ASJET

Al-Salam Journal for Engineering and Technology
Journal Homepage:
http://journal.alsalam.edu.iq/index.php/ajest
e-ISSN: 2790-4822 p-ISSN: 2958-0862

# Network Intrusion Detection Systems: Machine Learning-Based Attack and Remedy Strategies – A Review

**Meaad Ahmed[1] * and Qutaiba Abdullah[2]**

[1]Department of Computer Science, Tikrit University.
[2]Department of Cybersecurity, Tikrit University.

*Corresponding Author: Meaad Ahmed

DOI: https://doi.org/10.55145/ajest.2025.04.02.002
Received March 2025; Accepted May 2025; Available online August 2025

**ABSTRACT:** Network Intrusion Detection Systems (NIDS) have played an important role in protecting computer networks against illegal or unauthorized access and many cyberattacks. Given the advancement of machine learning (ML) approaches, NIDS have become more effective in detecting various complex anomalies in computer networks. However, the increasing complexity in adversarial attacks (AAs) poses a significant challenge to such systems. Cyberattacks are estimated to cost approximately $10.5 trillion annually by 2025, and this encourages researchers to improve and develop ML-based NIDS in order to address adversarial vulnerabilities in these systems to remain the systems resilient to modern attacks. In this review, we reveal the weaknesses of ML-based IDS to AAs in which many different attack techniques, such as evasion, poisoning, and generative adversarial networks (GAN) have been included. Also, this study presents and evaluates current possible defensive approaches against AAs, including but not limited to anomaly detection approaches, adversarial training, and feature selection. We also provide a comparative analysis of different ML models used in NIDS in order to further evaluate the system's susceptibility to sophisticated AAs. A discussion on future work to improve ML-based NIDS resilience against sophisticated attacks is also given.

**Keywords:** Machine learning, adversarial training, cybersecurity

## 1. INTRODUCTION

Nowadays, attackers can launch a wide range of security attacks against computer networks using a number of tactics due to the networks' extensive expansion and the new, growing applications on them [1]. These include unauthorized access to data and systems. Cyberattacks have increased significantly in diversity and scope, and they could deactivate customer services and enable illegal access to sensitive information [2]. The effectiveness of cyberattacks is significant not only for cultural safety and national security but also for national economic and financial institutions. Thus, it is important to prevent cyberattacks from external and internal sources and commercial and public systems [3]. Network operators can precisely identify security threats thanks to intrusion detection systems, which are essential to the network defense process [4]. According to a number of studies, network security is a critical concern at the moment, and intrusion detection systems (IDS) have been created to safeguard network security [5].

IDS are categorized as either host-based IDS (HIDS) or network-based IDS (NIDS) depending on where they are deployed. One device in the network houses a HIDS, which keeps an eye on its condition to spot any unusual activity [6]. NIDS, the first line of defense, is in charge of keeping an eye on network activity to identify threats or unusual activities that might be a component of an assault [7].

IDSs could be either signature-based or anomaly-based [8] in order to exploit indicators based on signatures that were previously taken from known attacks. For every new attack, a signature is created. Because of the largely growing variety and quantity of attacks, it is therefore expensive to keep an updated list of signatures. Unlike malicious behavior, anomaly-based approaches simulate typical network behavior. These methods have a significant false alarm rate since they can identify new normal behavior as malicious, even though they are capable of detecting new attacks [4] [9].

Anomaly detection has made extensive use of ML techniques in recent years. Though they solve some problems, machine learning-based security solutions also bring about new ones, like adversarial machine learning attacks. DL

models have been shown to be susceptible to "adversarial examples" (AEs), which occur when a carefully constructed data instance can cause the model to classify data incorrectly [7]. In order to identify possible network threats, recent developments in NIDS based on anomaly have been highly boosted by Deep Learning (DL) models. The main factor contributing to DL algorithms' success is their capacity to obtain highly non-linear and abstract representations by making full use of vast amounts of data [6]. It is worth mentioning that some ML based NIDSs can achieve high accuracy in attack detection rate but typically at the cost of increasing the computational complexity and execution time. However, lightweight detection systems can be implemented with lower execution time and cost via eliminating less important features. While such method can refine the efficiency of the system, it will reduce the accuracy slightly [10]. In particular, we examine white-box and gray/black-box attacks. White-box attacks give the adversary total access to all information about the ML-based NIDS, unlike gray/black-box attacks, in which the attacker has little to no knowledge of the system [11].

Given the increasingly expansion in networking systems and the incremental daily operations and tasks, the security threats and AAs have increased significantly and become difficult to be detected. Accordingly, NIDS has been presented to mitigate cyber threats. Unfortunately, conventional NIDS, e.g., signature-based approaches, fail to handle and resolve systems involved in sophisticated attacks with large volumes of data. ML-based NIDSs have been considered as a promising solution in order to improve the accuracy and the performance of the system in detecting anomalous events. Unfortunately, ML-based systems are shown to be subject to AAs in which malicious attacks can manipulate the primary input data in order to deceive the detection systems. In this work, we target to offer a complete analysis of AA techniques against ML-based NIDS and explore different possible defensive approaches. The main motivation of this review is to highlight the weaknesses and strengths of current NIDSs, evaluate the effectiveness of different adversarial methods, and possibly present mitigation techniques that can be used to refine the system's resilience. Through carefully analyzing game theory between attack and defense techniques, this work highlights current challenges and contributes insight for stronger frameworks against sophisticated threats.

## 2. BACKGROUND

Machine learning (ML) has been used to identify zero-day attacks and network stream anomalies that are hard to identify with conventional signature-based techniques. Finding patterns that are predictive and generalizable is the primary objective of machine learning [12]. The learning method used by the machine learning algorithm is categorized into three types as shown in Figure 1, and below is a detailed explanation of each type:

• Supervised Learning: This approach provides models for regression and classification using labeled data. These models can then predict outcomes for new, unlabeled data. Logistic Regression (LR) is a widely used supervised method for binary classification (where outcomes are labeled as 0 or 1). In LR, the log odds of the binary outcome are modeled as a linear function of the input variables, as defined by its underlying logistic function [12, 13].

• Unsupervised Learning: Unlike supervised learning, this method deals with unlabeled data, identifying hidden patterns or groupings without predefined outcomes. Examples include clustering, e.g., K-means, and dimensionality reduction, e.g., PCA. The second bullet point you provided mistakenly repeats supervised learning and LR, which do not apply here. If referring to a different concept, it should clarify unsupervised techniques [14].

• Reinforcement Learning: is a method of learning where an artificial intelligence (AI) agent uses trial-and-error to interact with its environment and then learns the best behavioral strategy based on the reward signals it has received from those interactions. The ability of RL to be applied to other scientific and engineering domains is among its most advantageous features [15].

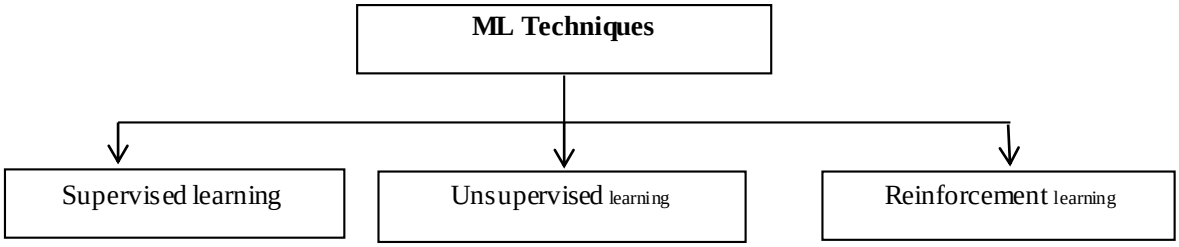


**FIGURE 2 Main ML approaches**

•Adversarial Machine Learning (AML) involves altering data in order to fool the machine learning model, producing the attacker's desired erroneous detection results. The term Adversarial Machine Learning (AML) was first used after researchers in the field of computer vision identified some blind spots in image classifiers that these adversarial samples used to trick the model [16].

Adversarial Attack Taxonomy:

1- Knowledge: In the adversarial threat model, the "knowledge" part refers to how much the attacker understands about the machine learning system. This knowledge can be categorized into three types: White-box (full knowledge of the system), Gray-box (partial knowledge), or Black-box (very little or no knowledge) attacks.

White-box attacks: in this case, the assailant is fully aware and has access to the internal parameters of the learning algorithm, the learned model, the training data, and the parameters that have been employed to train the model. An adversary with the precise information known by the developer or owner of the ML model under attack is represented by a white-box attack in most AA applications in the real world [11] [17].

Gray-box attack: The adversary has enough information to attack the machine learning system and make it fail, even though he or she lacks the precise knowledge that the model's creator possesses [11] [18].

Black-box attack: it is assumed that the machine learning system is completely unknown to the adversary. The adversary in this kind of attack is unaware of the learned model or the learning algorithm. One could argue that an attack that is truly black-box is impossible [11] [19].

2- Strategy: The assailant's strategy determines when and how they carry out their attack, with three approaches: Evasion (tricking the model after training), Poisoning (corrupting the training data), and Oracle (probing the model to extract information).

•Evasion attacks: During the testing or inference stage, it is also referred to as an attack at decision time, also known as exploratory attack. After the machine learning model has been learned, the attacker wants to skew its judgment. An optimization problem is usually calculated arithmetically in evasion attacks. Finding a small perturbation sigma that would raise the loss function is the aim of the optimization problem. At that point, the loss function would have changed sufficiently to cause the machine learning model to make an incorrect prediction. There are two types of evasion attacks: gradient-based attacks and gradient-free attacks [11] [20].

•Poisoning attacks: Poisoning attacks occur when an attacker adds malicious data to the training set to trick the ML model. The attacker inserts malicious examples that look like normal training data but are carefully designed to make the model learn wrong patterns. This changes how the model works without altering the original data labels or features. The result is a model that makes incorrect predictions when used [11] [21].

•Oracle attacks: occur when an attacker first uses a poisoning attack to access a model's API and create a corrupted version. This infected model keeps most of the original model's capabilities but can be used for harmful purposes like evasion attacks. Oracle attacks come in three types: (1) Extraction attacks steal the model's design (weights, hyperparameters) by studying its outputs [11] [22]; (2) Inference attacks let attackers identify specific data patterns from the training set; and (3) Inversion attacks where attackers try to recreate the original training data [23].

## 3. ATTACK AND DEFENSE TECHNIQUES ON ML-BASED NIDS

This main section delves into many articles that have leveraged AML in the NIDS, in which the research's goal is augmented to reveal current attack and defense techniques on NIDS. More specifically, we review current attack and defense studies on ML-based NIDS. A search using the keywords to elaborate this survey, ((adversarial OR AML OR "adversarial attack") AND ("intrusion detection systems" OR IDS OR NIDS OR "network intrusion detection systems") AND ("machine learning" OR ML)) has been conducted on different academic databases. Figure 2 shows the attack and defense techniques that will be discussed in detail in the incoming subsections.

The gathered investigations were examined and afterward grouped based on their purpose into two gatherings. The reviewed research falls into two main parts: (1) Studies on proposing methods to generate malicious network traffic that can bypass ML-based NIDS detection, along with evaluations of how well these systems withstand standard adversarial examples; and (2) Research concentrated on defenses and testing existing protection methods to strengthen NIDS against adversarial ML attacks.

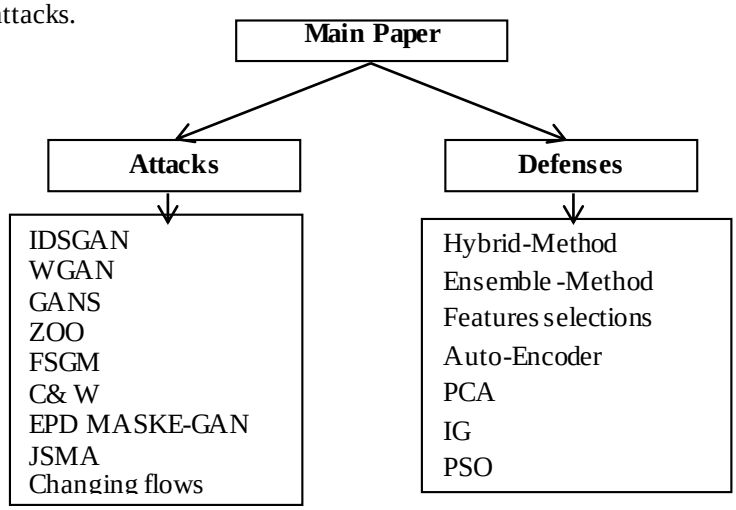


**FIGURE 2 Sections of research studies**

## 3.1 ATTACKS AGAINST ML-BASED NIDS

In this section, we discuss AML methodologies that have been widely utilized to fool machine learning models. Such methods use strategies and skills designed by attackers to trick the model and cause incorrect decisions.

### 3.1.1 CREATES AES TO ATTACK ML-BASED NIDS MODELS

Many different approaches have been proposed to generate strong attacks, namely Adversarial Examples (AEs), to deceive ML models, such as Generative Adversarial Networks (GANs), which is a type of ML model introduced in 2014. GAN can be used to generate new data instances from the original dataset. This generated data can be leveraged as attack or defense mechanisms on the NIDS model. GAN consists of two components; the first one called the generator (G) is used to create new instances from the original one, and the discriminator (D) is employed to distinguish samples whether they are from the actual dataset or from the generator.

A technique, called IDSGAN, was introduced in order to generate adversarial instances to fool the detection system through making incorrect outputs. Wasserstein GAN (WGAN) is integrated with IDSGAN to produce malicious samples that are hard for NIDS to detect. WGAN is comprised of 3 components: a black-box NIDS, a discriminator (D), and a generator (G). G produces infected instances by processing network traffic containing both random noise and attack records. The black-box NIDS is employed to identify benign records from the malicious ones through creating output labels that serve as target references for the model, where the discriminator employs the reference labels to train and guide the black-box NIDS [24]. The authors NSL-KDD dataset to evaluate different classifiers, including but not limited to Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR). According to the testing outcomes, the accuracy of all models ranged from 1% to 70%. It has been proven that the features that are updated to make the attack appear more accurate are limited to non-functional characteristics, which do not reflect the function associated with the type of attack, and therefore these features might be tweaked or preserved. In [25], the authors deployed polymorphic Distributed Denial of Service (DDoS) assaults with GAN to evaluate the capability of NIDS in recognizing adversarial samples and improve the training for the purpose of increasing GAN resistance. The polymorphic DDoS attacks were formed by combining previously created AEs with modified DDoS threat profile features, such as the count of features and feature switching. Then, they were fed to the GAN model. The outcomes indicated that the attack profile on a regular basis cannot be detected by NIDS, while keeping the model at a low false-positive ratio. In addition, it has been turned out that the defense systems that relied on incremental training were weak to unknown attacks. The experiment was assessed using RF. DT, LR, and NB classifiers with the CICIDS2017 dataset. In this study, switching features and changing the number of features yield detection rates of 5.23% and 3.89%, respectively. In [26], a technique, named Anti-Intrusion Detection Auto Encoder (AIDAE), has been introduced to create characteristics to deactivate NIDS. The proposed technique involved an encoder to convert a portion of the features to embedding space, and a large number of decoders in order to collect the continuous and discrete features. Also, GAN has been used to preserve the prior distribution of the embedded space. The proposed technique learns the typical feature distributed to create potent adversarial samples, and there is no need for NIDS in the training phase. Besides, the framework kept the relationship among the generated continuous and discrete features without change. The experimental setup involved three different datasets: CICIDS2017, NSL-KDD, and UNSW-NB15, with six different models: CNN+LSTM, LR, K-NN, RF, DT, and AdaBoost. Experimental outcomes showed that the created features are able to weaken the NIDS. This ensured that strong defensive techniques should be considered to prevent these threats.

Usama et al. proposed an adversarial ML attack utilizing GANs to generate adversarial variations in the traffic network. The proposal aims to evade the black-box NIDS. As a defensive approach, GANs have been leveraged in the training operation to make the classifier resist to adversarial attacks. In order to assess GAN based NIDS, the KDD99 dataset was leveraged. Several techniques have been selected as black-box NIDS to perform many tests and demonstrate the efficacy of the suggested GAN-based adversarial attack and training phase. The experiment results demonstrated that the GB achieved the best accuracy, 65.38 %, and SVM offered the lowest one, 43.44 %. After carefully training, the model was clearly enhanced, in which 86.64% accuracy was for LR and 79.31% was KNN [27].

To keep the model's performance under a specific training dataset, active learning can be used as a solution. Active learning is one of the family methods that can be used to improve training data collection in order to construct a minimal size training dataset and can still produce adequate performance. Shu et al. combined active learning with GANs to create strong adversarial examples to fool ML-based Black Box NIDS and evaluated these attacks on NIDS models. This technique was called Generative Adversarial Active Learning (Gen-AAL). It outperformed current adversarial threat techniques that necessitate a large amount of training dataset or need full knowledge about the NIDS model itself. Therefore, this approach required only a small set of queries to the targeted model for labeled instances to train the GAN and assumed no prior knowledge of the NIDS model. The proposed technique has been evaluated on the CICIDS2017 dataset using a gradient boosted decision tree NIDS. The results show that Gen-AAL is able to evade the NIDS classifier with a positive outcome of 98.86% only via using 25 labeled examples during the training phase [28].

In [29], an improved model called “attack GAN” for adversarial attacks has been proposed, and a new loss function is designed to accomplish a successful attack against the black-box NIDS. This work investigated the

effectiveness of GAN-based attacks and alternative attack methods employing NSL-KDD dataset. The authors proved that the impact of five ML/DL methods is significantly decreased under the introduced attack. It has been shown that potential of launching such attacks occurs a high harmful effect without the need for any secret details about the classification models. At last, when comparing GAN attacks to others, it has been discovered that the adversarial samples generated by GAN attack can correctly evade NIDS with 87.18% and have a higher attack rate of success than other existing adversarial attack methodologies, e.g., Fast Gradient Sign Method (FGSM) 17.76%, Project Gradient Descent (PGD) 29.78%, and CW attack 21.25%. Such analysis elucidates that the generated samples of the GAN assault are more powerful and successful. Yang et al. [30] performed three assaults leveraging NSL-KDD with DNN classifier. In this attack model, it is assumed that an attacker is aware of the feature extraction rather than the model: 1) A score-based approach (ZOO) that estimated symmetric difference quotients with gradients [31], along with a WGAN-based attack [32]; 2) A transfer-based assault that used adversarial transferability [33] to train a substitute classifier using the NSL-KDD dataset with a different architecture. However, the generated characteristics cannot be used to carry out the real attack because the NSL-KDD dataset is rather old.

In [34], authors introduced an approach, named Few-Features-Assault-GAN (FFA-GAN), to generate AEs to fool ML models. The proposed method is much faster than non-learning adversarial attack approaches, while also performing better than GAN-based attacks at preserving non-zero features. The improved performance comes from two key techniques: (1) using a masking strategy in the GAN's generator to limit how many features are changed in the perturbations (keeping alterations minimal), and (2) adjusting the loss function weights during training - first exploring a wide range of possibilities, then gradually focusing on a narrower range for refinement. The study used two types of benchmark datasets: (1) Structured datasets (CIC-IDS 2017 and KDD-Cup) which have smaller data sizes, and (2) Unstructured datasets (MNIST and CIFAR-10) which have much larger dimensions. These were selected to test the technique's performance on different data scales - with the structured network traffic datasets evaluating smaller-scale scenarios, and the high-dimensional image datasets assessing performance on larger, more complex data. The experiments on these datasets reveal that the FFA-GAN technique performs well in many classifiers. However, rather than changing the weights of the losses in a large range first and then minimizing the range subsequently, optimization techniques, like the population-based training method could be leveraged to optimize the weights of the losses throughout the training phase.

From the list of the given studies based on the GAN model in the field of network security, it has been indicated that the GAN is a very powerful technology in many scenarios related to NIDS. GAN models can be used to attack and compromise these detection systems by breaking the system or making it unsafe environment during the normal operations. Since GAN is a strong technology and can be used in many security applications, researchers should pay more attention to taking into account GAN attacks.

The Hierarchical Adversarial Attack (HAA) is known as a cutting-edge adversarial attack creating technique that was first presented by Zhou et al. [35]. The approach implemented a complex, multi-level black-box attack technique targeting Graph Neural Network (GNN)-based NIDS in IoT environments, while operating within strict computational/resource constraints. The method maintained predefined budget limitations throughout its execution of this layered attack plan against the GNN security systems. The authors' strategy involved identifying and changing important feature complements with the fewest possible disruptions, using a saliency map method to generate adversarial instances. The nodes with a higher attack vulnerability are then given priority using Random Walk with Restart (RWR) based on a hierarchical node selection method. The UNSW-SOSR2019 dataset [36] was used. The Jumping Knowledge Networks (JK-Net) [37] and Graph Convolutional Network (GCN) [38] were the two standard GNN models that the authors evaluated in their HAA method. When compromising the targeted GNN models, the authors also took into account three techniques: Greedily Corrected Random Walk (GCRW) [39], Resistive Switching Memory (RSM) [40], and Improved Random Walk with Restart (iRWR) [41]. The findings demonstrated that the HAA strategy, which is based on adversarial attacks, can reduce the accuracy of GNN classifiers by over 30%. The authors did not, however, investigate how well their HAA approach worked when an adversarial defense strategy was used.

The NIDSGAN proposed by Zolbayar et al. [42] has specified domain limitations. For more evasive attacks, they added more words to the loss function. The model used two key components in its optimization: (1) a primary loss function that reduced the differences between adversarial and original traffic features (making attacks harder to detect), combined with (2) a standard adversarial loss that trained the discriminator to treat malicious traffic as benign. Together, these ensured the adversarial examples remained both effective and stealthy by closely mimicking normal traffic patterns while bypassing detection. By introducing different quantities of contaminated instances into the training datasets at different rates, a poisoning attack technique was specifically developed for deep learning models. The impact of the assault on model performance was then investigated. The ability of NIDSs based on deep learning (DL) models to withstand unlimited white-box attacks was evaluated.

The findings demonstrated that raising the injection percentage and random amplified distribution from 1% to 50% could have a minor effect on the overall system's performance, with accuracy reaching 0.93. Nevertheless, the other measurement results— MSE, FPR, and PPV—scored 0.67, 0.29, and 0.082, respectively, underscoring the significant impact of data manipulation on the DL classifier. The findings demonstrated that the detection system was vulnerable to poisoning assaults, guaranteeing that appropriate defensive strategies were needed to mitigate such potent

attacks. According to the research, which was backed up by testing results, the generated poisoned data can significantly impact the mode's performance and are challenging to identify [43]. In [44], four attacks, ZOO, GAN, DeepFool, and KDE models, have been reevaluated using the datasets: CICIDS2019, CICIDS2018, and ADFA-LD. The evaluation was done on the trained ML based NIDS classifier in order to measure which attack is the most powerful. The experimental results ensured that DeepFool is the most promising attack.

### 3.1.2 CREATE ATTACKS WITH ANOTHER ADVERSARIAL ML TECHNIQUES

Moving to other types of attack scenarios, some examined studies come with efforts to attack IDS. The authors in [45] tried to compromise the IDS based on an "ANN" classifier by using AEs. These AEs are created via using the Fast Gradient Sign Method (FSGM) model. The goal of the presented work was to assess the capability of the anomaly detection system to resist this type of attack. The perturbation is chosen through differentiating the cost function. The experiment was evaluated using NLS KDD dataset. The outcomes indicated that the accuracy lowers from 0.99 to 0.53, implying that it is feasible to construct adversarial cases that will lead to a total misunderstanding of a prospective network attack. It is vital to build defensive procedures that can limit the quantity of incorrectly classified malware cases to make the IDS more resilient. In [46], a GAN-based algorithm was proposed to create AEs to train an efficient "NN" classifier. It has been shown that even if a classifier is built and trained with AEs that are generated from the original network dataset, the generated AEs can still affect the model and successfully destroy it. There are two steps that have been taken in this experiment. The first step is to create a strong NN classifier using GANs, in which the classifier was trained with AEs to improve its effectiveness. The second step is to leverage the FGSM to undertake AEs on an NN classifier in order to fool the model by making small changes in the data input samples. However, it is possible to develop defense systems against these types of network intrusion attacks by selecting important features and recalculating the attack success rate to mitigate the attack on the NN classifier using only important features. Another white–box attack [47] was developed for ML-based NIDS that uses a Multilayer Perceptron (MLP) network that was trained on two datasets to conduct binary classification and demonstrate the model's effectiveness. A model evasion attack versus MLP network using the AML approach in a white-box method, defined as the Jacobian-based Saliency Map Attack (JSMA), was performed. The researchers showed that the success of this attack and validation of its effectiveness in terms of accuracy were decreased by 22.52% and 29.87% for CICIDS and TRAbID, respectively. Moreover, the article also discussed factors that can be leveraged to prevent these types of attacks by limiting the size of the deep neural network, as a defensive system strategy for reducing the network's security vulnerabilities to adversarial sample production. However, since this approach assumes that the opponent has sufficient knowledge of the parameters needed for the model to implement an attack (white-box setting), which is not close to the real context, the experience should be used to assess the attack's effectiveness on other attack types (e.g., black-box setting) to indicate that evading network defensive systems and sustainable mitigation strategies is possible without very much activity. The researchers in [48] investigated the effectiveness of the FGSM and JSMA attack approaches in generating powerful targeted antagonistic instances that circumvent ML model. The AEs are created in a grey-box scenario with an MLP replacement classifier. Then, the resilience of certain NIDS classifiers is tested against the produced AEs, e.g., "SVM" with a linear kernel, "DT" based on the" CART" algorithm, "RF", and the most voting approach. The experimental findings reveal that all the classifiers are impacted more by using the JSMA method. The accuracy drops to 45%. The linear SVM is also reduced by 27% in accuracy, while RF is shown to provide the greatest withstand, where the accuracy was reduced by 18% and both the F1-score and AUC were reduced by 6%. The researchers have found that FGSM is not a strong technique to evade ML-based NIDS since the flow features were modified by 100%, while JSMA is a better realistic assault since only 6% changed in all features. The study's key flaws are that it assumes knowledge of the classifier's features, and therefore attack strategy produces the feature and vectors, rather than the AEs themselves.

Huang, Lee, et al. proposed a type of AEs that invested in the vulnerability of the DL classifiers in order to explore the influence of these attacks in the SDN environment [49]. The researchers collected the Packet-In messages and STATS reports to generate the SDN-based NIDS dataset. The suggested port scan attack detecting system is significantly dependent on the frequency of the delivered Packet-In messages and predefined features for the DL. Three typical DL models, "MPL", "CNN" and "LSTM", blended with 4 different adversarial testings have been examined and analyzed. Generally, the JSMA assault causes the greatest dramatic reduction in the capability of the used classifiers, ranging between 42% and 14%. Although the FGSM outcome didn't show its influence, it caused a major decrease in the LSTM accuracy by about 50%. The JSMA-RE attack reduces the accuracy of the MLP model by 35% without impacting the accuracy of the CNN or LSTM systems. In [50], a framework, named Evaluated Network Intrusion Detection System (ENIDS) to investigate the resilience of NIDS based on DL, was introduced. The proposed approach used four target models, DNN, SVN, RF, and LR over the benchmark dataset NSL-KDD. Then, four advanced attack methods, e.g., PGD attack, SPSA attack, L-BFGS attack, and Momentum Iterative FGSM, have been leveraged to produce adversarial samples. The results of this extensive experiment showed that the DNN model has the least ROC score of 0.37 and the MI-FGSM assault has the highest success rate. Debicha et al. evaluated the impact of AEs on DL-based NIDS [51]. The effectiveness of adversarial training as a defense mechanism is tested against several attacks, FGSM, PGD, and BIM. The FGSM reduced the prediction performance from 99.61% to 14.13% when the

NSL-KDD dataset was used, whereas PGD and BIM reduced it moreover to 8.85% on the same dataset. After applying the adversarial training, the model's resilience was improved against AEs, and it did so at the expense of accuracy on "non-attacked" data. The work in [52] focused on investigating the effectiveness of several evasion attacks, e.g., ZOO, PGD, and DeepFool, and made a comparison between the resilient ML and DL algorithms, e.g., C4.5, KNN, ANN, CNN, as well as RNN, in classifying "encrypted traffic". The effectiveness of these classifiers was evaluated in oppositional and adversarial attack situations. These classifiers are evaluated on two datasets for network traffic, SCX VPNN on VPN, and NIMS. The authors realized that without an adversarial environment, on average DL algorithms performed better than ML algorithms in terms of classification. Whereas, in the state of AEs, the resilience of ML and DL would depend on the kind of implemented attack. An analysis that focuses on evaluating the fragility of NIDSs based on ML techniques against AEs is presented [53]. The study involved an RF classifier that used network flows to classify between "botnet" and "benign" samples. The analysis was based on the CTU dataset, a public and tagged dataset encompassing both legal and botnet traffic gathered in a realistic and big organization. The duration of the flows, the overall number of sent packets, and the number of outgoing (Src) or incoming (Dst) bytes are all changed. The findings of this experiment showed that RF classifiers are vulnerable to adversarial instances and even though when extremely minor perturbations (combinations of 1 second, 1 byte, 1 packet) were considered, the detection rate dropped by higher than 20% and up to 50% for utilizing only some of the feature sets. The flaw in this strategy is that if an attacker uses more than four features, he or she would require a huge number of perturbations, which will affect not only the botnet but also the logic of the botnet. In [54], the study is concentrated on adversarial attacks that aim to impact the detection and prediction capabilities of ML models. The authors considered serious types of poisoning and evasion attacks targeting security solutions devoted to malware, NIDS, and spam and explored the possible damage that the adversary can cause to a cyber-detector. Then, the authors presented some existing original defensive techniques to mitigate such kinds of attacks. The experiment was evaluated utilizing CTU-13 dataset. The recall scores for MLP ranged from 93% to 97% under typical conditions. Where the KNN performs the worst, while RF performs the best. After the adversarial retraining enhances the performance of the classifiers, resulting in recalls of 0.49 for KNN and 60% for RF. Meanwhile, the recall measure for MLP and RF increases to 76% and 89%, accordingly due to utilizing feature removal, which removes disturbed features prior to actual training. The study proposed a network traffic attack that used Mutual Information (MI) to generate adversarial perturbations. The approach trains a replacement model to mimic the target system, enabling effective black-box attacks without needing internal knowledge of the original model. [55]. The method was tested against "SVM" and "DNN" classifiers using the UNB-CIC Tor dataset. According to the experimental data, DNN and SVM prediction accuracies decrease from 96.3% to 2% and 96.4% to 63.95%, respectively.

An improved boundary-based method that created adversarial DoS samples that could go around ANN-based NIDS was presented by Peng et al. [56], where they examined the characteristics of DoS attacks. By changing both continuous and discrete aspects of DoS settings, the proposed technique maximized Mahalanobis distance. It used optimization techniques and query outputs to operate in a black-box environment while accounting for DoS flow characteristics. The KDDCup99 and CICIDS2017 datasets were used to evaluate the performance of the model against the ANN model. The experimental results indicated that the proposal could reduce the prediction from 90% to 49% and generate adversarial DoS cases using only a few queries.

The work in [57] has tended to poison the NIDS system by incorporating algorithms as a new technique to gradually add antagonist samples. In this research, the Edge Pattern Detection (EPD) algorithm was proposed to create a newer poisoning technique that attacks multiple ML algorithms by generating AEs that were close to the discriminant boundary that was identified by the classifiers but classified as benign ones. Also, the authors solved the disadvantage of limited AEs obtained by BPD by introducing a Batch-EPD Boundary Pattern (BEBP) detection algorithm. Then, a moderate poisoning strategy, known as chronic poisoning attack, was presented to alter the proportion of training data each time of learning models. As a result of gradually adding antagonistic samples, the efficiency of NIDSs in detecting suspicious samples decreases significantly after many iterations of poisoning. The proposed poisoning technique on synthetic and original datasets outperformed the prior works on NIDSs. In [58], The study introduced to judge whether adversarial threats could deceive models in an SDN system. The authors built an anomaly-based NIDS called 'Neptune,' which used traffic flow features and multiple ML models. They also developed an adversarial testing tool, 'Hydra,' to see how evasion attacks affect Neptune in terms of reducing its ability to detect malicious traffic. The results showed that perturbing only a few input features significantly decreased the accuracy of a Neptune SYN flood DDoS attack across a number of classifiers, e.g., KNN, SVM, LR, and RF, while KNN illustrated the strongest resilience to these adversarial perturbations. Other classifiers, SVM, LR, and RF, exhibited similar vulnerability, with comparable drops in detection performance.

An adversarial threat technique has been presented by Han et al., which automatically changing the given traffic in both black-box and grey-box attacks and maintaining the operation of traffic features [59]. With the use of multiple ML/DL models and none payload features, the presented attack could be used to evaluate the resilience of different NIDSs. This technique approximated these examples to the misclassification boundary using PSO, e.g., an optimization algorithm, after using GAN to create the AEs. The introduced technique produced evasion greater than 97% in the half of the cases, according to the reported results using the Kitsune and CICIDS2017 datasets. However, leveraging GANs

to create AEs was time consuming and computationally difficult. As a result, this approach is not effective or feasible for real-time attacks in the application of IoT networks.

The limitations of current evaluation techniques that evaluated the Adversarial Training (AdvTrain) defensive technique leveraging gradient based adversarial attacks have been highlighted by Fan et al. [60, 61]. The authors proposed a new adversarial attack technique, named nongradient attack (NGA), and presented a new evaluation standard called composite criterion (CC), which took into account both attack success rate and accuracy. In order to provide adversarial examples outside the decision boundary, the NGA technique employed a search strategy. These samples retained their misclassification features while being repeatedly modified toward the given original data points. To systematically evaluate the effectiveness of the AdvTrain technique, the researchers conducted the experiments using two widely employed datasets, CIFAR-100 and CIFAR-10 [62]. The primary technique employed in this assessment was to compare AdvTrain's efficacy against four classifiers— C&W, PGD, BIM, and FGSM. The work came to the conclusion that NIDSes based on DNN of IoT traffic could not be robust enough. The dependability of DNN-based NIDSes should be evaluated more precisely in both AdvTrain and normal defensive techniques by utilizing NGA and CC. Then, the authors acknowledged the convergence speed limitation of the proposed NGA approach at the conclusion of the study and committed to improving it in the future work. Note that detection approaches against sophisticated poisoned data, typically created employing big data with closely related features, are needed, particularly for sensitive data relevant to the healthcare system [63].

The main focus of the research in [64] was to exanimate DNN-based NIDS against sophisticated evasion attacks, such as using the Jacobian Saliency Map Attack [65], Projected Gradient Descent [66], Carlini & Wagner, and the Fast Gradient Sign Method [67]. By introducing adversarial disturbed instances into the system, the primary goal of this work was to modify DL-based NIDS in order to incorrectly categorize network malicious as benign network traffic. Consequently, the NIDS's performance has been reduced in terms of AUC, F-score, accuracy, precision, and recall. In the majority of the most recent adversarial attacks, the C&W attack was shown to be the most potent among others. The reduced classification report and confusion matrix that the NIDS produced were quite comparable to those of other strong attack algorithms, such as FGSM, JSMA, and PGD. Under the C&W attack with the usage of the CICIDS2017 dataset, the AUC score is 63.41%, which is higher than that of FGSM (59.23%), PGD (58.48%), and JSMA (68.04%).

Table 1 provides a summary of existing studies on both poisoning and evasion attacks against ML-based NIDS.

According to the most given current advanced studies, the GAN attack has been considered the most powerful and dangerous type of adversarial attack against ML—based NIDS. These attacks can trick the system into misclassifying threats by creating realistic but fake data. However, a recent study published in December 2024 [44] provided new experimental results. It showed that DeepFool and KDE-based attacks are actually more effective than GAN attacks in bypassing ML-based NIDS. This means that DeepFool and KDE attacks can fool the system more easily or more often than GAN attacks, making them a bigger threat than previously thought.

**Table 1 Summarized studies on adversarial attacks (poisoning and evasion)**

| | Category | Year | Ref. | Dataset | Models | Attack Type |
|---|---|---|---|---|---|---|
| Attacks | Generations of Adversarial Attacks | 2018 | [24] | NSL-KDD | IDSGAN+ KNN, SVM, DT | Evasion |
| | | 2020 | [25] | CICIDS2017 | GAN, RF, LR, NB, DT | Evasion, Poisoning |
| | | 2021 | [26] | NSL-KDD, UNSW-NB15, and CICIDS2017 | CNN+LSTM , LR, RF, K_NN, DT, ADABoost. | Evasion |
| | | 2019 | [27] | KDD99 | DNN, KNN, LR, SVM, NB, RF, DT, GB | Evasion |
| | | 2020 | [28] | CICIDS2017 | GAN, Gradient Boosted DT | Evasion |
| | | 2021 | [29] | NSL_KDD | FGSM, C&W, PGDSVM, DT, RF, NB, and DNN | Evasion |
| | | 2018 | [30] | NSL-KDD | ZOO, WGAN, C&W, DNN | Evasion |
| | | 2020 | [34] | KDD-Cup 1999 and CIC-IDS 2017, MNIST and CIFAR-10 | MASKE-GAN AND 4 CLASSIFIERS LR, XGBOOST, DT, MLP | Evasion |
| | | 2021 | [36] | UNSW-SOSR2019 | GNN | Poisoning |
| | | 2023 | [42] | NSL-KDD, CICIDS2017 | KNN, SVM, DNN | Evasion |
| | | 2024 | [43] | CSE, CICIDS2017 | DL | Poisoning |
| | tack Generation ba | 2018 | [45] | NSL-KDD | FSGM, ANN | Evasion |
| | | 2020 | [46] | IEEE big data 2019 | GANS AND FGSM | Evasion |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 2017 | [48] | NSL-KDD | Majority voting ensemble, DT, MLP, RF, SVM, JSMA. | Poisoning |
| | | 2018 | [49] | Theirs | MLP, JSMA-RE, CNN, LSTM, FGSM, and JSMA. | Evasion |
| | | 2019 | [50] | NSL-KDD | RF, SVM, PGD, SPSA, DNN, LR, L-BFGS, and MIFGSM. | Evasion |
| | | | | | | |
| | | 2021 | [51] | NSL-KDD | BIM, FGSM, PGD , DNN | Evasion |
| | | 2021 | [52] | SCX VPNNonVPN, NIMS | Deepfool, PGD, Zoo, CNN, DNN, KNN, RNN, C4.5 | Evasion |
| | | 2020 | [47] | CICIDS and TRAbID | MLP Classifier and JSMA for generated AEs | Evasion |
| | | 2018 | [53] | CUT | RF | Evasion |
| | | 2019 | [54] | CTU-13 | RF, MLP, and K-NN | Evasion, Poisoning |
| | | 2019 | [56] | KDDCup99 and CICIDS2017 | DNN | Evasion, Poisoning |
| | | 2019 | [55] | UNB-CIC | SVM, DNN | Evasion |
| | | 2018 | [57] | KDDCcup99, NSL-KDD, KYOTO 2006 | EPD, BEPD, LR, SVM, NB | Poisoning |
| | | 2019 | [62] | CICIDS2017, ARPA SYN flood set | LR, SVM, KNN, RF | Evasion |
| | | 2020 | [59] | Kitsune, CICIDS2017 | Kitsune, IF, LR, SVM, DT, MLP | Evasion |
| | | 2022 | [62] | CIFAR-10, CIFAR-100 | DNN | Poisoning |
| | | 2024 | [66] | CICIDS2017 | DL | Evasion |
| | | 2022 | [3] | CSE.CICIDS2018 | RF | Evasion |
| | | 2024 | [44] | CSE.CICIDS2018, ADFA-LDs, CSE.CICIDS2019 | DNN | Evasion |

## 3.2 DEFENSE ML-BASED NIDS MODELS AGAINST ADVERSARIAL ATTACKS

In this part, we will summarize the studies that provided numerous strategies and explore what defenders may do to improve the security of ML-based NIDSs versus adversarial assaults in both forms (poisoning and evasion). Defenses, such as data sanitization and robust methods mitigate label-flipping attacks, but they are vulnerable against feature-space perturbations. Hybrid approaches combining anomaly detection and adversarial training show promising solutions, but unfortunately, they require further validation. In this subsection, we will discuss each one in detail.

### 3.2.1 DEFENSE AGAINST ADVERSARIAL PERTURBATIONS

Many studies were applied on AML to increase the resilience of the ML models against different attacks. The approach belonging to [68] introduced an approach known as Adversarial Learned Anomaly Detection (ALAD), based on bi-directional GANs, that adversarially knew the features' distribution in order to expose anomalies. The "ALAD" method then used recreation errors depending on these adversarially features to specify if a data sample is malicious. ALAD relies on the latest advances to ensure latent space, data space, and cycle consistencies while stabilizing the GAN during the training phase, and the results showed that the proposed technique further helped to largely detect anomalies. The performance was evaluated using the Arrhythmia tabular, KDD99, CIFAR-10, and SVHN datasets. These improvements boosted defensive technique capability highly, compared to previous works, proving their value.

A bidirectional-GAN (BiGAN) was used to develop a protection technique for NIDS versus AEs, which was composed of a (G), (D), as well as Encoder (E), named ASD [68]. The E mapped input data samples to the latent space during the training phase. Through training, the (G) learned the normal example distribution of data, and the (E) determined the incoming sample's potential form, which was then used by the (G) to generate the de-noised recreated sample, while the (D) was used to distinguish real input samples from the fake samples, generated by (G). The AEs were detected by the proposed ASD, and it relied on the discriminator's capability to determine if the entry was a normal sample. This method estimates the sample's reconstructing error and discriminator matching error after the

training. The malicious data was then deleted, leaving just the normal data to be fed into the classifier. For a full experimental assessment, the research adopted the NSL-KDD dataset. The actual test dataset with the created AEs, e.g., FGSM, PGD, and MI-FGSM, are integrated to a new dataset for further evaluation to examine the capability of defensive mechanism ASD to identify AEs. In the adversarial setting, the efficiency of DNN-based NIDS with ASD significantly improved. The FGSM enhanced by 11.85% and The PGD by 26.46%. However, the effect of ASD showed a small enhancement in detecting MI-FGSM adversarial examples, indicating that more work is needed to prevent this attack.

A good strategy was implemented through generated examples to reinforce minority class (specific amount of sample instances) and address the issue of class imbalance [69]. The introduced method relied on the "Divide-Augment-Combine" DAC" method, in which the examples were first divided by k-means (divide), then next data was supplemented on a group basis utilizing GAN (augment), and lastly, extended examples were inserted into a traditional classifier to form a learning model. Two public datasets UNSW-NB15 and IDS-2017 were used, and it has been pointed out that the suggested approach improves the performance of spotting abnormalities in the network by 9.6% for UNSW-NB15 and 21.5% for IDS-2017.

Moreover, for increasing detection effectiveness in anomaly-based NIDS, a new hybrid GAN-based oversampling technique has been proposed, including three primary phases: feature extraction using Information Gain and PCA, data clustering using DBSCAN, and data generation using WGAN-DIV [70]. Three HTTP datasets are leveraged for performance assessment: NSL-KDD-HTTP, UNSW-NB15-HTTP, and Kyoto2006-Plus-HTTP. The proposed technique utilized several ML algorithms, such as SVM, RF, XGBOOST, LR, KNN, and DT, with SMOTE a traditional oversampling for comparisons. The XGBoost classifier got the best F1 score in all given datasets compared with the five NIDS classifiers. Also, when compared to the SMOTE approach, the introduced model had equivalent or even superior performance. However, experimentation required more complete evaluation, such as data distribution consistency and diversity, as well as an increase in the existing model's stability.

In the domain of cyber security, by using the neuron activations at test time, the work in [71] identified the adversarial assaults using four recognized evasion attack techniques: Fast Gradient Sign, Basic Iterative Method, Carlini and Wagner attack, and Projected Gradient Descent. The researchers gathered the test time of an “ANN” model that trained on a subset of the CICIDS2017 dataset, as well as the neural activations of AEs. These activations have been employed in order to train and evaluate five different ML models to expose adversarial samples, attaining a recall of 0.99 with two of them, “RF” and “KNN” classifiers in adversarial attacks. The findings pointed out that the prospect of developing an adversarial attack detector does not impair the protected model's classification results, paving the way for more research into network defense as well NIDS as depending on ML techniques. The resistance of DL-based NIDS to AEs was examined by AbouKhamis et al. [72]. A DNN model was trained on AE using the UNSW-NB15 dataset with the min-max technique. To create adversarial samples that can maximize the loss, the max technique was applied. However, in order to reduce the loss of adversarial samples during the training, the min method was used as a defensive technique. Bit Coordinate Ascent (BCAS), Multi-Step Bit Gradient Ascent (BGAS), Randomized Rounding Approach (rFGSMS), and Deterministic Approach (dFGSMS) were employed to create the AEs. To increase the model's resilience to AEs, a model was trained using both benign instances and the AEs that each technique created. Furthermore, a clean dataset was employed to train a natural model. Four sets of AEs for each approach were used to assault the five constructed models during the testing process. Out of all the adversarial attack methods, the model that was trained using AEs produced by dFGSMS had the lowest overall evasion rates. In the meantime, the (BGAS) performed better than any attack technique in all five constructed models. PCA was used to eliminate invaluable features in the dataset during the second experiment setting.

Research on dimensionality minimization in DL was concluded to improve the robustness of DL-based NIDS towards evasion attempts. Another defense measure, the min-max formulation presented by Abou Khamis and Matrawy, was augmented with tailored inputs through model training [73]. Five white-box assaults, FGSM, CW, BIM, PGD, and DeepFool, were employed to create strong AEs. The usefulness of the min-max concept was tested using “ANN”, “CNN”, and “RNN” classifiers on UNSW-NB15 and NSD-KDD datasets. The experiments showed that the accuracy percentages of the models were improved significantly. The transferability of black box AEs across different NIDSs based on ML, employed distinct ML approaches in black-box cases, has been investigated by Debicha12 et al. [74]. To avoid a DNN model, adversarial samples were created using FGSM and PGD. The adversarial samples have been investigated and applied to well-known ML algorithms, such as SVM, DT, LR, RF, and LDA. Because of their composition of differentiable parts, the outcomes indicated that the DNN has been shown to be the most degraded one in accuracy compared to the other classifiers.

To mitigate the impact of adversarial perturbations, a more suitable solution of an innovative method that uses the defensive distillation methodology was proposed [75]. This methodology showed a better performance in cybersecurity detection missions, considering the random forest algorithm. The uniqueness of the proposal was based on two phases: the production of probability labels from the hard target class, and the supervised model trained with the obtained probability labels to conduct cyber detection. CTU-13, a public collection of diverse datasets, was used for the experimental tests. An exhaustive campaign of experiments conducted clearly illustrated that the introduced approach was better than the previous works in two ways: it improved the recognition rate by up to 25% in instances with

adversarial manipulated input data, and it achieved equivalent or superior accuracy in scenarios without adversarial attacks. A DDoS self-defense technique that was resilient to adversarial attacks was presented by Benzaid et al. [76]. In order to mitigate and detect DDoS attacks in the application-layer, the DL and SDN have been employed. In terms of server response time and system load, the proposed technique has been shown to perform very well. The CICIDS2017 dataset and DDoS traffic traces were utilized to train the proposal, which was constructed using MLP. As a defensive approach, adversarial training was used, in which the model was trained using AEs that are produced by the FGSM technique based attack. Similarly, researchers [77] utilized GAN to expose DDoS in SDN systems. This way was very effective since GAN can use adversarial training and naturally generate strong adversarial traffic. The experiments have been tested on actual SDN traffic, and the outcomes showed that GAN outperformed LSTM, MLP, and CNN classifiers in terms of detection rate, where the CICDDoS 2019 dataset was used for testing purposes.

In 2022, Raghuvanshi et al., transformed all of the NSL-KDD dataset's symbols into numerical expressions in order to obtain the entire data in detail. They then used Principal Component Analysis (PCA) to extract valuable features from the obtained dataset, where LR, SVM, and RF were employed to classify the same dataset. According to the constructed models' performance outcomes, the RF, LR, and SVM provided 85%, 78%, and 98% accuracies, respectively, indicating good performance [78]. To further elevate the impact of the NIDS, Faker and Dog dou integrated DL model with large data. They conducted an analysis utilizing three models to classify the network traffic data: The Deep Feed Forward Neural Network (DFFNN) classifier, Random Forest, and Gradient Boosting Tree. The main outcome of this proposal compared to [79], the DNN classifier was shown to provide accuracy of 97.01% in multiclass and 99.16% in binary classes with better precision. An AI-powered NIDS solution intended for the IoT environment was introduced via Siganos et al. in 2023. The model's SHapley Additive exPlanations technique with classifiers derived from DL and algorithms was used to build and explain the features. Transparency was restored to the NIDS's black-box operating style. The evaluation of this NIDS revealed its proficiency in explainable AI, which is needed for sophisticated AI systems, in addition to the performance improvement. An experimental evaluation of the proposed framework was achieved utilizing two balanced datasets, including IEC 60870-5-104. AdaBoost, FR, DT, RBF, XGBoost, LR, Naive Bayes, SVM, Linear SVM, Quadratic Discriminant Analysis, and DNN were the ML methods used in this work, and the results showed that an F1-score of 66% for RF was the most accurate among all others [80].

### 3.2.2 DEFENDED STRATEGIES LEVERAGING MISCELLANEOUS TECHNIQUES

Moving to other strategies based on feature space, a detection technique to evaluate the system against the FGSM attack method was presented [81]. This technique targets to achieve a balance between reducing the possible assaults on the classification and feature space outcomes. The work has been assessed using the CICIDS2017 dataset. The researchers investigated how vulnerable data features were to modification attacks. Then, defense solutions for algorithmically generated AEs were produced. Also, Rfe feature reduction method was utilized to remove those weak features during the FGSM assault, while the rest feature enhanced classifier resiliency. Given a few features, the proposal provided less improvement on the accuracy; however, when all features were applied, the accuracy was greatly enhanced, except accuracy under assault seldom reached only 60%. The experimental results ensured that feature selection could be leveraged to increase model accuracy under the FGSM attack.

A lot of research has invested in emerging methods and techniques to refine the development of a modeling performance for ML, including ensemble approaches, which depended on integrating several models called ensemble members and trained on the same training data in order to provide a single improved predictive model and hybrid techniques that powering up the system models. Other techniques used feature selection methods, such as PCA, which chose a subset of the variables that preserves the largest amount of data available from the total given dataset, where a genetic algorithm (GA) was also used to determine the best group during the evaluation. For selection features, the first step was to generate a population on a subset of potential evidence, and then this subset was evaluated using the predictive model of the desired target. Given these considerations, the authors in [82] proposed a hybrid technique that combined two ML techniques, RF with Classification and Regression Trees (CART) to identify many probable threats, where effective feature selection and classification were conducted. By using relevance score, the RF technique was utilized to minimize the “42” features in the UNSW-NB15 dataset to only ”11” most essential features; however, it has been shown that the maximum accuracy was achieved when all 42 features were incorporated. As a result, there is a trade-off involving accuracy and complexity. Although further investigation has been involved to determine more valuable features to be “13”, the model still had a decent accuracy of 87.74% when trained with the top 13 features, but the training time was less than when utilizing the whole original collection of features. When compared to current algorithms, the findings revealed that the introduced technique performed better. Note that, the Principal Component Analysis (PCA) can be used to increase the system's overall performance by converting the closely “correlated” features of the dataset into a collection of “uncorrelated” features, then employing the new additional features to the dataset. The features reduction method, PCA, applied on the KDDCup99 dataset, has been utilized in order to detect anomalies, and the experiment was carried out via computer networks leveraging various methods” LR”,” NN”, and” DT” [83]. An ensemble learning was also used to further improve the detection rate. Afterward, the findings of the classifiers have been mixed, through gathering the most sensitive data from all models throughout the training as well

as testing phases. Then, depending on individual classifier results, a weighted majority voting mechanism was employed to find if the instance was malicious. The accuracy in the testing and training was 91.66% and 92.08% for both NN and DT, 96.13% and 96.66% for LR, and 89.83% and 90.67% for NN. In [84], the same dataset and PCA method in previous research were leveraged with the random forest (RF) as a classifier to develop an efficient NIDS. The PCA aided in the organization of the dataset (KDDCup99) by lowering its dimensionality, while the RF aided in classification. The collected results showed that the suggested strategy outperformed existing approaches, such as "SVM"," NB", and" DT", in regarding to accuracy. The performance time "min" was 3.24 minutes, the accuracy rate was 96.78%, and the error rate was 21%, according to the findings obtained by the presented approach. The work in [85] incorporated feature engineering, Chi-square statistical approach (ChiX2) addition to PCA, to identify the best collection of features with the highest accuracy in order to reduce the complexity and the training time of the ML models. The usefulness of ML approaches, such as "SVM" and "ANN" for detecting intrusions in the cloud environment, was investigated, and the models were trained and tested using the UNSW-NB-15 dataset. The findings showed that ANN performed marginally better compared to "SVM". The best SVM model had an accuracy of 68%, while the best ANN model had an accuracy of 72%. Furthermore, for about the same feature category, "SVM" had a precision of 46%, while "ANN" had a precision of 78%. Therefore, "ANN" showed to provide a better probability of detecting aberrant traffic. Likewise, in ANN, the Connection Features model surpassed the SVM model by approximately 20%F when combining ANN with ChiX2 versus SVM with PCA.

Some studies tended to use Genetic Algorithm (GA) as an optimization method for feature selection purposes, aiming to enhance the results and increase attack detection. In [86], ensemble learning methodologies (Boosting and bagging approaches), such as Distributed Random Forest "DRF", Gradient Boosting Machine "GBM", XGBoost as well as DNN, were presented. The genetic algorithm was used for feature selection, and the popular NSL-KDD dataset was employed for evaluation purposes. The findings indicated that the proposal outperformed various traditional ML models, and the "DNN" model exceeds earlier results after employing a genetic process to choose features; however, the work has several limits. Only a few bits of the dataset with 10 iterations were used in the operation of GA. As a result, if the entire dataset is used with additional iterations, the results could be enhanced. In [87], Network Anomaly Detection System (NADS) was proposed, and different datasets and classifiers were used. The proposal was intended to detect activities that indicated a network traffic assault. The KDD99 dataset was used to test the model, and the feature selection method was achieved using GA. The GA is often used to produce multiple individuals to discover which features of the individual provided a better outcome when it came to understanding the network traffic behavioral pattern. By using the specified features, an ideal feature subset was attempted to be acquired in anomaly determination. Weka classifiers were used during the training and testing of the datasets to determine the optimum fitness value and individual features that were more successful at detecting anomalies that occur in real time. The result showed that J48 classification algorithm had the highest accuracy of 91.1% and picked the fewest number of features of 22. Likewise, when using the training dataset to train the best models, the NSGA-II method with the J48 classifier further refined the accuracy to 97.03% and picked the same number of features (22).

In [88], a NIDS has been used with a small feature number to reveal malicious. The system extracted features via GA idea to increase utilized resources and minimize the computational time complexity. GA is used to remove the Irrelevant data from the dataset, the results were improved when a specific number of features was utilized. Given the results of the mean values to all datasets, the TPR was also improved once the feature ranking technique had been done. Spearman's rank correlation coefficient was leveraged to increase the effectiveness of ML-based NIDSs [89]. An ensemble learning strategy that incorporates the benefits of each individual detection algorithm, as well as the feature selection process, was adopted. Seven individual classifiers were compared to find the most acceptable fundamental classifiers for ensemble learning. The ensemble models "LR", "DT", and "GB", and the CSE-CIC-IDS2018 dataset were incorporated for evaluation purposes. The use of Spearman's rank correlation coefficient aided the choice of 23 out of the dataset's 80 original characteristics. The proposed model's accuracy was 98.8% with a reasonably short detection time that decreased from 34 minutes and two seconds to 10 minutes and 54 seconds, respectively. Compared to NN techniques, such as "DNN", "RNN", and" ANN", the suggested model performed better. In [90], different NIDS datasets to assess the model's performance with the removal of strongly associated features were presented as a defensive strategy. The proposal was used for anticipating well-known and zero-day communication network attacks in near-real-time, called ANIDINR (anomaly-based NIDS), and to reduce the proportion of undiscovered attacks, false alarm rate, and total time. Many enhancements were made to boost the model's performance by cleaning the datasets further and lowering the complexity of the models. The correlation analysis enabled the removal of strongly associated features from datasets, while the RF impact analysis revealed how the Mean Squared Error (MSE) rose when features were arbitrarily modified. This research was evaluated using KDD and NSL-KDD datasets as well as five models. Finally, the cleaned datasets were used to train several ML models (SVM, RF, and XGBoost), and 2 Deep Learning models (Neuralnet and Keras). The findings showed that the XGBoost model, which was trained on the KDD, achieved the highest results in terms of accuracy.

In [91], Bhosale et al. proposed filter-based hybrid feature selection algorithm (HFSA). Most relevant features were kept and utilized to create classifiers for corresponding classes. Along with the HFSA method, the authors first offered a methodology for identifying cyber-attack brute force by changing the HFSA and classification algorithms.

The NB classifier was used to classify the data. When compared to other approaches, the performance of the suggested method demonstrated its efficiency. A hybrid FIDS, combined with the "J48 DT" and "SVM" of ML techniques, was designed as a defense technique [92]. Particle Swarm Optimization (PSO) was utilized to choose significant features from KDD CUP 1999 dataset for training and testing, and the dataset was partitioned into three different rates: 60:40, 70:30, and 80:20 percentages. The experiment's result revealed 99.1% accuracy, 99.6% detection rate, and 10% FAR for 60:40 datasets, while 99.2% accuracy, 99.6% detection rate, and 90% FAR for 70:30 datasets, and 99.1%, 99.6%, and 90% FAR for 80:20 datasets, respectively. A new feature selection termed Mutation Cuckoo Fuzzy (MCF) was used to pick the optimal feature subsets and MVO-ANN for classification purposes in an anomaly-based detection system [93]. To identify the ideal weights and biases of ANN-MLPs, the training technique took into account the MVO's capabilities in terms of high exploration and exploitation. This approach was used to address the problem of NIDS using NSL-KDD dataset. The suggested technique chose 22 features out of 41 as the most critical features that developed the performance of an oddity NIDS significantly. The accuracy, detection rate, and false alarm rate are assessed in two experiments: using two feature selection methods, MCF and MVO-ANN, and without feature selection, MVO-ANN. The comparison scenario showed the usefulness of feature selection and its influence on effectiveness. The findings showed that the suggested technique outperformed MVO-ANN with a precision of 98.16% and produced better and more stable outcomes in terms of the given metrics and run time.

Detection model, researchers [94] used feature reduction, specifically RFR, to exclude the features that the adversarial evasion attempt was aiming for. An ensemble of many ML models, e.g., SVM, RF, DNN, and LR, which was robust to evasion assaults was found using several feature sets of smaller size. The KDDCup99, CICIDS, and DARPA datasets were utilized for the tests, and the evasion dataset was created using the Hydra tool. The experimental results showed that the ensemble model successfully identified a number of evasion attacks that are impossible to identify with a given single classifier that has all of the features. GNN architecture was presented by Yumlembam et al. [95] in order to train and strengthen an Android malware detection system. The proposal showed how well GNN worked to create graph embeddings utilizing the centralized properties of an API graph in conjunction with "Permission" and "Intent" to enhance malware classification. This method has been called VGAE-MalGAN, and it can be used to dynamically create an adversarial Android malware API graph and efficiently add nodes and edges to an existing API graph. These have been used to deceive the GNN-based malware classifier that was previously trained using GNN. The malware API graph's original semantics will still be kept. Note that both of a generator and a replacement detector have been used to build the VGAE-MalGAN. A GraphSAGE model was employed to serve as the replacement detector, while the Generator was the modified variation graph. AutoEncoder showed how retraining the model can strengthen it against VGAE-MalGAN attacks and obtained excellent attack detection rate, where the Drebin and CMaldroid datasets were used. In this study, two-phase defensive technique against Carlini & Wagner (C&W), the most potent optimization-based adversarial attack, was used. Training and testing are the two defensive steps. The modified adversarial training with Gaussian Data Augmentation (GDA) was used during the training phase and the resulting adversarial method was subjected to the Feature Squeezing (FS) approach during the testing step before being sent to the robust NIDS model for final classification. The employed two-phase defensive technique was successfully evaluated using the most recent dataset, CIC-DDoS-2019 [96]. A DeepFool based defensive approach was proposed to mitigate the impact of the incorporated attacks: DeepFool, KDE, ZOO, and GAN models, using CICIDS2018, CICIDS2019, and ADFA-LD datasets. The experimental findings showed that the proposal can perform better than other models in terms of achieving higher detection rate [44]. A summarization of existing defensive ML-based NIDS models against poisoning and evasion attacks is given in Table 2.

The given state-of-the-art studies have shown that GAN models can help reduce many serious attacks in ML-based NIDS. However, a newer study experimentally found that DeepFool and KDE-based defensive techniques can operate even better than GANs if they are used properly. When these methods are applied, they can make the system stronger and more resistant to different types of serious attacks, and this in turn gives a better protection level [44].

**Table 2 Summarization of defensive approaches**

| | Category | Year | Ref. | Dataset | Models |
|---|---|---|---|---|---|
| | Defenses against Adversarial Perturbations | 2018 | [68] | Tabular datasets, KDD99 and Arrhythmia. Image datasets SVHN and CIFAR-10 | GAN |
| | | | | | |
| | | 2020 | [69] | UNSW-NB15 and IDS-2017 | GAN, K-Means |
| | | 2020 | [70] | NSL-KDD-HTTP, UNSW-NB15-HTTP, Kyoto2006-Plus-HTTP | IG, PCA, WGAN-DIV, SVM, RF, XGBOOST, LR, KNN, DT |
| | | 2020 | [71] | CICIDS2017 | ANN, RF, SVM, XGBOOST, KNN, FSGM, CW, PGD, BIM. |
| | | 2020 | [72] | UNSW-NB15 | BCAS, BGAS, dFGSMS, Rfgsms, DNN |

| | | | | | |
|---|---|---|---|---|---|
| | | 2020 | [73] | UNSW-NB15, NSL-KDD | CNN, Recurrent NN, Deepfool, FGSM, PGD, DNN, BIM, and C&W. |
| | | 2021 | [74] | NSL-KDD | RF, SVM, DNN, DT, LDA, LR, FGSM, and PGD. |
| | | 2020 | [75] | CTU-13 | RF |
| | | 2020 | [76] | CICIDS2017 | MLP |
| | | 2021 | [77] | CICIDS2019 | LSTM, MLP, GAN, CNN |
| | | 2022 | [78] | NSL-KDD | DNN, SVM |
| | | 2023 | [80] | CICIOT Dataset 2022, IEC60879-5-104 | DNN, DT, RF, LR, Adaboost |
| | **Defenses against Miscellaneous** | 2021 | [81] | CICIDS2017 | Systematic feature selection, FSGM, RFE |
| | | 2020 | [82] | UNSW-NB15 | Hybrid RF, CART |
| | | 2018 | [83] | KDD99 | Ensemble of LR, DT, NN |
| | | 2020 | [84] | KDD99 | PCA, RF |
| | | 2020 | [86] | NSL-KDD | DRF, GBM, DNN, XGBOOST |
| | | 2019 | [85] | UNSW-NB15 | SVM, ANN, PCA, ChiX2 |
| | | 2019 | [88] | NSL-KDD, AWID | AE-RL |
| | | 2019 | [87] | KDD Cup 99 | ANN, GA, DT |
| | | 2020 | [89] | CSECIC- IDS2018 | LR, DT, GBoost, Spearman's rank |
| | | 2018 | [92] | KDD Cup 99 | HFSA, NB |
| | | 2020 | [90] | NSL-KDD, KDD-Cup 1999 | SVM, RF, XGBOOST, Keras, Neuralent |
| | | 2020 | [93] | NSL-KDD | MCF, MVO-ANN |
| | | 2020 | [91] | KDD Cup 99 | J48 DT, SVM, PSO |
| | | 2021 | [94] | CICIDS2017, DARPA, KDDCup99 | SVM, LR, DNN, RF |
| | | 2022 | [95] | Drebin, CICMaldroid 2020 | GAN, GNN |
| | | 2024 | [96] | CICIDS2019 | Feature Squeezing |
| | | 2022 | [3] | CSE.CICIDS2018 | RF |
| | | 2024 | [66] | CSE.CICIDS2018, ADFA-LD, CICIDS2019 | DNN |

## 4. DISCUSSION

The main target and findings from this review paper reveal that although ML-based NIDSs provide remarkable potential in detecting malicious activities in computer networks, they are still susceptible and vulnerable to different adversarial techniques. This study classifies these attacks into three primary types, as follows:

•Evasion Attacks: These attacks occur at the inference stage, where attackers manipulate given input data to evade the detection system. Different techniques, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) have proven to be highly effective in deceiving ML models by introducing and incorporating minimal perturbations to given input data.

•Poisoning Attacks: Unlike evasion attacks, poisoning attacks target the training phase of the ML models. By inserting infected (malicious) data into the training dataset, the attackers can negatively impact on the learning process in order to render the model misclassify threats in real-world applications. This highly compromises the reliability of NIDS over time.

•GAN-Based Attacks: The attack of GANs is considered one of the most promising threats in creating adversarial examples that can successfully bypass NIDS. GANs usually create synthetic network traffic that is similar to the original ones while containing malicious data that are hard to be detected.

There are several techniques have been proposed as defensive approaches to mitigate these attacks, as follows:

•Adversarial Training: Training ML models with adversarial attacks can improve the robustness of the system; however, this method requires additional computational costs.

•Feature Selection Approaches: Methods, such as PCA and GA, have been shown to highly improve NIDS accuracy via selecting the most valuable and preferable features and eliminating redundant ones.

•Anomaly Detection via GANs: Some studies have proposed using GANs for anomaly detection purposes, where the discriminator in there can learn to differentiate between normal and malicious traffic.

The main limitations of current existing approaches can be summarized, as follows:

•Adaptive Attack Weaknesses: Defensive techniques often consider static adversaries, leaving the entire system exposed to evolving attack methods, e.g., Deepfool and GANs.

•Computational Overhead: Techniques, such as ensemble and adversarial training approaches, require significant resources, and this in turn limits the scalability.

•Dataset Bias: The high dependency on outdated datasets and benchmarks significantly reduces relevance to sophisticated threats. To address this, researchers should consider leveraging large-scale and up-to-date datasets in adversarial ML research since older datasets often fail to reflect current attack techniques in real-world applications.

•Poor Generalizability: Many ML models perform very well on specific datasets, e.g., NSL-KDD, but fail to adapt to dynamic real-world network environments.

Note that adversarial training strengthens systems by revealing them to well-known attack patterns during the training, including GAN-generated attacks and evasion techniques. Such approach can help models identify and block these threats. However, it depends heavily on historical data, and therefore it cannot defend against never-seen threats before, such as novel attack methods and zero-day exploits. Unfortunately, this leaves large gaps in system protection. Anomaly detection, on the other hand, monitors for unusual behavior, e.g., data patterns and unexpected network traffic, to flag possible new attacks. Although this makes the system adaptable to new threats, the alerts could lack clarity, including unusual activity detected without specifics. This forces security teams to spend additional time investigating ambiguous warnings. Such delays negatively impact the system's response and allow fast-moving attacks to bypass the detection. Table 3 shows a comparative analysis of different defense approaches, including trade-offs between adaptability, robustness, and required resources among different defensive approaches.

**Table 3 Summarization of comparative analysis on defensive approaches**

| Defenses | Attack Types Addressed | Computational Cost | Robustness | Main Limitations |
|---|---|---|---|---|
| Anomaly Detection | Novel/unknown attacks | Different from Low to High | Detects zero-day anomalies | High false positives reduce reliability |
| Feature Selection | Static poisoning attacks | Moderate | Reduces attack surface for basic poisoning | Vulnerable to adaptive poisoning strategies |
| Adversarial Training | Evasion, GAN-based attacks | High | Effective against known evasion patterns | High resource require; limited adaptability to novel evasion approaches |

Given the aforementioned techniques into consideration in order to provide a high level of resilience, no single defensive technique has been shown to offer full protection against all adversarial attacks. Therefore, a multi-layered security approach integrating multiple defensive approaches may help to offer a better level of protection. Also, the door is still open for the hyper-heuristic-based ML methods to show their ability to offer excellent secure systems against sophisticated attacks.

## 5. CONCLUSION

This review study highlights the increasingly growing in adversarial attacks against ML-based NIDS and the continuous arms race between attack defense techniques. The paper shows the real need for a more adaptive security system that is able to prevent or at least mitigate sophisticated threats. Even though adversarial training and feature selection approaches have shown to be one of the best promising techniques in protecting the system, they still remain expensive and cannot be generalized well in all different attack types. Future research works should mainly focus on hybrid defense approaches that combine multiple techniques in order to ensure a complete protection system against adversarial attacks. Moreover, real world datasets that can be collected from different large network servers should be utilized in order to validate the NIDS performance more effectively. Experts in cybersecurity and ML areas can collaborate to develop resilient and secure NIDS.

## FUNDING

None

## ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their efforts.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest

## REFERENCES


[1] Q. Alasad et al., "Resilient and secure hardware devices using ASL," ACM J. Emerg. Technol. Comput. Syst., vol. 2, no. 17, pp. 1–26, 2021.

[2] K. P. Immastephy and A. A., "A systematic review on network intrusion detection system based on machine learning and deep learning approach," in E3S Web Conf., EDP Sciences, 2024.

[3] S. Alahmed et al., "Mitigation of black-box attacks on intrusion detection systems-based ML," Computers, vol. 7, no. 11, p. 115, 2022.

[4] M. K. Jmila and H. A., "Adversarial machine learning for network intrusion detection," Computer Networks, vol. 214, p. 109073, 2022.

[5] R. Ferdiana, "A systematic literature review of intrusion detection system for network security: Research trends, datasets and methods," in Proc. 4th Int. Conf. Informatics Comput. Sci. (ICICoS), 2020.

[6] K. D. and M. A. He, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," IEEE Commun. Surv. Tutorials, vol. 25, no. 1, pp. 538–566, 2023.

[7] C. M. Alatwi and H. A., "Adversarial machine learning in network intrusion detection domain: A systematic review," arXiv:2112.03315, 2021.

[8] A. T. T. Saheed et al., "A novel hybrid ensemble learning for anomaly detection in industrial sensor networks and SCADA systems for smart city infrastructures," J. King Saud Univ. - Comput. Inf. Sci., vol. 35, no. 5, p. 101532, 2023.

[9] A. T. and Y. S. Abdulganiyu, "A systematic literature review for network intrusion detection system (IDS)," Int. J. Inf. Security, vol. 22, no. 5, pp. 1125–1162, 2023.

[10] Q. M. H. and S. A. Alasad, "Performance and complexity tradeoffs of feature selection on intrusion detection system-based neural network classification with high-dimensional dataset," in Int. Conf. Emerging Technol., 2022.

[11] O. Ibitoye et al., "The threat of adversarial attacks on machine learning in network security—a survey," arXiv:1911.02621, 2019.

[12] M. R. B. and M. R. McComb, "Machine learning in pharmacometrics: Opportunities and challenges," Br. J. Clin. Pharmacol., vol. 88, no. 4, pp. 1482–1499, 2022.

[13] T. J. G. and A. R. Jiang, "Supervised machine learning: A brief primer," Behavior Therapy, vol. 51, no. 5, pp. 675–687, 2020.

[14] A. R. F. and P. G. Károly, "Unsupervised clustering for deep learning: A tutorial survey," Acta Polytech. Hungarica, vol. 15, no. 8, pp. 29–53, 2018.

[15] A. G. P. and S. C. Shakya, "Reinforcement learning algorithms: A brief survey," Expert Syst. Appl., vol. 231, p. 120495, 2023.

[16] J. Liu et al., "Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems," IEEE Commun. Surv. Tutorials, vol. 24, no. 1, pp. 123–159, 2021.

[17] Y. Zhang and P. L., "Defending against whitebox adversarial attacks via randomized discretization," in Proc. 22nd Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, 2019.

[18] N. Carlini et al., "On evaluating adversarial robustness," arXiv:1902.06705, 2019.

[19] W. Hu and Y. T., "Black-box attacks against RNN based malware detection algorithms," in AAAI Workshops, 2018.

[20] A. Pattanaik et al., "Robust deep reinforcement learning with adversarial attacks," arXiv:1712.03632, 2017.

[21] E. Tabassi et al., "A taxonomy and terminology of adversarial machine learning," NIST IR, 2019.

[22] M. Jagielski et al., "High accuracy and high fidelity extraction of neural networks," in 29th USENIX Security Symp. (USENIX Security 20), 2020.

[23] M. S. J. and T. R. Fredrikson, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. 22nd ACM SIGSAC Conf. Computer and Communications Security, 2015.

[24] Z. Y. S. and Z. X. Lin, "IDSGAN: Generative adversarial networks for attack generation against intrusion detection," in Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2022.

[25] R. Chauhan and S. H., "Polymorphic adversarial DDoS attack on IDS using GAN," in 2020 Int. Symp. Networks, Computers and Communications (ISNCC), 2020.

[26] J. Chen et al., "Fooling intrusion detection systems using adversarially autoencoder," Digital Communications and Networks, vol. 3, no. 7, pp. 453–460, 2021.

[27] M. Usama et al., "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019.

[28] D. Shu et al., "Generative adversarial attacks against intrusion detection systems using active learning," in Proc. 2nd ACM Workshop on Wireless Security and Machine Learning, 2020.

[29] S. Zhao et al., "attackgan: Adversarial attack against black-box ids using generative adversarial networks," Procedia Computer Science, 2021.

[30] K. Yang et al., "Adversarial examples against the deep learning based network intrusion detection systems," in MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), 2018.
[31] P.-Y. Chen et al., "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in Proc. 10th ACM Workshop on Artificial Intelligence and Security, 2017.
[32] M. S. Arjovsky, L. Bottou, and others, "Wasserstein generative adversarial networks," in Int. Conf. Machine Learning, PMLR, 2017.
[33] N. Papernot et al., "Distillation as a defense to adversarial perturbations against deep neural networks," in 2016 IEEE Symposium on Security and Privacy (SP), 2016.
[34] C. Feng et al., "Few features attack to fool machine learning models through mask-based GAN," in 2020 International Joint Conference on Neural Networks (IJCNN), 2020.
[35] X. Zhou et al., "Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system," Internet of Things Journal, vol. 12, no. 9, pp. 9310–9319, 2021.
[36] A. Hamza et al., "Detecting volumetric attacks on IoT devices via SDN-based monitoring of MUD activity," in Proc. 2019 ACM Symposium on SDN Research, 2019.
[37] K. Xu et al., "Representation learning on graphs with jumping knowledge networks," in Int. Conf. Machine Learning, PMLR, 2018.
[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
[39] J. Ma and Q. M. Diao, "Towards more practical adversarial attacks on graph neural networks," Advances in Neural Information Processing Systems, vol. 33, pp. 4756–4766, 2020.
[40] Z. Sun et al., "In-memory PageRank accelerator with a cross-point array of resistive memories," Transactions on Electron Devices, vol. 67, no. 4, pp. 1466–1470, 2020.
[41] X. Zhou et al., "Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," Transactions on Emerging Topics in Computing, vol. 9, no. 1, pp. 246–257, 2018.
[42] B.-E. Zolbayar et al., "Generating practical adversarial network traffic flows using NIDSGAN," arXiv preprint arXiv:2203.06694, 2022.
[43] S. Alahmed et al., "Impacting robustness in deep learning-based NIDS through poisoning attacks," Algorithms, vol. 17, no. 4, p. 155, 2024.
[44] M. Ahmed et al., "Re-evaluating deep learning attacks and defenses in cybersecurity systems," Big Data and Cognitive Computing, vol. 8, no. 12, p. 191, 2024.
[45] A. G. K. Warzyński, "Intrusion detection systems vulnerability on adversarial examples," in 2018 Innovations in Intelligent Systems and Applications (INISTA), 2018.
[46] A. S. Piplai et al., "Adversarial attacks to bypass a GAN based classifier trained to detect network intrusion," in 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conf. on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 2020.
[47] M. Ayub et al., "Model evasion attack on intrusion detection systems using adversarial machine learning," in 2020 54th Annual Conference on Information Sciences and Systems (CISS), 2020.
[48] M. Rigaki, "Adversarial deep learning against intrusion detection classifiers," 2017.
[49] C.-H. Huang et al., "Adversarial attacks on SDN-based deep learning IDS system," in Mobile and Wireless Technology 2018: Int. Conf. on Mobile and Wireless Technology (ICMWT 2018), 2019.
[50] Y. Peng et al., "Evaluating deep learning based network intrusion detection system in adversarial environment," in 2019 IEEE 9th Int. Conf. on Electronics Information and Emergency Communication (ICEIEC), 2019.
[51] I. e. a. Debicha, "Adversarial training for deep learning-based intrusion detection systems," arXiv preprint arXiv:2104.09852, 2021.
[52] R. D. S. a. A. M. Maarouf, "Evaluating resilience of encrypted traffic classification against adversarial evasion attacks," in 2021 IEEE Symposium on Computers and Communications (ISCC), 2021.
[53] G. a. M. C. Apruzzese, "Evading botnet detectors based on flows and random forest with adversarial samples," in 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), 2018.
[54] G. e. a. Apruzzese, "Addressing adversarial attacks against security systems based on machine learning," in 2019 11th international conference on cyber conflict (CyCon), 2019.
[55] M. e. a. Usama, "Black-box adversarial machine learning attack on network traffic classification," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019.
[56] X. W. H. a. Z. S. Peng, "Adversarial attack against dos intrusion detection: An improved boundary-based method," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 2019.
[57] P. e. a. Li, "Chronic poisoning against machine learning based IDSs using edge pattern detection," in 2018 IEEE International Conference on Communications (ICC), 2018.

[58] J. a. S. S.-H. Aiken, "Investigating adversarial attacks against network intrusion detection systems in SDNs," in 2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), 2019.
[59] D. e. a. Han, "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," Journal on Selected Areas in Communications, vol. 8, no. 39, pp. 2632-2647, 2021.
[60] M. e. a. Fan, "Toward Evaluating the Reliability of Deep-Neural-Network-Based IoT Devices," Internet of Things Journal, vol. 18, no. 9, pp. 17002-17013, 2021.
[61] E. L. R. a. J. K. Wong, "Fast is better than free: Revisiting adversarial training," arXiv preprint arXiv:2001.03994, 2020.
[62] A. V. N. a. G. H. Krizhevsky, "The CIFAR-10 dataset," 2014.
[63] N. Q. A. a. M. A. Fatehi, "Towards adversarial attacks for clinical document classification," Electronics, vol. 1, no. 12, p. 129, 2022.
[64] K. A. Z. a. S. H. Roshan, "Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system," Computer Communications, no. 218, pp. 97-113, 2024.
[65] S. S. G. a. A. B. Kumar, "BB-Patch: BlackBox Adversarial Patch-Attack using Zeroth-Order Optimization," arXiv preprint arXiv:2405.06049, 2024.
[66] X. e. a. Tian, "Dynamic geothermal resource assessment: Integrating reservoir simulation and Gaussian Kernel Density Estimation under geological uncertainties," Geothermics, no. 120, p. 103017, 2024.
[67] D. e. a. Golovin, "Gradientless descent: High-dimensional zeroth-order optimization," arXiv preprint arXiv:1911.06317, 2019.
[68] H. e. a. Zenati, "Adversarially learned anomaly detection," in 2018 IEEE International conference on data mining (ICDM), 2018.
[69] M. e. a. Al Olaimat, "A learning-based data augmentation for network anomaly detection," in 2020 29th International Conference on Computer Communications and Networks (ICCCN), 2020.
[70] D. D. K. a. Y. O. Li, "Improving attack detection performance in NIDS using GAN," in 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), 2020.
[71] M. M. C. a. R. K. Pawlicki, "Defending network intrusion detection systems against adversarial evasion attacks," Future Generation Computer Systems, no. 110, pp. 148-154, 2020.
[72] R. M. S. a. A. M. Abou Khamis, "Investigating resistance of deep learning based IDS against adversaries using min-max optimization," in ICC 2020-2020 IEEE international conference on communications (ICC), 2020.
[73] R. a. A. M. Abou Khamis, "Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs," in 2020 international symposium on networks, computers and communications (ISNCC), 2020.
[74] I. e. a. Debicha, "Detect & reject for transferability of black-box adversarial attacks against network intrusion detection systems," in International Conference on Advances in Cyber Security, 2021.
[75] G. e. a. Apruzzese, "Hardening random forest cyber detectors against adversarial attacks," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 4, pp. 427-439, 2020.
[76] C. M. B. a. T. T. Benzaïd, "Robust self-protection against application-layer (D) DoS attacks in SDN environment," in 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020.
[77] M. e. a. Novaes, "Adversarial Deep Learning approach detection and defense against DDoS attacks in SDN environments," Future Generation Computer Systems, no. 125, pp. 156-167, 2021.
[78] G. Apruzzese et al., "Modeling realistic adversarial attacks against network intrusion detection systems," Digital Threats: Research and Practice (DTRAP), vol. 3, no. 3, pp. 1-19, 2022.
[79] A. Raghuvanshi et al., "Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming," Journal of Food Quality, vol. 1, no. 2022, p. 3955514, 2022.
[80] M. Siganos et al., "Explainable AI-based intrusion detection in the Internet of Things," in Proceedings of the 18th International Conference on Availability, Reliability and Security, 2023.
[81] A. McCarthy et al., "Feature vulnerability and robustness assessment against adversarial machine learning attacks," in 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2021.
[82] Z. Chkirbene et al., "Hybrid machine learning for network anomaly intrusion detection," in 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020.
[83] A. Mirza, "Computer network intrusion detection using various classifiers and ensemble learning," in 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018.
[84] S. Waskle et al., "Intrusion detection system using PCA with random forest approach," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.
[85] N. Aboueata et al., "Supervised machine learning techniques for efficient network intrusion detection," in 2019 28th International Conference on Computer Communication and Networks (ICCCN), 2019.

[86] A. Rai, "Optimizing a new intrusion detection system using ensemble methods and deep neural network," in 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI), 2020.

[87] E. Uysal et al., "Network anomaly detection system using genetic algorithm, feature selection and classification," in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019.

[88] A. Punitha et al., "A feature reduction intrusion detection system using genetic algorithm," in 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019.

[89] Q. Fitni et al., "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," in 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 2020.

[90] B. Serinelli et al., "Training guidance with KDD Cup 1999 and NSLKDD data sets of ANIDINR: Anomaly-based network intrusion detection system," Procedia Computer Science, no. 175, pp. 560-565, 2020.

[91] K. Bhosale et al., "Data mining based advanced algorithm for intrusion detections in communication networks," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018.

[92] A. Kumari et al., "A hybrid intrusion detection system based on decision tree and support vector machine," in 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020.

[93] S. Sarvari et al., "An efficient anomaly intrusion detection method with feature selection and evolutionary neural network," IEEE Access, no. 8, pp. 70651-70663, 2020.

[94] A. Ganesan et al., "Mitigating evasion attacks on machine learning based NIDS systems in SDN," in 2021 IEEE 7th International Conference on Network Softwarization (NetSoft), 2021.

[95] R. Yumlembam et al., "IoT-based Android malware detection using graph neural network with adversarial defense," IEEE Internet of Things Journal, vol. 10, no. 10, pp. 8432-8444, 2022.

[96] M. Roshan et al., "Boosting robustness of network intrusion detection systems: A novel two phase defense strategy against untargeted white-box optimization adversarial attack," Expert Systems with Applications, no. 249, p. 123567, 2024.