraq
[3] Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA; jiann-shiun.yuan@ucf.edu
[4] Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA; alawad@wayne.edu
* Correspondence: qutaibaeng@knights.ucf.edu

Abstract: Intrusion detection systems (IDS) are a very vital part of network security, as they can be used to protect the network from illegal intrusions and communications. To detect malicious network traffic, several IDS based on machine learning (ML) methods have been developed in the literature. Machine learning models, on the other hand, have recently been proved to be effective, since they are vulnerable to adversarial perturbations, which allows the opponent to crash the system while performing network queries. This motivated us to present a defensive model that uses adversarial training based on generative adversarial networks (GANs) as a defense strategy to offer better protection for the system against adversarial perturbations. The experiment was carried out using random forest as a classifier. In addition, both principal component analysis (PCA) and recursive features elimination (Rfe) techniques were leveraged as a feature selection to diminish the dimensionality of the dataset, and this led to enhancing the performance of the model significantly. The proposal was tested on a realistic and recent public network dataset: CSE-CICIDS2018. The simulation results showed that GAN-based adversarial training enhanced the resilience of the IDS model and mitigated the severity of the black-box attack.

Keywords: intrusion detection systems; machine learning; random forest; generative adversarial networks; recursive features elimination; principal component analysis; adversarial instances or examples

Citation: Alahmed, S.; Alasad, Q.; Hammood, M.M.; Yuan, J.-S.; Alawad, M. Mitigation of Black-Box Attacks on Intrusion Detection Systems-Based ML. *Computers* **2022**, *11*, 115. https://doi.org/10.3390/computers11070115

Academic Editor: Paolo Bellavista

Received: 6 July 2022
Accepted: 15 July 2022
Published: 20 July 2022

1. Introduction

Nowadays, the use of the Internet, in general, and reliance on cloud-based resources is growing at an exponential rate. Operations are concentrating on their core businesses while transferring their information technology (IT) services to the cloud. Many more factors encourage businesses to use internet-based offerings. Likewise, malicious traffic has increased at a rapid rate [1]. Today's cyberattacks are becoming more diversified and broad. The purpose of these assaults is to obtain unauthorized access to remote data or to create service interruptions for consumers. These attacks have a tremendous influence not only on the economy and finances of a country but also at the national level, in addition to cultural security [2,3]. Therefore, such assaults should be prevented from both inside and outside as well as from governmental and private institutions [4–6]. As a result, it is critical to rely on automated powerful systems for quickly and reliably identifying threats. Interestingly, intrusion detection systems (IDS) have been considered an excellent solution to further boosting the security level of a system [7].

IDS is a type of security software that observes network traffic and gives warnings once an unusual behavior is discovered [8]. Generally, there are two kinds of IDS: host-

based intrusion detection systems (HIDS) and network-based intrusion detection systems (NIDS) [9]. HIDS, known as a passive component, focuses on a single machine [10]. On the other hand, NIDS, known as an active component, is used to analyze network packets as well as to monitor and safeguard a system from network risks [9]. The focus of this research was on NIDS, since the proposed IDS analyzes the flow of data among computers, such as network traffic, detects unusual behavior, and defends nodes from complex assaults.

Machine learning (ML) algorithms are frequently employed in network security to further enhance the capabilities of an IDS in identifying attacks due to the nature of its learning abilities. They have shown outstanding performance as powerful and successful defense mechanisms [11]. However, the adoption of ML in this domain poses severe challenges to cyber defense, with adversarial machine learning (AML) assaults being one of the most critical [12]. AML has lately arisen as a serious threat to the success of such systems. An adversarial attacker can weaken network protection by exploiting flaws in the ML methods. By generating small disturbances into network traffic, attackers can utilize the vulnerability and cause NIDS to damage [13]. The fabrication of samples is meant to disrupt the ML algorithm by eliciting outcomes favorable to the attacker, which is one of the malevolent behaviors. In the cyber security area, these flaws are crucial, since an undetected intrusion may compromise an entire enterprise [14]. Therefore, it is essential to develop ML-based approaches that can protect the systems from adversarial attacks.

Among different ML algorithms, generative adversarial networks (GANs) have been extensively leveraged with adversarial attacks. A GAN is a type of deep learning techniques in which two neural networks compete against each other in a two-player game. It has demonstrated the advantage of ML in producing higher-dimensional data, such as images, audio, and text, since it was initially released in 2014 [15]. Many research papers have employed GANs to either enhance IDS or to design novel attack instances such as generating adversarial malware samples [16,17]. However, there are only a few works on GAN-based intrusion detection [18].

GAN-based adversarial training (AT) can be utilized as a defense mechanism. AT methods inject adversarial examples (AEs) into the training data to make sure that the machine learning technique imparts adversarial perturbations as much as possible [18]. Therefore, they enhance the generalization and resilience of the machine learning techniques, since these models are trained on clean and adversarial examples [19]. In AT, AEs are generated by GANs and then added to the clean dataset [20]. This approach shows that the added AEs elevated the accuracy of the trained model by approximately 10% [20]. However, as the same GAN model is used during the training (i.e., defending) and testing (i.e., attacking), the accuracy will go down once the proposal tests with different attack models. Further, the size of the epochs utilized in [20] was only 100, which is not sufficient to generate strong AEs.

To overcome the aforementioned shortcomings, we present a GAN-based approach for the adversarial training of ML models against AEs generated using a black-box attack, namely, the zeroth order optimization (ZOO) attack. Then, we evaluated the resilience of the designed system by presenting the ZOO black-box attack method to define adversarial perturbation in the data network. Note that an opponent did not have access to the specifics of the ML-based IDS, which made the experiment more realistic, and we took this into consideration during the implementation of the ZOO attack. Our finding was that the employment of a GAN as a defensive strategy makes ML-based IDS more resilient to previously unseen and unknown adversarial perturbations.

1.1. Motivation

Over the years, computer networks have grown swiftly, contributing greatly to social and economic progress. However, compared to other sectors, network security applications of ML confront a significant concern regarding active adversarial attacks [21,22]. This happens due to the adversarial nature of ML applications in network security. In a battle among both attackers and defenders that may be described as an arms competition, ML

systems are continually probed by adversaries with inputs that are specifically meant to evade the system and generate a false prediction. Furthermore, malicious attacks have become more common, and ML models' defense and resistance against them must be addressed. Several studies in text and image recognition fields have looked at the danger and provided viable countermeasures. Unfortunately, not much research on the NIDS sector that addresses the problems of adversarial attacks has been undertaken [12]. In addition, the learning model and dataset quality are both closely connected to the efficiency of the IDS. Many researches have been dependent on datasets that have significant shortcomings such as simulated traffic (i.e., not from an actual production network), anonymity, redundancy, and outdated attack traffic, e.g., denial of service (DDoS) [20,23]. Other studies have concentrated on the adversary knowledge factor, such as white-box attacks, and shown that such attacks are strong in targeting a system under the assumption that opponents have full access and knowledge of the classifier [24,25]. In practice, having such an ability by an attacker seems to be elusive. It has been proven that a GAN is a very serious and powerful attack compared with other existing attacks [26]. Contrary to white-box attacks, GAN-based black-box attacks are considered weak, as attackers have no knowledge or only have superficial information about the victim classifier. A GAN-based adversarial ML attack has been proposed and validated on a black-box IDS, and it turned out that GAN is a powerful technology for bypassing an IDS due to the fact of its potential to generate data that have a similar distribution to the original dataset [20]. In general, there is a lack of research studies investigating and evaluating the effectiveness of existent adversarial defensive mechanisms. Accordingly, it is necessary to ensure the resilience of the proposed methods against adversarial attacks and to pay more attention to proposing attack-agnostic defense mechanisms that address the increasing variety of adversarial attacks, rather than focusing only on a narrow range of attacks [27]. Therefore, the above reasons served as motivation to propose the main contribution in this paper.

1.2. The Contribution of the Paper

The major contribution of this paper is as follows:

- We used a GAN to generate strong adversarial examples, for the first time, from the CSE-CIC-IDS2018 dataset, which was introduced by the Canadian Institute of Cybersecurity (CIC), called the Communications Security Establishment and CIC 2018 (CSE-CIC-IDS2018) Dataset. The strong adversarial examples are generated using a GAN with 2000 epochs;
- We designed a defensive model for NIDS-based random forest classifier and enhanced the proposed model using GAN-based adversarial training, where the generated adversarial examples are used for training the model and measuring the model resistance in two phases. The first phase was utilized to train the proposed technique on a non-crafted dataset, and the second phase was related to improving the robustness and accuracy of the first phase by retraining the proposed model on a combined dataset that included a non-crafted dataset, generated dataset from a GAN;
- Our proposal was further improved by carefully training our model with valuable features selected by PCA with the generated adversarial examples;
- We implemented a black-box-based ZOO attack to evaluate the resistance of the proposed random forest model in which this attack was capable of generating adversarial examples that the model had never seen before.

To the best of our knowledge, there is no recent work concentrated mainly on the improvement of ML-based IDS as a defense model and evaluating the resilience of the model by thwarting new unseen attacks.

The remainder of this paper is structured as follows: Section 2 tackles, in brief, the random forest model, GAN-based defense technique, black-box-based ZOO attack, and feature reduction methods. Section 3 tackles the curriculums of related work including adversarial attack and defense techniques. Our proposed methodology is described in detail in Section 4. The experimental setup and results of the proposal are illustrated in

Section 5. A comparison with other prior work is given in Section 6. Finally, we provide the conclusion in Section 7.

2. Background

In this section, we present the fundamentals of the random forest classifier and explain the GAN architecture in detail. Afterward, we describe the realistic threat model scenario, ZOO attack, which was considered in our work, and then we explain the current feature selection and reduction methods.

2.1. Random Forest (RF) Classifier

RF is one of the most powerful methods employed to solve classification and regression issues in machine learning. It is a class of supervised classification algorithms. The random forest requires two steps: one is to tune the random forest configuration, and the other is to predict the incoming results obtained from step one [28]. The random forest algorithm is implemented based on building multiple decision trees; each one represents a classifier. Every tree in the forest is sampled from the original dataset to create a sub-dataset. Then, subsets of data are placed in each decision tree, and each decision tree produces results. The result of the final decision is determined via a vote by all decision trees. A tree does not select all the features, instead only some features are randomly chosen; then, from the chosen features, only the optimal features are selected. Because of this randomness, its variance decreases, and a better overall classification model is also produced [29].

2.2. GAN-Based Defense Methodology

A GAN is a deep learning approach that is composed of two NNs, each one against the other in a game setting as shown in Figure 1 [30]. It has been studied in depth in the field of security, as a GAN is capable of generating new unseen threats. The usage of a GAN as a defense mechanism renders the model more robust against future attacks. The main objective of a GAN is to detect unknown or unseen attacks and protect systems from various vulnerabilities [18]. In a zero sum game context, a GAN has two NNs competing against each other. One is leveraged for producing regression and labeled as a generator (G), while the second is labeled as a discriminator (D). Usually, the purpose of the generator is to take random noise (V) as input, transform it using the NNs, and create false instances, whereas the aim of the discriminator is to use a NN to separate the infected data generated via the generator from the actual one [31,32]. When the process reaches equilibrium, the discriminator is unable to recognize between real and bogus data. The generator, therefore, accepts random noise (V) as input and produces actual instances as output. That is to say, the generator has found how the data is distributed [26,33]. The adversarial loss for both G and D is given in Equations (1) and (2), respectively [34].

$$L_G = E_{M \in S_{\text{attack}}, N} D(G(M, N)) \tag{1}$$

$$L_D = E_{s \in B_{\text{benign}}} D(s) - E_{s \in B_{\text{attack}}} D(s) \tag{2}$$

In the above equations, S refers to the data collected from the generator and leveraged to train the discriminator, while the variable E is the expected volume of the produced data that is indicated to be an attack or benign. B_{benign} is a variable for the benign data, and B_{attack} is the attack data.

2.3. Black-Box-Based ZOO Attack

The ZOO-attack-based method was first introduced in [35] to generate adversarial examples (AEs). Note that white-box-attack-based methods differ from black-box attacks, as black-box methods do not rely on the gradient information of the target model. The black-box attack process represents a targeted misclassification by which the data are crafted to generate AEs. The generation of such examples relies on the modification of optimization parameters and on a conjecture of confidence, rather than the gradation.

When an attacker generates these examples, he or she utilizes them to violate IDS [36]. In this paper, the threat model settings assume an attacker only queries the model for relevant labels and has no access to the IDS model, including its hyperparameters. The goal of such an attacker is to generate AEs that are hard to detect via the IDS model, and this makes the model vulnerable to many threats.

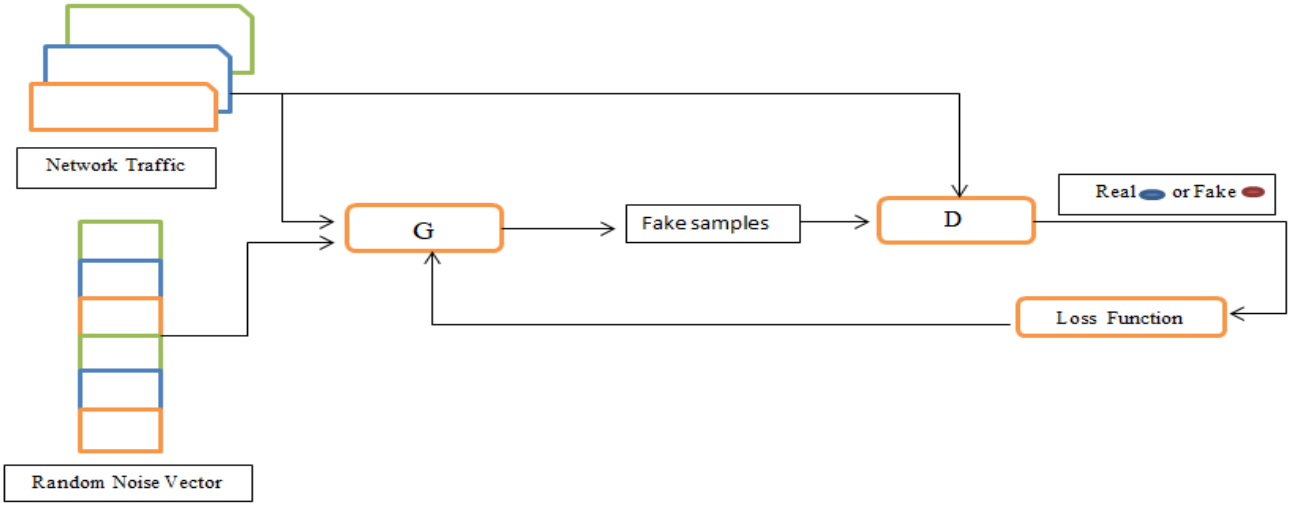


Figure 1. A block diagram of the GAN model.

2.4. Reduction Techniques

To further increase the IDS model's resilience, the most valuable features should be extracted from the collected dataset. This will also help to decrease the data's dimensions and the model's complexity. Such a method is known as a reduction method in which only valuable features are chosen during the classification. In this work, two reduction techniques, known as "PCA" and "RFE", are tackled in the following.

- **Principle component analysis (PCA)** is widely employed to extract preferable features and compress them, in which the dimensions of the feature are reduced. Note that this also leads to the diminishment of the computational time and the model's complexity. The subsets of the feature set are extracted via PCA, and this helps diminish the search range [37]. In fact, the general usage of PCA is to extract important features for traffic analysis [38];
- **Recursive feature elimination (Rfe)** is utilized to select some valuable features out of all of the features in the dataset. Only features with high ranking are selected, and reset features (e.g., those with low ranks) are eliminated one by one. Rfe technique removes duplicated features and extracts only preferable and valuable features from all dataset features. The goal of Rfe is to choose the best subsets of valuable features [10].

3. Literature Review

Given the most recent resurgence of DL effectiveness models and ML approaches, studies in different domains have been accomplished to resolve prominent challenges in the various realms across the world [39]. ML and DL models have been widely leveraged in data generation, network security classification, network attack modification, and forecasting. This section tackles prior work on AML attacks and possible defense techniques for NIDS.

3.1. Adversarial Attack Approaches

Researchers have investigated adversarial attacks and shown how easily they may fool ML models [40]. White-box and black-box threats are two types of adversarial attacks. The former necessitates that the white-box attack has access to the variables of the detector,

whereas the black-box setting does not [25]. Lin et al. proposed a framework, namely, IDSGAN, to produce AEs that can fool the IDS model by making prediction mistakes. IDSGAN employs a Wasserstein generative adversarial network (WGAN) to produce malicious traffic records that are hard to detect by IDS. A WGAN is composed of three major parts: a generator (G), discriminator (D), and a black-box IDS. The G generates hostile unlawful data from the incoming mixed malicious records with noise. The black-box IDS is used to foretell the malicious records from the normal ones by producing predicted labels as targets, in which the D uses these targets to impart the black-box IDS [41]. In 2019, researchers proposed an AML attack via the utilization of GANs to create a large adversarial variation in the original network dataset. This attack aimed to evade a black-box IDS. Then, GANs were used as a defense mechanism during the training phase to render the system more robust against adversarial threats. The KDD99 dataset, which is extensively used to measure IDS performance, was used to examine the proposed GAN. The experiments showed that the highest accuracy was 65.38 percent for gradient boosting (GB), and the lowest accuracy was 43.44 percent for support vector machine (SVM). After training with the GAN, the classifiers' performance improved, where the accuracy rate reached 86.64 and 79.31 percent for LR and KNN, respectively [20]. Later in 2021, a new aggressive framework, called anti-intrusion detection auto encoder (AIDAE), was proposed, where GAN was used to create features for deactivating IDS. This framework has an encoder that converts some characteristics to embedding space, and many decoders to gather discrete and continuous characteristics. Then, a GAN is employed to impart the previous distribution of the embedded space. The framework learns the typical feature spread to produce irregular features, and this does not require IDS feedback during the training operation. In addition, the proposal maintained the correlation between the created discrete and continuous characteristics. The test was carried out on the NSL-KDD, UNSW-NB15, and CICIDS2017 datasets, with six classifiers (i.e., LR, K-NN, DT, AdaBoost, RF, and CNN+LSTM). The experimental results demonstrated that the generated characteristics were capable of weakening the baseline IDS, implying that researchers need to take into consideration defending against such attacks in the future [42].

3.2. Recent Work on Defense

Many researchers have introduced novel defense techniques to prohibit various existing threats via leveraging GANs to render stronger IDS. In 2018, the author, Mirza, introduced an ensemble learning method to enhance system resilience. The results in all classifiers were merged by collecting the most valuable information from all classifiers (e.g., LR, NN, and DT) during both the training and testing phases. Afterward, a weighted majority voting mechanism was applied to each individual classifier, and the results were released to determine whether each sample was abnormal. The general accuracy for the training and testing was 96.66% and 96.13% for LR, 90.67% and 89.83% for NN, and 92.08% and 91.66% for DT [43]. In the same year, Zenati et al. proposed an anomaly detection method-based bidirectional GAN, called adversarial learned anomaly detection (ALAD). The GAN learned the distribution of the features to perform the anomaly detection goal. Afterward, recreation errors based on the adversarial features were leveraged to specify whether a given sample was malicious or benign. ALAD is constructed on the last level to guarantee "data-space", "latent-space", and "cycle-consistencies" and to stabilize a GAN during training. It was proven that the proposal elevates anomaly detection performance significantly compared to state-of-the-art studies, where the KDD99 and Arrhythmia tabular datasets and the SVHN and CIFAR-10 picture datasets were employed during the evaluation [44]. A novel intelligent IDS was introduced in which a lower number of features was used to detect intrusions. The genetic algorithm (GA) is leveraged to extract preferable features in order to minimize resource usage and time complexity. After employing the GA to remove the redundant and unnecessary data from the dataset, the GA output predicts the best features via using a specific number of comparisons. The true positive rate was enhanced when feature ranking was accomplished according to the results obtained from

averaging the values in the dataset [45]. In 2021, McCarthy et al. proposed a defensive strategy for measuring feature susceptibility to AEs generated by the fast gradient signed method (FGSM) attack. The FGSM is a combination of white-box and misclassification models, which is used to trick a NN model via rendering incorrect predictions. The presented strategy aims to strike a balance between classification results and diminishing potential attacks on the feature space. The proposal was evaluated on the CICIDS2017 dataset. The authors found that there are data features vulnerable to attack. Defense solutions are given for algorithmically generated AEs. In addition, Rfe was employed to eliminate the vulnerability features that had the largest absolute difference during the FGSM attack. Furthermore, regular feature selection for training enhanced the model's durability against AEs. With limited features, the method achieved high accuracy. When all features were taken into account, the model had the highest accuracy; however, the accuracy under assault seldom reached 60%. The results indicate that incorporating feature selection increases the accuracy rate of the model when an FGSM attack exists [25].

4. Proposed Research Methodology

In this section, we present in detail our proposed technique's structure, the dataset's preparation and preprocessing, and the evaluation metrics.

4.1. Model Structure

A framework diagram of the proposed model is demonstrated in Figure 2. It consists of three main parts. The first part is data preprocessing, which is used to prepare the original data for the ML models and apply feature reduction methods to improve the accuracy and reduce the complexity. Specifically, we used PCA and RFE to reduce the data's dimensions by selecting only the relevant features that are needed for the classification task. The second part is the defender model, which consists of the classifier model and the GAN model. The classifier is an ML model used for binary classification. The GAN model aims to generate adversarial examples (AEs) from the original dataset using arbitrary latent vector (noise vector) and retrain our classifier on the new dataset (original and synthetic dataset) to make it more resistant and powerful against known and unknown attacks in the future. The last part, i.e., the attacker model, is a black-box attack method. This model generates new AEs that aim to evade the detection system, the defender model, and influence on the predictions of the classifier to determine its robustness.

4.2. Dataset

The CSE-CIC-IDS2018 is an intrusion detection dataset created by the Communications Security Establishment and Canadian Institute for Cybersecurity on AWS (Amazon Web Services), located at Fredericton, Canada, in 2018 [46]. The IDS2018 is the updated version of the IDS2017 dataset and the latest and most comprehensive intrusion dataset, collected for launching real attacks, which is publicly available. The dataset includes the necessary standards for the attack dataset and contains many different kinds of attacks. This dataset also comprises network traffic, system logs, and 80 features [47]. To better model the attacks, a topology with a machine diversity similar to real-world networks was created [48]. The infrastructure of the network included 50 attacker machines, 420 victim machines, and 30 servers. The details of the dataset's features are provided in Table 1.

The intrusions in the CSE-CICIDS2018 dataset were normalized into two kinds, namely, benign and malicious. The number of benign and malicious network traffics is given in Table 2.

4.3. Data Preprocessing

The CSE-CICIDS2018 dataset consists of over 1,000,000 records. The dataset consists of the original traffic in the packet capture (pcap) files, the logs, the preprocessed labels, and the feature-selected comma-separated values (CSV) files. The CSV files are categorized into two classes benign (class-0) and malicious (class-1). The dataset does not contain blanks

or errors. Therefore, we applied some preliminary data processing procedures which are presented as follows:

1. **Numerical standardization:** To provide data consistency, the data were standardized using the technique of obtaining the Z-Score in which the standard deviation was set to 1, and the average value of each feature was set to 0.
2. **Outliers:** We deleted two features (i.e., the timestamp (date and time) and Fwd packets features) from the CSE-CIC-IDS2018 dataset, because they have a neglected influence on the model training. Therefore, the total number of features was 78.
3. **Replacement of default values:** In the leveraged dataset, the packet length Std feature has a value of infinity. We fixed this by changing its value to 0 in the database.

For all the experiments in Section 5, 75% of the dataset was employed to train the ML model, and the remaining 25% was employed to test the model. However, 70% of the dataset was used to train the GAN model, and the remaining 30% was used to test the model.

Table 1. Description of the CSE-CIC-IDS2018 features.

Count	Description
4	Basic features of network connections
11	Features of network packets
5	Features of network flow
22	Statistic of network flows
17	Content-related traffic features
3	Features of network sub-flows
18	General purpose traffic features

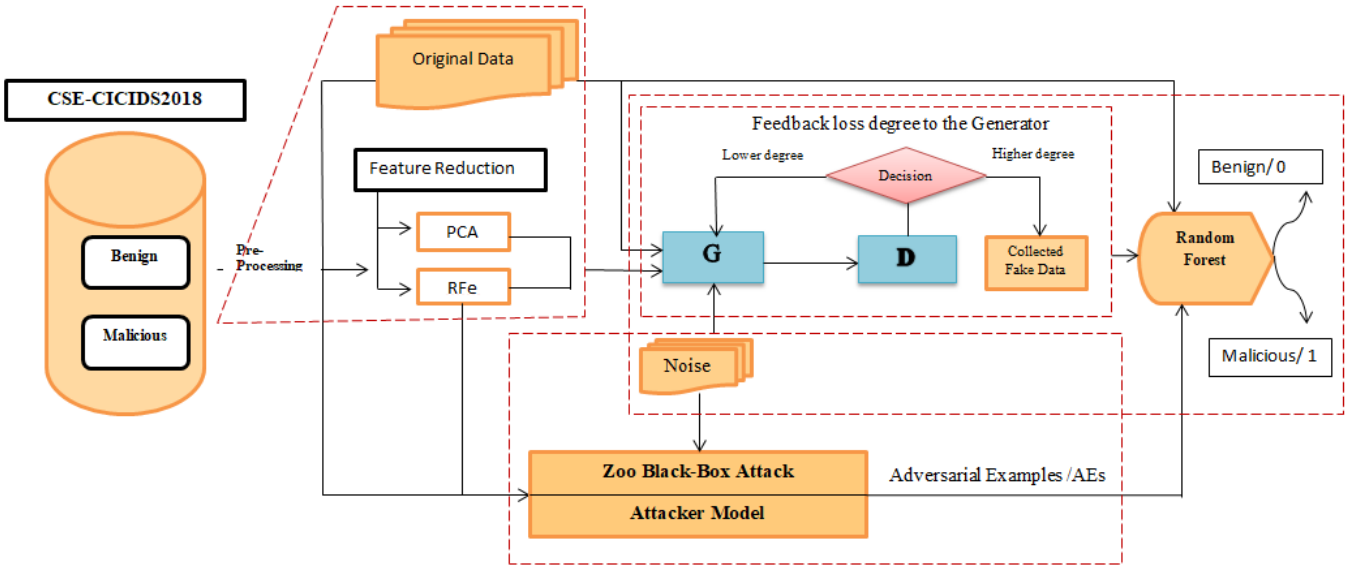


Figure 2. The overall framework of our proposal.

Table 2. Benign and malicious network count.

Types	Count
Benign	762,384
Malicious	286,191

4.4. Evaluation Metrics

There are several classification metrics for IDS. The confusion matrix (CM) of a two-class classifier was used to compute the performance metrics because, in our work, the experiments were conducted broadly to distinguish between malicious and normal records. The abbreviations of the CM are as follows:

- TP: Normal events are correctly classified by the model;
- TN: Malicious attacks are successfully identified by the model;
- FP: Normal events are incorrectly distinguished to be an anomaly;
- FN: Malicious attacks are incorrectly recognized via the model as a normal event.

The performance of our classifier model could be obtained via the following standards: accuracy (AC), precision (PC), recall (RC), F1-score, AUC-ROC, and MSE.

- ACC: The proportion of all predicted instances, including normal or abnormal, that is correctly predicted by the IDS. It is one of the longest-used metrics to measure IDS performance, and it can be very useful when the classes are imbalanced.

$$\text{ACC} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \quad (3)$$

- Precision: The ratio of normal records that are correctly identified by the IDS to all records that the IDS identified as normal.

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \quad (4)$$

- Recall: The percentage of all normal records correctly identified by IDS.

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \quad (5)$$

- F1-score: The balance between precision and Recall, and it is expressed as the harmonic mean of the two metrics.

$$\text{F1} = 2 \times (\text{Precision Recall}/\text{Precision} + \text{Recall}) \quad (6)$$

- AUC-ROC: This indicates how much or to what extent a machine learning model is capable of detecting or classifying various categories of scenarios as we intended.
- MSE: This is the average squared error between the model's predictions and the actual outcomes.

5. Results

This section presents three experimental setups with the results. In the first setup, we used all 78 features of the dataset as the model's input. However, we used dimensionality reduction methods (i.e., RFE and PCA) to reduce the number of features in experiments II and III, respectively. Each experiment included three parts: training the ML model on the original dataset; using GAN to generate adversarial examples and retrain the ML model on the original and generated dataset; evaluating the performance of the ML model when the black-box ZOO attack was applied.

5.1. Experiment I

In the first experiment, we used all 78 features to train our IDS classifier, the random forest. The classifier was used to classify the dataset into two classes (i.e., benign and malicious). The hyperparameters of the random forest classifier are illustrated in Table 3.

The second part of this experiment was to increase the ability of this model to handle more than a real dataset. Therefore, the proposed GANs with all 78 features was built to generate adversarial examples to increase the defense mechanism. The architecture of the proposed GAN consisted of two neural networks (i.e., generator (G) and discriminator (D)). The G neural network model had three layers and a Relu activation function, including an input layer with 79 units to meet the formula of the input vector after preprocessing. The

hidden layer consisted of 100 units, and the output layer of the proposed generator had 79 (78 features, 1 label), which are referred to as the vectors of the fake record generated from noise V. On the other hand, the proposed D network was designed to classify (fake or real) data generated by the G network. It was also used to update the noise vectors depending on the feedback loss function from the network. This D network consisted of three layers: an input layer with 78 units followed by the activation function Relu; a hidden layer with 100 units followed by a dropout layer with a dropping rate of 0.4 used to avoid the overfitting problem; finally, the sigmoid output layer was used for binary classification: 0 for real, 1 for fake. In this experiment, the Adam optimizer was used to update the trainable parameter at a learning rate of 0.001 as shown in Table 4.

Table 3. Random forest hyperparameters.

Hyperparameters	Values
Estimators	20
Criterion	Gini Index
Minimum samples leaf	1
Minimum samples split	2

Table 4. GAN hyperparameters.

Optimizer	Adam
Learning rate	0.001
Batch size	512
Epochs	2000
Loss function	Binary cross-entropy
Latent dimension	Depending on the input vector

After 2000 epochs, the D completely failed to classify the output from the G network, which is known as fooling the D model to distinguish between the real and generate fake samples. The loss rate of G reached the lowest value at 0.002, while the D loss rate reached 17.12 as shown in Figure 3. The process of training the GAN led to the generation of 230,000 samples. After this process, the generated data were merged with the real data and used to retrain the proposed RF model. Table 5 shows the classification results of the proposed RF model with all 78 features.

Table 5. Evolution metrics of our IDS classifier (RF) including accuracy, precision, recall, F1-score, and MSE for Experiment I.

RF	Accuracy	Precision	Recall	F1-Score	MSE
Before GAN	0.863	1.00	0.84	0.91	0.13
After GAN	0.85	0.99	0.80	0.88	0.14
After ZOO attack	0.69	0.62	0.74	0.67	0.30

The third part of this experiment was to evaluate the proposed IDS model using a black-box attack. We used the ZOO method as an attacker model to generate adversarial examples that were used for launching against the proposed RF classifier. Our goal was to assess the ability and susceptibility of the system after the adversarial training process. We modified this method and used it on one vector. The adversarial setting of the proposed ZOO model is explained as follows: the Adam optimizer with $\beta 1 = 0{:}8$; $\beta 2 = 0{:}899$ was applied to minimize the loss with a learning rate of 0.001 and $\in = 0.0000001$. The maximum epoch was set to 2000. The classification results of the proposed RF classifier after applying

the ZOO attack are shown in the third row of Table 5. Figures 4a–c and 5a–c summarize the confusion matrices (CMs) and AUC-ROC of all three parts of experiment I.

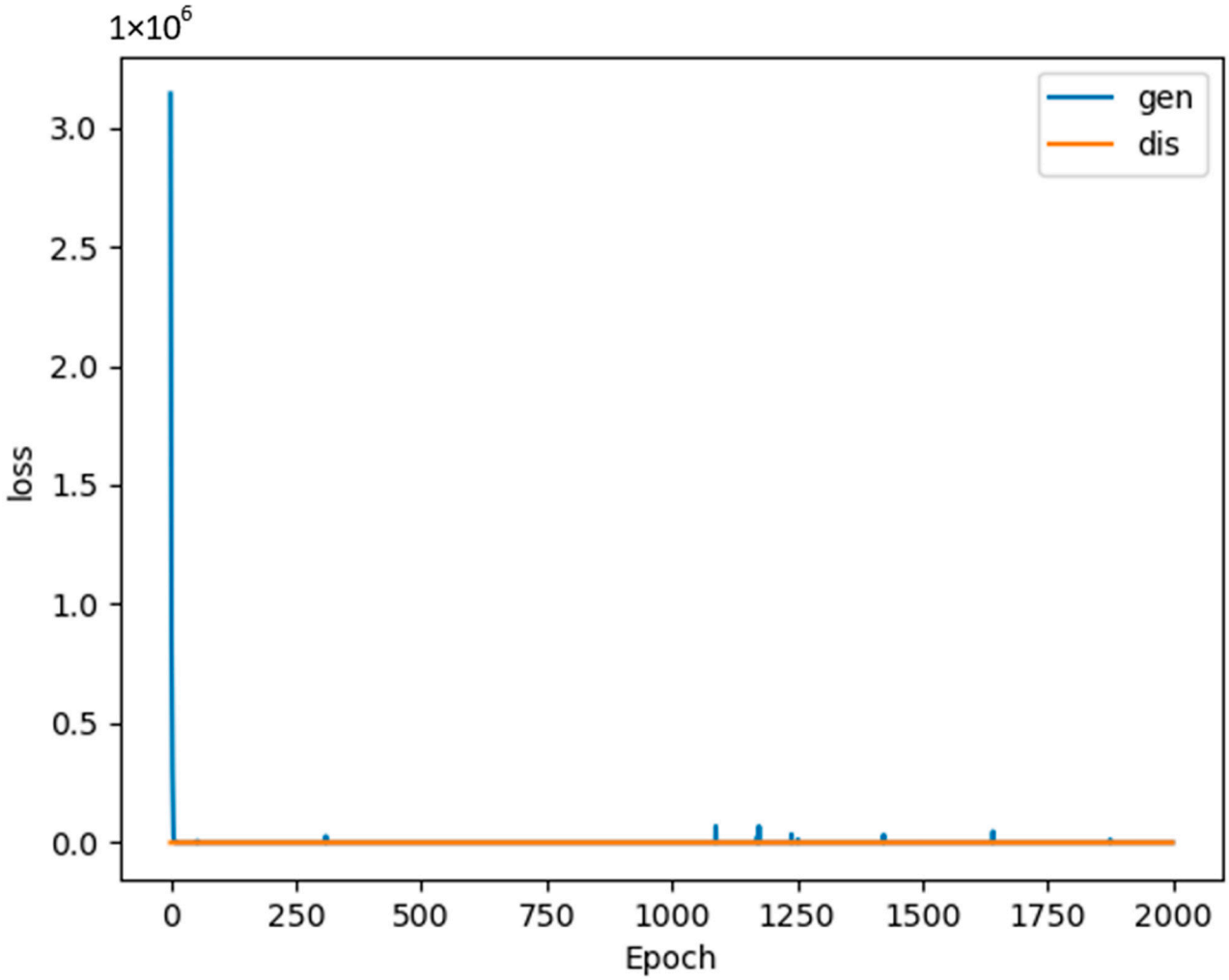
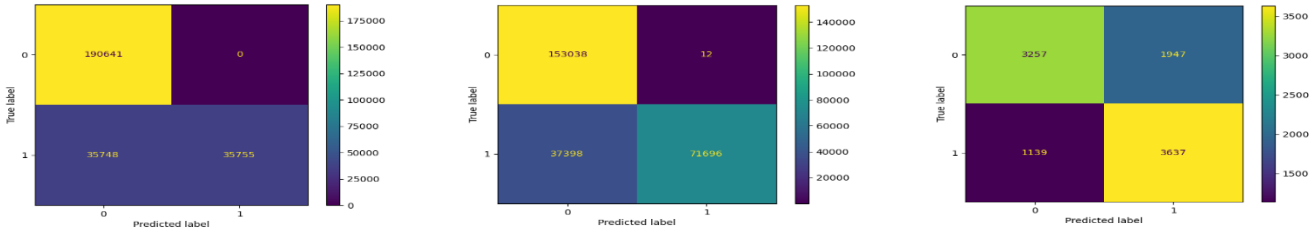


Figure 3. Losses between the generator and discriminator.

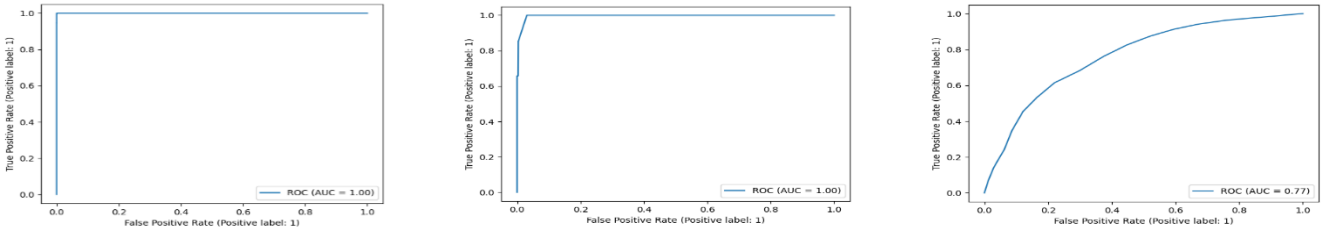


(**a**) RF before GAN (**b**) RF after GAN (**c**) RF after ZOO attack

Figure 4. CMs of the evaluation of RF on (**a**) clean data; (**b**) after adversarial training; (**c**) after launching a black-box attack in Experiment I.

Table 5 summarizes the classification results of the proposed random forest classifier (IDS) when 78 features were used. Even though RF before the GAN provided good results compared to RF after the GAN and ZOO, the performance could be improved by removing unrelated or redundant features from the dataset. Therefore, it was necessary to choose the best and most effective features from the dataset to improve the performance of our IDS classifier. As can be seen in Table 5, the generated data by the GAN resulted in a decline in accuracy. This may indicate that these generated samples were somehow symmetrical to the real data and, therefore, identification by the trained model is hard. Moreover,

the selected values of the parameters (e.g., epochs) help the GAN model produce strong adversarial examples. It is worth mentioning that even though the ROC was close to 1, as seen in Figure 5b, it did not help much in detecting infected samples generated by the GAN. Moreover, the accuracy of the RF was significantly decreased under the influence of the black-box attack (ZOO), because the samples generated by the ZOO attack could not be easily detected by the IDS. This confirms two things: (1) the process of the ZOO adversarial black-box attack generated new and strong infected instances that were hard for our model to detect, and this reduced the model's efficiency; (2) RF had the advantage of good accuracy on the original dataset with imbalanced high diminutions. Based on the aforementioned, the RF was successful because it did not have the problem of nominal data and did not overfit the data. Note that our proposal was still efficient at detecting unknown attacks as shown in Figure 5c.



(**a**) AUC-ROC of the natural model (**b**) AUC-ROC after GAN (**c**) AUC-ROC after the attack

Figure 5. AUC-ROCs of our RF classifier that describe the extent of the ML model in classifying in Experiment I.

5.2. Experiment II

The objective of this experiment was to train and test the proposed IDS model on fewer features using feature reduction methods, because we concluded from previous experiments that taking all of the features did not provide an optimal performance probably, as some of the features were unrelated and redundant. Accordingly, we turned to leveraging methods to minimize the features to obtain a better result. Specifically, we applied the RFE feature selection method to the original dataset to select only eight features. The RFE method was efficient in choosing the robust features and neglecting the weaker ones. Furthermore, it reduced the dependencies and the interlinear relationships that may exist in the dataset. The most important eight features described in Table 6 were utilized to retrain the proposed RF classifier. The first row of Table 7 illustrates the classification results of the RF classifier with the eight input features.

Table 6. Overview of the 8 selected features based the Rfe method.

Count	Feature Name
2	Protocol
13	Bwd Pkt Len Min
20	Flow IAT Max
22	Fwd IAT Tot
23	Fwd IAT Mean
30	Bwd IAT Max
54	Pkt Size Avg
78	Idle Min

Similar to experiment I, we applied the strategy of adversarial training based on the GAN to generate new samples based on the same eight selected features. We also tested the model when the ZOO model was used to attack the classifier. The proposed GAN was

modified to use eight features instead of 78. After 2000 epochs, the G's success in fooling the D can be seen in Figure 6. The proposed RF model was retrained with 240,000 adversarial samples that were generated by the GAN. The classification results of the proposed RF after applying the GAN are shown in the second row of Table 7. While the results after applying the ZOO attacker are shown in the third row of Table 7. Figures 7a–c and 8a–c summarize the confusion matrices and AUC-ROC of all three parts of experiment II.

Table 7. Evolution metrics of our IDS classifier (RF) including accuracy, precision, recall, F1-score, and MSE for experiment II.

RF	Accuracy	Precision	Recall	F1-Score	MSE
Before GAN	0.85	1.00	0.83	0.90	0.14
After GAN	0.905	0.81	0.98	0.88	0.09
After ZOO attack	0.487	1.00	0.48	0.64	0.51

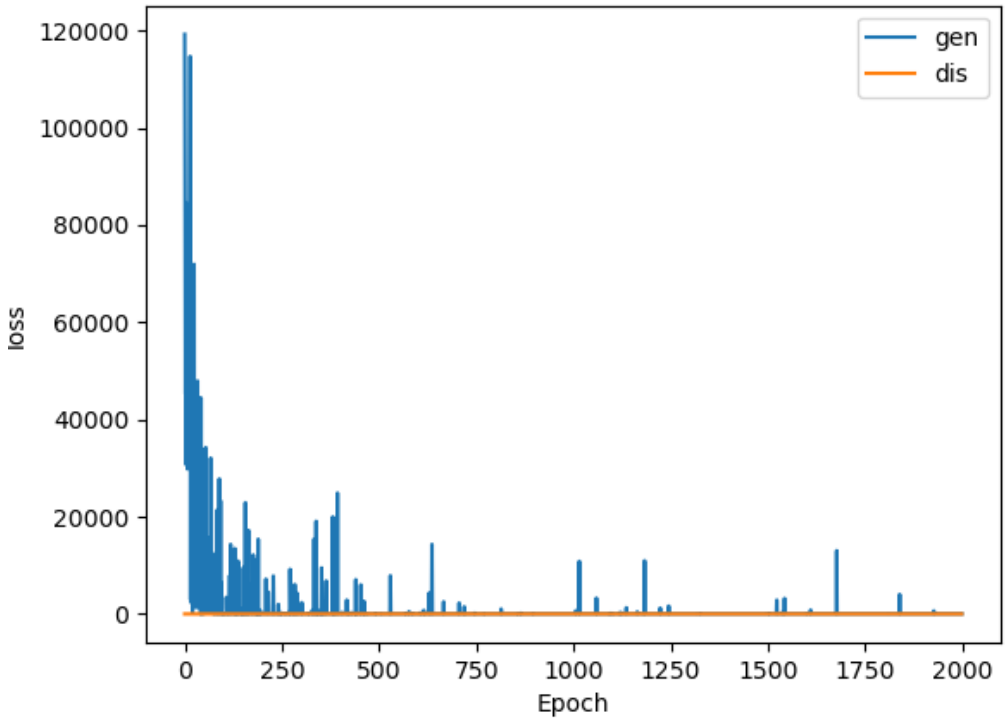
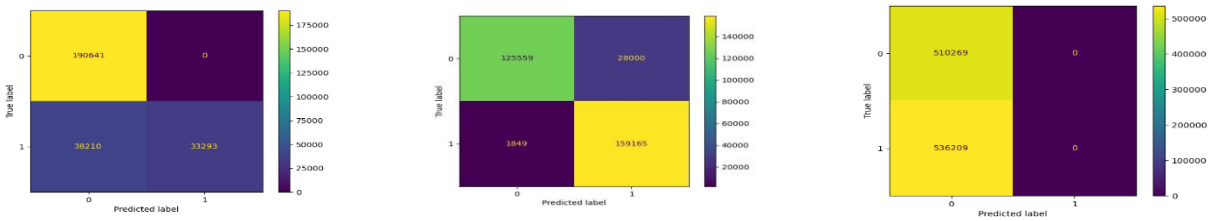


Figure 6. Computation between the generator and discriminator on 8 features.

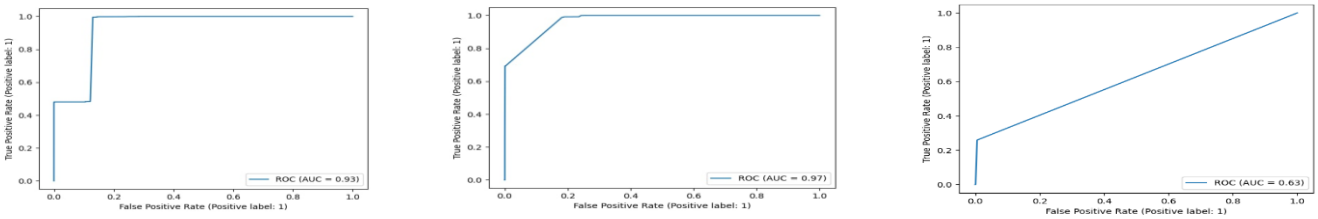


(**a**) RF before GAN (**b**) RF after GAN (**c**) RF after ZOO attack

Figure 7. CMs of the evaluation of the RF on (**a**) clean data; (**b**) after adversarial training; (**c**) after launching the black-box attack in experiment II.

According to the results in Table 7, the accuracy did not differ much from the accuracy of the random forest in experiment I. However, after applying the GAN and retraining the RF model, the accuracy improved compared to the GAN in experiment I. The MSE also improved from 0.14 to 0.09, as well as the FN in the confusion matrix. In addition,

an improvement in the ability of the classifier ROC curve from 0.93 to 0.97 is shown in Figure 8b. The performance of the RF model against the ZOO attacker in experiment II was lower than in experiment I. Where the accuracy dropped from 0.69 to 0.48. This suggests selecting more robust features from the original dataset by using a different feature selection method. Therefore, we performed the third experiment.



(**a**) AUC-ROC of the natural model (**b**) AUC-ROC after GAN (**c**) AUC-ROC after the attack

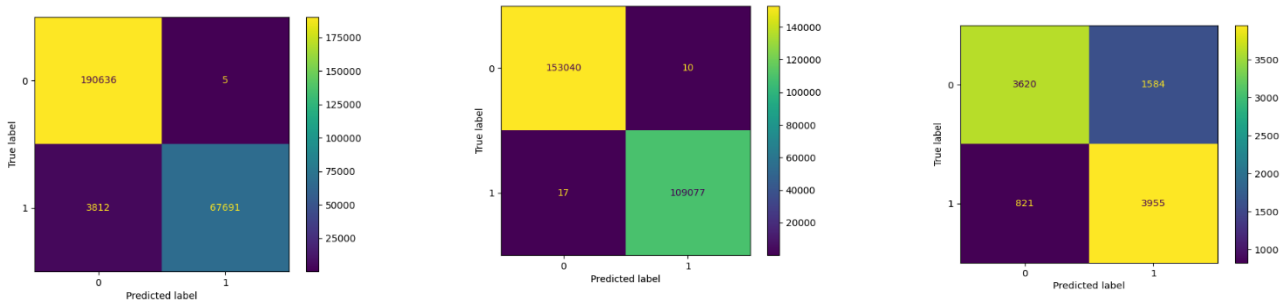
Figure 8. AUC-ROCs of our RF classifier that describe the extent the ML model in classifying in experiment II.

5.3. Experiment III

In this experiment, we used a different a feature selection method to improve the classifier's performance and handle the issue in experiment II when the ZOO attack was applied. Specifically, we used the PCA feature selection method to select the most efficient features from the original 78 features. The experimental results, illustrated in Table 8, indicate the RF classifier with the top 15 features selected by PCA with the highest variance. Note that in experiment II, we tried a different number of features chosen by Rfe, but the best accuracy was achieved when the number of selected features by Rfe was eight. The selected features by PCA were used to generate GAN-based adversarial examples. The performance of the RF classifier with GAN-based examples is shown in the second row of Table 8. Similar to the previous two experiments, we evaluated the RF classifier using the ZOO attack. The results are shown in the third row of Table 8. Figures 9a–c and 10a–c summarize the CMs and AUC-ROCs of all three parts of experiment II.

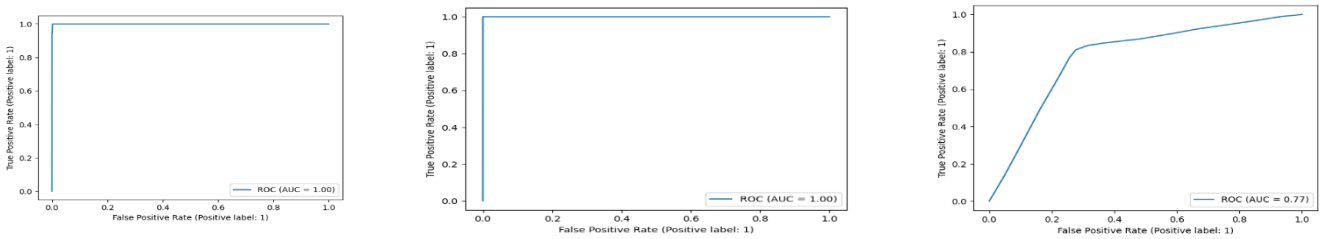
Table 8. Evolution metrics of our IDS classifier including accuracy, precision, recall, F1-score, and MSE for Experiment III.

RF	Accuracy	Precision	Recall	F1-Score	MSE
Before GAN	0.863	0.99	0.98	0.98	0.13
After GAN	0.999	0.99	0.99	0.99	0.0001
After ZOO attack	0.759	0.74	0.81	0.77	0.24



(**a**) RF before GAN (**b**) RF after GAN (**c**) RF after ZOO attack

Figure 9. CMs of the evaluation of the RF on (**a**) clean data; (**b**) after adversarial training; (**c**) after launching the black-box attack in experiment III.



(**a**) AUC-ROC of the natural model (**b**) AUC-ROC after GAN (**c**) AUC-ROC after the attack

Figure 10. AUC-ROCs of our RF classifier that describe the extent of the ML model in classifying in Experiment III.

The improvement in the RF classifier results shown in Table 8 indicate that the PCA method was more robust than the Rfe method. The RF with PCA accuracy was 0.863, while the RF with RFE accuracy was 0.85. The results also show that using GAN with the features selected by PCA achieved the highest accuracy of 99.9 and a significant decrease in the MSE from 0.014 to 0.0001. The excellent results of this experiment were not limited to the adversarial training but also extended to the black-box attack. Specifically, the system repelled the ZOO attack and obtained the highest accuracy of 0.759; the accuracies of experiment I and II when ZOO attack is applied were 0.69 and 0.487, respectively.

6. Comparison with Previous Studies

In this section, we compare our proposed ML-based IDS with state-of-the-art ML/DL-based IDS that only evaluate the model' performance without measuring the model resistance. Each of the prior work was implemented with different methods (e.g., some of them used a single model and others used multiple models) on same dataset. The results show that our proposal offers better accuracy compared to other existing works as shown in Table 9. It is worth mentioning that we did not compare our work with work that handled class imbalance issues by modifying the original dataset (e.g., the work in [47,49]), because it was beyond the scope of this paper, and it could be a complement to our proposed model. Although DL has been proven to be effective in the field of NIDS, our proposed ML model achieved better results. This might be because DL algorithms deal very well with complex tasks that require discovering relationships among a large number of different features. However, our experimental results show that reducing the number of features of our targeted task led to improving the overall accuracy. In addition, a recent study showed that using such techniques (i.e., using PCA to minimize the dimensionality of the dataset) with IDS reduced the performance of the model.

Table 9. Comparison with some related works on the CSE-CICIDS2018 dataset.

Authors	Year	Models	Acc
Usama, Asim et al. [20]	2019	LR	0.866
Amaizu, Nwakanma et al. [50]	2020	DNN	0.764
Fitni and Ramli [11]	2020	Ensemble model	0.988
Sawadogo, Bassolé et al. [51]	2021	CNN	0.975
Our proposed method	2022	RF	0.999

We compared our work with a related work [20] to measure the proposed model resistance against different attacks, where they used GAN as a defense method based on adversarial training. The test was conducted before and after applying GAN and ZOO attacks with and without applying adversarial training. Our model outperformed the work in [20] in all testing stages when these two attacks were applied as shown in Table 10. We obtained better results due to the technique used for feature selection and number of epochs. Specifically, we used PCA for feature selection instead of dividing the features into

functional and nonfunctional as done in [20]. Moreover, since the epochs used in [20] were only 100, this would not be enough to generate sufficiently strong fake samples for training the model effectively. To address this dilemma and generate strong samples, we increased the number of epochs to 2000.

Table 10. Comparison with other related work when different attacks are applied.

Technique	Before GAN	After Attack GAN	Attack GAN after AT	After ZOO Attack after AT
Reference [20]	0.89	0.6538	0.86	0.678
Our proposal	0.86	0.6683	0.87	0.759

7. Our Findings and Future Work

To identify abnormal and malicious behavior in networks, IDS have been used. Many ML techniques have been utilized to adopt different types of such systems to protect the network. Improving the system's performance and analyzing large amounts of network traffic requires providing robust and efficient systems to counter possible unknown attacks. To cover this issue, this paper proposed a theoretical-game-based approach to create a defensive system and train adversarially based on a GAN to ensure the system's reliability against black-box attacks. This system was then attacked to evaluate its strength and resilience in capturing the samples that were distorted by the adversary. This process used a three-stage framework for each experiment. All of the features were used, and then feature selection methods were performed to determine the right features for good results with less complexity and execution time. Evaluation of these experiences was conducted on the recent CSE-CICIDS2018 dataset. The outcomes showed that the quality of the data that our IDS trained on and the features that were selected as well as the rate of perturbed samples by the attacker were factors that influenced the results of the system. The PCA method was the best with lower implementation times compared to other trials. The use of a GAN as a defense technique is a good decision to protect networks from modern attacks. For future actions, we recommend that a GAN can be used in security domains other than image and encryption areas to train the system to defend itself against adverse attack scenarios. It can also be applied to deep learning techniques to determine their effectiveness with high-dimensional data. Proposing new defense methodologies against such attacks is necessary. While we have focused on the problem of binary classification in this work, it is important to extend this research to the problem of multiclass classification to classify separate types of attacks, and this will be one of our key future works. This could be important in reducing the complexity and execution time of the ML model.

Author Contributions: Investigation, M.A.; Methodology, S.A. and Q.A.; Project administration, Q.A.; Resources, M.M.H.; Supervision, Q.A. and J.-S.Y.; Validation, S.A.; Visualization, M.M.H.; Writing—review & editing, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yihunie, F.; Abdelfattah, E.; Regmi, A. Applying machine learning to anomaly-based intrusion detection systems. In Proceedings of the 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 3 May 2019; pp. 1–5.
2. Ahmad, S.; Arif, F.; Zabeehullah, Z.; Iltaf, N. Novel Approach Using Deep Learning for Intrusion Detection and Classification of the Network Traffic. In Proceedings of the 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Tunis, Tunisia, 22–24 June 2020; pp. 1–6.
3. Alasad, Q.; Lin, J.; Yuan, J.-S.; Fan, D.; Awad, A. Resilient and secure hardware devices using ASL. *ACM J. Emerg. Technol. Comput. Syst. (JETC)* **2021**, *17*, 1–26. [CrossRef]
4. Shin, S.; Lee, I.; Choi, C. Anomaly dataset augmentation using the sequence generative models. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1143–1148.
5. Sarvari, S.; Sani, N.F.M.; Hanapi, Z.M.; Abdullah, M.T. An efficient anomaly intrusion detection method with feature selection and evolutionary neural network. *IEEE Access* **2020**, *8*, 70651–70663. [CrossRef]
6. Alasad, Q.; Yuan, J.-S.; Subramanyan, P. Strong logic obfuscation with low overhead against IC reverse engineering attacks. *ACM Trans. Des. Autom. Electron. Syst. (TODAES)* **2020**, *25*, 1–31. [CrossRef]
7. Caminero, G.; Lopez-Martin, M.; Carro, B. Adversarial environment reinforcement learning algorithm for intrusion detection. *Comput. Netw.* **2019**, *159*, 96–109. [CrossRef]
8. Liao, H.-J.; Lin, C.-H.R.; Lin, Y.-C.; Tung, K.-Y. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **2013**, *36*, 16–24. [CrossRef]
9. Serinelli, B.M.; Collen, A.; Nijdam, N.A. Training guidance with KDD cup 1999 and NSL-KDD data sets of ANIDINR: Anomaly-based network intrusion detection system. *Procedia Comput. Sci.* **2020**, *175*, 560–565. [CrossRef]
10. Sah, G.; Banerjee, S. Feature Reduction and Classifications Techniques for Intrusion Detection System. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 28–30 July 2020; pp. 1543–1547.
11. Fitni, Q.R.S.; Ramli, K. Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems. In Proceedings of the 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia, 7–8 July 2020; pp. 118–124.
12. Alatwi, H.A.; Aldweesh, A. Adversarial Black-Box Attacks Against Network Intrusion Detection Systems: A Survey. In Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 10–13 May 2021; pp. 0034–0040.
13. Ayub, M.A.; Johnson, W.A.; Talbert, D.A.; Siraj, A. Model evasion attack on intrusion detection systems using adversarial machine learning. In Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 18–20 May 2020; pp. 1–6.
14. Apruzzese, G.; Andreolini, M.; Colajanni, M.; Marchetti, M. Hardening random forest cyber detectors against adversarial attacks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 427–439. [CrossRef]
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 2, pp. 2672–2680.
16. Hu, W.; Tan, Y. Generating adversarial malware examples for black-box attacks based on GAN. *arXiv* **2017**, arXiv:1702.05983.
17. Salem, M.; Taheri, S.; Yuan, J.S. Anomaly generation using generative adversarial networks in host-based intrusion detection. In Proceedings of the 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 8–10 November 2018; pp. 683–687.
18. Dutta, I.K.; Ghosh, B.; Carlson, A.; Totaro, M.; Bayoumi, M. Generative Adversarial Networks in Security: A Survey. In Proceedings of the 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 28–31 October 2020; pp. 0399–0405.
19. Silva, S.H.; Najafirad, P. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv* **2020**, arXiv:2007.00753.
20. Usama, M.; Asim, M.; Latif, S.; Qadir, J. Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 78–83.
21. Piplai, A.; Chukkapalli, S.S.L.; Joshi, A. NAttack! Adversarial Attacks to bypass a GAN based classifier trained to detect Network intrusion. In Proceedings of the 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), Baltimore, MD, USA, 25–27 May 2020; pp. 49–54.
22. Alhajjar, E.; Maxwell, P.; Bastian, N. Adversarial machine learning in network intrusion detection systems. *Expert Syst. Appl.* **2021**, *186*, 115782. [CrossRef]
23. Rigaki, M. Adversarial Deep Learning Against Intrusion Detection Classifiers. Master's Thesis, Luleå University of Technology, Luleå, Sweden, 2017.

24. Khamis, R.A.; Shafiq, M.O.; Matrawy, A. Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–7.
25. McCarthy, A.; Andriotis, P.; Ghadafi, E.; Legg, P. Feature Vulnerability and Robustness Assessment against Adversarial Machine Learning Attacks. In Proceedings of the 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 14–18 June 2021; pp. 1–8.
26. Zhao, S.; Li, J.; Wang, J.; Zhang, Z.; Zhu, L.; Zhang, Y. attackGAN: Adversarial Attack against Black-box IDS using Generative Adversarial Networks. *Procedia Comput. Sci.* **2021**, *187*, 128–133. [CrossRef]
27. Alatwi, H.A.; Morisset, C. Adversarial Machine Learning In Network Intrusion Detection Domain: A Systematic Review. *arXiv* **2021**, arXiv:2112.03315.
28. Waskle, S.; Parashar, L.; Singh, U. Intrusion detection system using PCA with random forest approach. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; pp. 803–808.
29. Kanimozhi, V.; Jacob, T.P. Artificial Intelligence outflanks all other machine learning classifiers in Network Intrusion Detection System on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. *ICT Express* **2021**, *7*, 366–370. [CrossRef]
30. Mahbod, A. Towards Improvement of Automated Segmentation and Classification of Tissues and Nuclei in Microscopic Images Using Deep Learning Approaches. Ph.D. Thesis, Medical University of Vienna, Wien, Austria, 2019.
31. Bourou, S.; el Saer, A.; Velivassaki, T.-H.; Voulkidis, A.; Zahariadis, T. A review of tabular data synthesis using gans on an ids dataset. *Information* **2021**, *12*, 375. [CrossRef]
32. Feng, C.; Shang, Y.; Jincheng, H.; Bo, X. Few features attack to fool machine learning models through mask-based GAN. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
33. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 214–223.
34. Chauhan, R.; Heydari, S.S. Polymorphic Adversarial DDoS attack on IDS using GAN. In Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 20–22 October 2020; pp. 1–6.
35. Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 15–26.
36. Yang, K.; Liu, J.; Zhang, C.; Fang, Y. Adversarial examples against the deep learning based network intrusion detection systems. In Proceedings of the MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, USA, 29–31 October 2018; pp. 559–564.
37. Nskh, P.; Varma, M.N.; Naik, R.R. Principle component analysis based intrusion detection system using support vector machine. In Proceedings of the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 20–21 May 2016; pp. 1344–1350.
38. Brauckhoff, D.; Salamatian, K.; May, M. Applying PCA for traffic anomaly detection: Problems and solutions. In Proceedings of the IEEE INFOCOM 2009, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 2866–2870.
39. Apostolidis, K.D.; Papakostas, G.A. A survey on adversarial deep learning robustness in medical image analysis. *Electronics* **2021**, *10*, 2132. [CrossRef]
40. Tian, J. Adversarial vulnerability of deep neural network-based gait event detection: A comparative study using accelerometer-based data. *Biomed. Signal Processing Control.* **2022**, *73*, 103429. [CrossRef]
41. Lin, Z.; Shi, Y.; Xue, Z. Idsgan: Generative adversarial networks for attack generation against intrusion detection. *arXiv* **2018**, arXiv:1809.02077.
42. Chen, J.; Wu, D.; Zhao, Y.; Sharma, N.; Blumenstein, M.; Yu, S. Fooling intrusion detection systems using adversarially autoencoder. *Digit. Commun. Netw.* **2021**, *7*, 453–460. [CrossRef]
43. Mirza, A.H. Computer network intrusion detection using various classifiers and ensemble learning. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.
44. Zenati, H.; Romain, M.; Foo, C.-S.; Lecouat, B.; Chandrasekhar, V. Adversarially learned anomaly detection. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 727–736.
45. Punitha, A.; Vinodha, S.; Karthika, R.; Deepika, R. A Feature Reduction Intrusion Detection System using Genetic Algorithm. In Proceedings of the 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 29–30 March 2019; pp. 1–7.
46. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* **2018**, *1*, 108–116.
47. Liu, L.; Wang, P.; Lin, J.; Liu, L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access* **2020**, *9*, 7550–7563. [CrossRef]
48. Kilincer, I.F.; Ertam, F.; Sengur, A. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Comput. Netw.* **2021**, *188*, 107840. [CrossRef]
49. Karatas, G.; Demir, O.; Sahingoz, O.K. Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE Access* **2020**, *8*, 32150–32162. [CrossRef]

50. Amaizu, G.C.; Nwakanma, C.I.; Lee, J.-M.; Kim, D.-S. Investigating Network Intrusion Detection Datasets Using Machine Learning. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 21–23 October 2020; pp. 1325–1328.
51. Sawadogo, L.M.; Bassolé, D.; Koala, G.; Sié, O. Intrusions Detection and Classification Using Deep Learning Approach. In Proceedings of the International Conference on Research in Computer Science and its Applications, Virtual, 17–19 June 2021; Springer: Cham, Switzerland, 2021; pp. 40–51.