future internet

MDPI

Review

Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense

Afnan Alotaibi [1] and Murad A. Rassam [1,2,*]

[1] Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
[2] Faculty of Engineering and Information Technology, Taiz University, Taiz 6803, Yemen
* Correspondence: m.qasem@qu.edu.sa

Abstract: Concerns about cybersecurity and attack methods have risen in the information age. Many techniques are used to detect or deter attacks, such as intrusion detection systems (IDSs), that help achieve security goals, such as detecting malicious attacks before they enter the system and classifying them as malicious activities. However, the IDS approaches have shortcomings in misclassifying novel attacks or adapting to emerging environments, affecting their accuracy and increasing false alarms. To solve this problem, researchers have recommended using machine learning approaches as engines for IDSs to increase their efficacy. Machine-learning techniques are supposed to automatically detect the main distinctions between normal and malicious data, even novel attacks, with high accuracy. However, carefully designed adversarial input perturbations during the training or testing phases can significantly affect their predictions and classifications. Adversarial machine learning (AML) poses many cybersecurity threats in numerous sectors that use machine-learning-based classification systems, such as deceiving IDS to misclassify network packets. Thus, this paper presents a survey of adversarial machine-learning strategies and defenses. It starts by highlighting various types of adversarial attacks that can affect the IDS and then presents the defense strategies to decrease or eliminate the influence of these attacks. Finally, the gaps in the existing literature and future research directions are presented.

Keywords: adversarial machine learning; intrusion detection systems; adversarial attacks; machine learning; deep learning; network security

Citation: Alotaibi, A.; Rassam, M.A. Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. *Future Internet* **2023**, *15*, 62. https://doi.org/10.3390/fi15020062

Academic Editor: Franco Davoli

Received: 22 December 2022
Revised: 15 January 2023
Accepted: 29 January 2023
Published: 31 January 2023

1. Introduction

Machine learning (ML) approaches are changing our perceptions of the world and affecting every aspect of our lives in the age of technology, such as autopilot, facial recognition, and spam detection. A distinctive feature of ML is that instead of designing the solution with coding, the programmer creates a method to discover the key to a problem using samples of other issues that have been solved. ML techniques can produce satisfactory results in many situations since machine-generated features are typically more reliable and representative than hand-crafted features [1]. Furthermore, ML procedures' training and evaluation phases are generally constructed assuming they are executed in a secure environment [1]. Therefore, the use of ML has expanded drastically, especially in cybersecurity, depending on providing a secure environment for users and institutions. Furthermore, due to the efficiency of ML in automatically extracting useful information from massive databases, the use of ML in cybersecurity meets the development of cyberattacks [2].

Various network attacks affect the users' or institutions' network systems, such as denial of service (DoS), distributed denial of service (DDoS), and SQL injection. Thus, cybersecurity specialists propose and utilize various types of defensive methods against network attacks, such as firewalls, intrusion detection systems (IDS), and intrusion

prevention systems (IPSs). Such defense methods are used to detect or deter unauthorized network attacks that may affect network users in a harmful way.

Furthermore, to serve cybersecurity specialists and beneficiaries, researchers proposed building their defenses based on ML techniques to improve the defense methods since a variety of network security techniques are increasingly using ML approaches for enhancements [1], for example, an IDS. Figure 1 shows some network security applications that can be improved with ML to protect the network against cyberattacks.

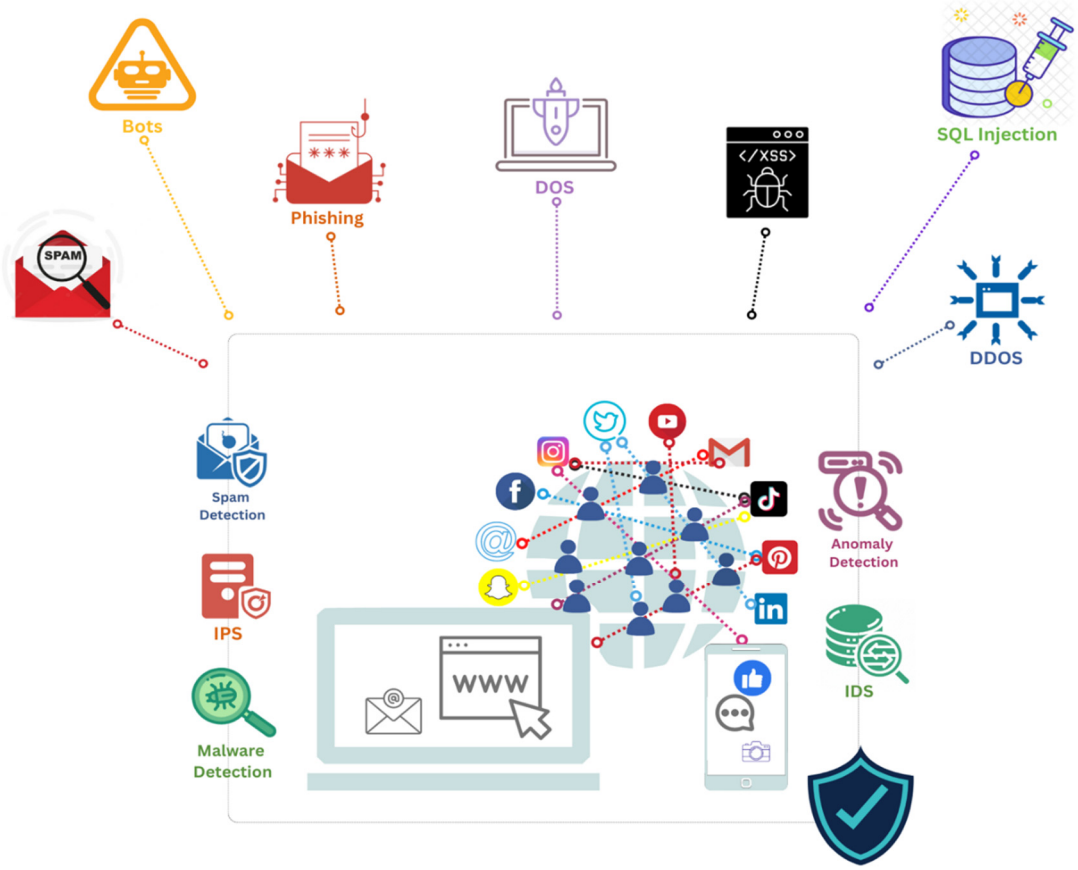

Figure 1. Network security applications against cyber attacks.

IDS is one of the cybersecurity domains where machine learning is suitable. It is a type of computer security software that seeks to identify a wide range of security breaches, from attempted intrusions by outsiders to system penetrations by insiders [2]. In addition, the IDS evaluates the information from many sources and produces alerts when specific criteria are met [3].

In order to improve the IDS and make it more reliable against network attacks, the cybersecurity specialists suggested building it with ML, which achieves an effective result in classification and assists in resolving malware detection issues. ML-based IDSs can identify system anomalies with high precision, according to [4]. Consequently, ML-based IDSs yield several benefits, including increased accuracy and the detection of new attacks [5]. Furthermore, according to [6], an ML-based IDS produces superior results, recommending a filtering approach based on a support vector machine (SVM) classifier and the NSL-KDD intrusion detection dataset to detect suspicious network intrusion.

ML systems are increasingly trusted in cyber-physical systems [7], including factories, power plants, and the oil and gas industries. In such complex physical surroundings, a threat that manages to get through a weak system could be harmful [8]. Despite the dependence on and faith in ML systems, attackers who want to avoid ML-based system

discovery processes may use the inherent nature of ML, learning to recognize patterns, as a possible attack component [9]. As a result, attackers craft malicious inputs called "adversarial samples." Adversarial samples are constructed by intentionally adding minor perturbations to initial inputs, which results in the misclassification of the ML/DL models [10]. Adversarial machine learning (AML) based on the National Institute of Standards and Technology (NIST) is divided into four attacks, which are: evasion, extraction, poisoning, and inference [11].

Hence, the misclassification of ML initially appeared approximately two decades ago and has piqued researchers' interest. The researchers in [12] deceived spam classifiers into injecting some changes into an email. Moreover, it is even older than 38 years, according to the authors in [13], when they showed that false fingerprints might be made with plastic-like materials to deceive biometric identity recognition systems.

Along with ML technology's significant advancement in network security, it exposes a new attack surface for attackers. Accordingly, the IDS is susceptible to adversarial attacks since it is built on ML, which could be compromised by crafting adversarial input against ML/DL models such as the artificial neural network (ANN), the deep neural network (DNN), and the support vector machine (SVM), affecting its accuracy and robustness. Furthermore, research has also demonstrated that adversarial samples could affect ML-based IDSs [10,14]. As a result, ML can also be fooled, necessitating some protection mechanisms. Additionally, the system becomes susceptible due to communication on the open network, which also gives enemies a massive attack surface [15].

In contrast, the adversarial sample inputs to ML deceive the model, causing the model to provide an incorrect result. Thus, IDSs based on ML may be harmed, affecting classification. Consequently, the ML classifier in cybersecurity is used to defend against and detect malicious attacks, but the big issue here is who will protect the defenders and how ML can withstand these attacks and provide correct categorization.

Therefore, this challenge drives researchers to improve the resilience of ML algorithms. This paper presents an overview of ML methods and clarifies adversarial attacks on IDSs. Additionally, it provides a thorough literature review on the security and robustness of ML/DL models when applied to the development of IDSs. Above all, it is essential to emphasize that this study aims to provide a thorough overview of the impact of adversarial samples raised by using ML and DL in IDSs and to present potential solutions to these problems. To sum up, the particular contributions of this paper are as follows:

- We analyze related surveys in the field of AML.
- We present a general overview of the use of ML on an IDS in order to enhance its performance.
- We clarify all types of adversarial attacks against ML and DL models and the differences between them, in addition to the challenges that face the launch of adversarial attacks.
- We display the adversarial attacks launched against ML/DL-based IDS models in particular.
- We present the different types of defense strategies to address adversarial attacks.
- We investigate the gaps in the related literature and suggest some future research directions.

The remainder of this paper is structured as follows: Section 2 discusses related existing surveys. Then, an overview of ML is presented in Section 3. Next, adversarial ML is introduced in detail in Section 4.

Next, the studies that implemented the adversaries against IDSs are presented in Section 5. Then, the benchmark dataset is shown in Section 6. Next, the defense strategies against adversarial attacks in the two domains of computer vision and network security are presented in Section 7. After that, the challenges and future directions are given in Section 8. Finally, this paper is concluded in Section 9.

2. Related Surveys

Many surveys present AML in various domains, such as computer vision, which recently received much attention, and network security. Major previous studies focused on adversarial attacks against ML and DL in various domains or the computer vision domain, such as in [16–18]. Additionally, other surveys take this topic from a game perspective, making it more straightforward for the reader, such as [19], which presented a general view of the arms race between adversarial attacks and defense methods and how they constantly try to defeat each other. In addition, [20] presented more details about adversarial attacks and defense methods from a cybersecurity perspective.

Furthermore, ML security has received much attention, with many researchers mentioning the dangers of adversarial attacks on ML and the defense methods described in [21]. This survey clarified the various types of adversarial attacks and the defense methods to protect ML. However, this study highlighted the ML adversaries and primary defenses; it was not specialized in specific ML methods in cybersecurity, such as malware detection.

On the other hand, the authors of [22] had to dig deeper into the network security domain. This study has more than the original view. It presents detailed information for network security applications in ML and adversarial attacks against them, in addition to defense methods against these attacks. However, it is not connected to something special such as phishing or spam detection. The research in [9] presented adversarial attacks in cybersecurity, such as intrusion detection, which provided a more detailed perspective, discussed attacks, and offered some defense strategies. In general, the researchers found this study helpful in providing the basis for the issue of adversaries and defenses against ML-based network applications. Despite this study's insightful ideas, its main focus is on keeping adversarial attacks functioning so they can continue avoiding ML classifiers.

To our knowledge, no prior survey reviewed adversarial attacks against ML/DL-based IDSs. Therefore, this survey highlights adversarial attacks made particularly against IDS and earlier research that created adversaries for ML-based IDSs. It also describes the benchmark datasets used in most of these studies. Additionally, it analyzes state-of-the-art defense strategies to improve the robustness and accuracy of ML-based IDSs and suggests using defenses applied to computer vision on ML-based IDSs. Finally, it clearly describes these adversarial attacks so that it is simple for the researchers to choose one to defend the IDS. Table 1 demonstrates the main differences between the previous surveys.

Table 1. Comparison between related surveys.

Ref.	Year	Highlights	Domain	General Contribution	Applications of ML in Network Security	Adversarial Attack's Methods	Solutions For Adversarial Attacks	Open Research Issues
[21]	2018	It examined ML system threat models and outlined alternative attack and defense strategies.	ML/DL methods in various domains.	- Attacks capabilities - Defense methods	✖	✓	✓	✖
[18]	2018	It thoroughly overviewed adversarial assaults on deep learning in computer vision.	ML/DL methods in computer vision.	- It examined the possibility of adversarial attacks against deep learning and offered countermeasures. - It presented articles that crafted adversarial attacks in the physical world.	✖	✓	✓	✓

[17]	2018	It explored some of the state-of-the-art adversarial attacks and suggested countermeasures.	ML/DL methods in various domains.	- It presented a taxonomy of adversarial-learning-related issues. - It reviewed alternative attack and threat models.	✖	✓	✓	✓
[23]	2018	It provided a thorough introduction to various topics related to adversarial deep learning.	ML/DL methods in various domains.	- It provided theoretical underpinnings for AML. - Typical offensive and defensive tactics.	✖	✓	✓	✓
[16]	2019	It gave a thorough summary of the research on the security characteristics of ML algorithms in hostile environments.	ML/DL methods in various domains.	- It analyzed the ML security model. - It presented adversarial attack techniques. - It also suggested potential future research that will be important for creating safer ML models.	✖	✓	✓	✓
[19]	2019	It provided a thorough overview of all game theory in AML.	Adversarial game-theoretic model in various domains.	- It thoroughly analyzed various game-theoretic models utilized in adversarial learning. - It also discussed creating learning algorithms that are impervious to active adversaries.	✖	✓	✖	✓
[20]	2019	It presented the current and recent methods used to strengthen an ML system against adversarial attacks utilizing the computational framework of game theory.	ML/DL methods in cybersecurity tasks.	- It concentrated on game-theory-based methods for enhancing the resistance of ML systems against adversarial attacks. - It discussed open research issues related to the capabilities of attacks, such as transferability.	✖	✓	✖	✓
[22]	2019	It introduced the taxonomy of ML in network security applications. In addition, it presented several adversarial attacks on ML in network security and provided two categorization algorithms for these assaults.	ML/DL methods in network security applications.	- It offered a novel technique for categorizing adversarial attacks in network security. - It described the adversarial risk in terms of computer and network security. - In addition, it provided defense strategies based on attack methods.	✓	✓	✓	✓
[5]	2019	It discussed building IDS with the ML and DL models, potentially improving IDS performance.	ML/DL methods in IDS.	- It defined the taxonomy and concept of IDSs. - Measurements and benchmark datasets were provided, along with the	✓	✖	✖	✓

				ML methods often employed in IDSs.			
[23]	2021	It briefly outlined the obstacles involved in using ML/DL approaches in various healthcare application domains from a security and privacy perspective.	ML/DL methods in healthcare.	- It provided potential approaches for ML security and privacy protection in healthcare applications. - It offered insight into the future directions for future research and the existing research obstacles.	✖	✓	✓	✓
[9]	2022	It presented the AML with an adversary's perspective in the cybersecurity domain and NIDS.	ML/DL methods in cybersecurity tasks.	- It provided the basis for the issue of adversaries and defenses against ML-based network applications.	✓	✓	✓	✓

3. Intrusion Detection System Based on ML

In general, machine learning techniques can be divided into two categories [5].

3.1. Supervised Machine Learning

Supervised learning depends on meaningful information in labeled data. The most common goal in supervised learning (and, therefore, in IDS) is classification. Nevertheless, manually labeling data is costly and time-intensive. As a result, the fundamental barrier to supervised learning is the lack of adequate labeled data.

3.2. Unsupervised Learning

Unsupervised learning recovers useful feature information from unlabeled data, making training material much more straightforward. On the other hand, unsupervised learning approaches often perform worse in terms of detection than supervised learning methods. Figure 2 shows the most prevalent ML techniques used in IDSs.

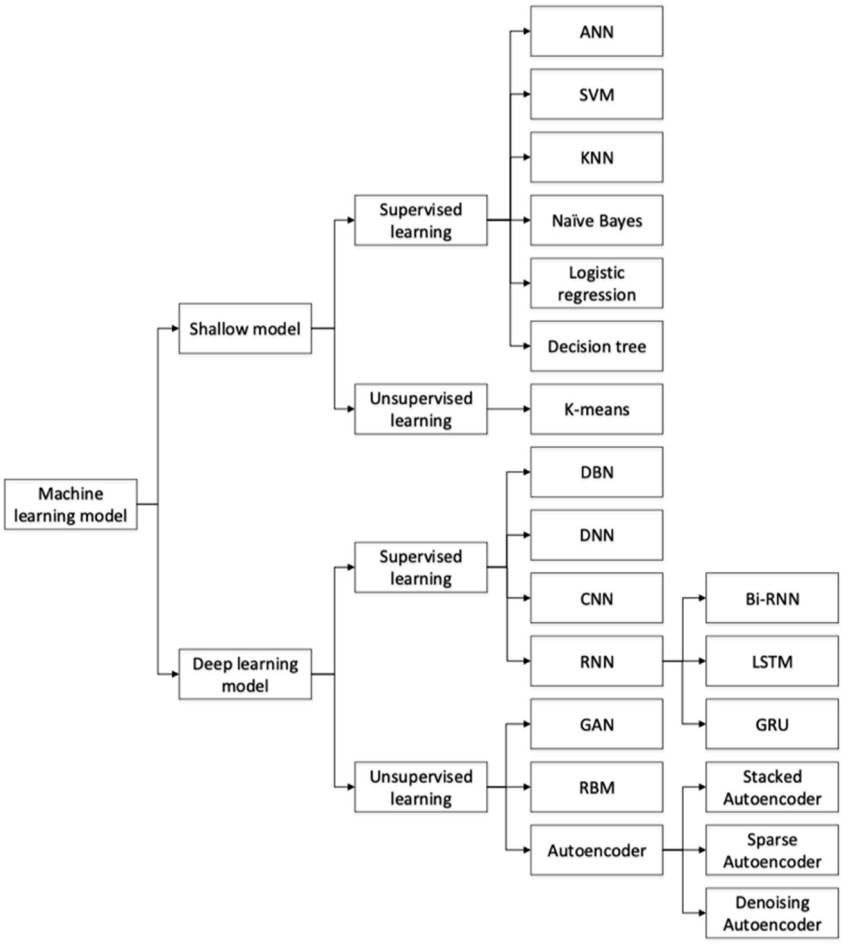

Figure 2. ML methods in IDSs [5].

IDS is a type of computer security software that seeks to identify a wide range of security breaches, from attempted break-ins by outsiders to system penetrations by insiders [2]. Furthermore, the essential functions of IDSs are to monitor hosts and networks, evaluate computer system activity, produce warnings, and react to abnormal behavior [5]. Moreover, one of the significant constraints of typical intrusion detection systems (IDS) is filtering and decreasing false alarms [24]. In addition, many IDSs improve their performance by utilizing neural networks (NN) for deep learning. Furthermore, deep neural network (DNN)-based IDS systems have been created to improve tremendous data learning, processing, and a range of assaults for future prediction [25].

Various machine learning techniques have been used to build intrusion detection models; the following paragraphs summarize the most commonly used techniques.

3.3. Artificial Neural Network (ANN)

An ANN is designed to function in the same way as human brains. An ANN comprises numerous hidden layers, an input layer, and an output layer. Units in neighboring strata are interconnected. Furthermore, it has an excellent fitting ability, particularly for nonlinear functions.

3.4. Deep Neural Network (DNN)

The parameters of a DNN are initially learned using unlabeled data in an unsupervised feature learning stage, and then the network is tweaked using labeled data in a supervised learning stage. The unsupervised feature learning step is mainly responsible for DNN's remarkable performance.

Furthermore, DNN plays a crucial role in cybersecurity; therefore, DNN could understand the abstract, high-level properties of APT assaults even if they use the most complex evasion strategies [26].

3.5. Support Vector Machine (SVM)

In SVMs, the goal is to locate a hyperplane of maximum margin separation in the n-dimensional feature space. Because a small number of support vectors control the separation hyperplane, SVMs can produce satisfactory results even with small-scale training data. SVMs, on the other hand, are susceptible to noise around the hyperplane. SVMs excel at solving linear problems. Kernel functions are commonly used with nonlinear data. The original nonlinear data can be split using a kernel function that transfers the original space into a new space. SVMs and other machine-learning algorithms are rife with kernel trickery.

3.6. Generative Adversarial Network (GAN)

A GAN model has two subnetworks, one for the generator and one for the discriminator. The generator's goal is to create synthetic data that looks like actual data, whereas the discriminator's goal is to tell the difference between synthetic and natural data. As a result, the generator and discriminator complement each other [5,27].

Furthermore, GANs are a trendy study area at present. They are being utilized to augment data in attack detection, which helps alleviate the problem of IDS dataset scarcity. GANs, on the other hand, are adversarial learning algorithms that can improve model detection accuracy by including adversarial samples in the training set.

A comprehensive survey of supervised and unsupervised learning techniques used in IDS can also be found in [5].

4. Adversarial Machine Learning

In AML, an opponent tries to trick the system into selecting the incorrect course of action. In other words, it causes the ML model to misclassify the data, producing inaccurate results. The adversarial sample is a critical element of an adversarial attack. An input to an ML model that has been altered constitutes an adversarial sample. An adversarial sample is a single data point that, for a given dataset containing attributes x and a label y, leads a classifier to predict a different label on x′ from y even if x′ is almost identical to x. One of the various optimization techniques referred to as "adversarial attack techniques" is used to produce adversarial samples. To create adversarial samples, an optimization problem must be solved to identify the minimal perturbation that maximizes loss for the neural network.

The adversary's optimization objective is to calculate a perturbation with a tiny norm that would change the classifier's output.

where the disturbance is δ [22].

Furthermore, the whole process of AML and the adversaries' samples are illustrated in Figure 3.

Figure 3. The whole process of AML [22].

This paper covers AML in two components: the adversarial game-theoretic and adversarial threat models. The adversarial threat model is detailed in three components:

adversaries' capabilities, adversaries' challenges, and potential threats, as demonstrated in Figure 4.

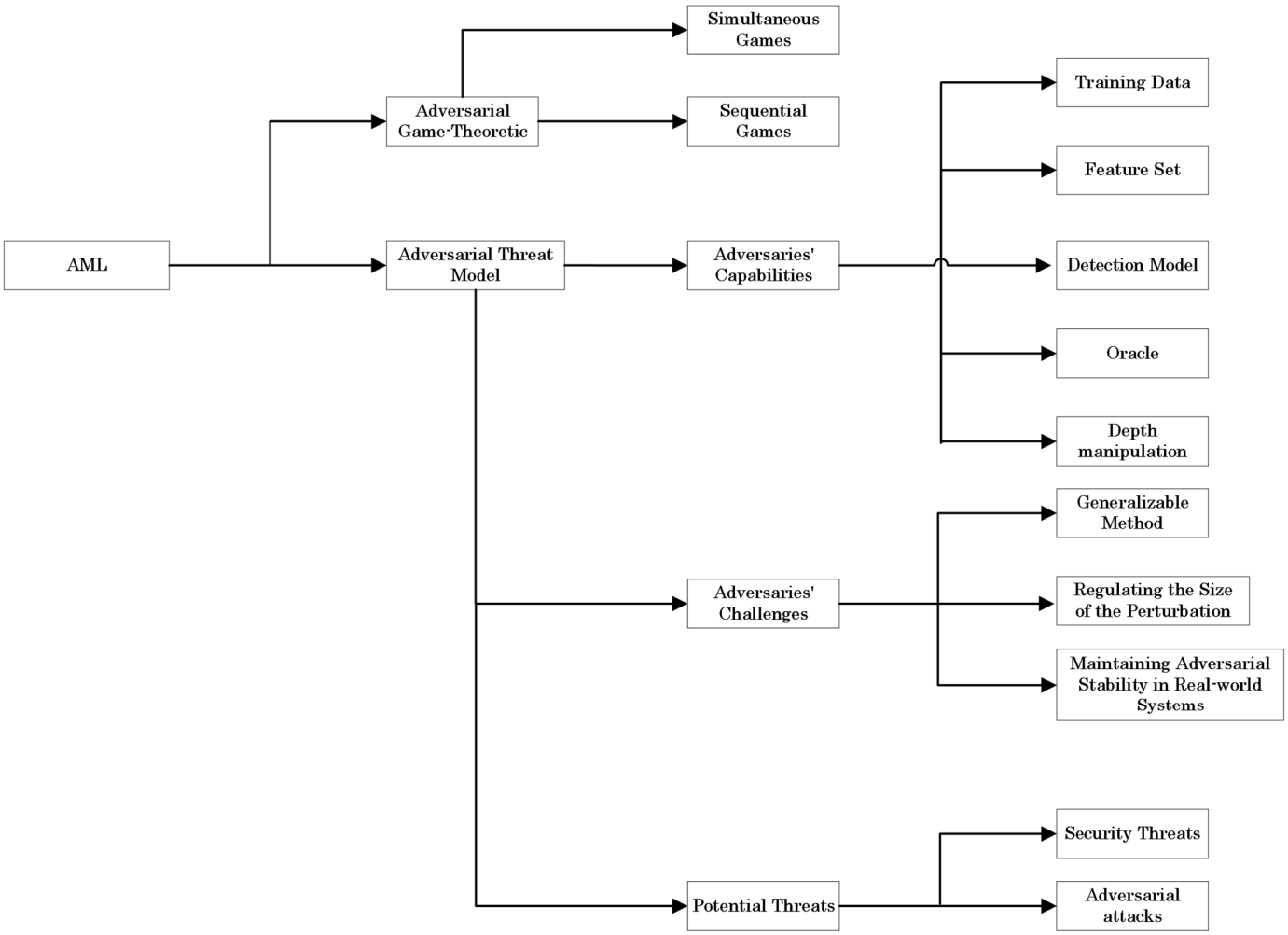

Figure 4. AML components.

4.1. Adversarial Game-Theoretic

Adversarial learning is a type of ML in which two entities, the learner and the adversary, try to develop a prediction mechanism for data relevant to a specific problem but with various goals. The goal of learning the prediction mechanism is for the learner to predict or classify the data accurately. On the other hand, the adversary's goal is to force the learner to make inaccurate predictions about inputs in the future [19,28].

Game theory is an attractive technique for adversarial learning because it allows the mathematical representation of the behavior of the learner and the adversary in terms of defensive and attack methods, as well as figuring out how to reduce the learner's loss from adversarial examples [28].

In general, the exciting objective of this game theory is to figure out how to achieve equilibrium between the two players (classifier and adversarial). To put it another way, how can we keep the adversary from influencing the classification? As a result, the fight between opponents and ML is never-ending, similar to an 'armed race' [20].

In contrast, the learner in this work, for example, is an IDS-based ML classifier that classifies traffic as "benign" or "malicious." On the other hand, the attacker creates adversarial samples to influence the IDS's accuracy, allowing it to misinterpret benign traffic as malicious and vice versa. Moreover, AML, based on a game theory perspective, is divided into two categories: simultaneous and sequential games.

In the simultaneous game, each player selects his or her approach without knowing what the other player seems to be doing. In the other game, one player takes on the role of leader and decides on a plan before the other players, who then play optimally against the leader's approach. In summary, the attacker will know what affects the model most based on recognizing the model first, i.e., manipulating the features. This type can be divided into Stackelberg games, where the adversary acts as a leader. In this game, the classifier is the follower; for example, the IDS classifier will follow the leader, which is the adversary. The classifier (follower) attempts to discover the adversary features to enhance the ability to discover the adversarial methods.

In addition to Stackelberg games, the other category is a learner as a leader. In this game, the adversary is the follower; for example, in the ML-based IDS, the adversary will follow the IDS classifier to discover its strategies to craft a suitable adversarial sample that could affect the ML-based IDS model.

In the Sequential game theory, the attacker attempts to learn about the model before crafting his attacks, which is an analogy to a white-box attack since it is based on the attacker's knowledge of the model [28].

4.2. Adversarial Threat Model

4.2.1. Adversary Capabilities

If an attacker has direct or physical access to the defense system, any cybersecurity protection can ultimately be defeated [29]. Thus, five particular things are under the attacker's control to implement adversarial attacks against the ML/DL model [30]:

Training Data:

It denotes the availability of the dataset used to train the ML models. It can be read-only, write-only, or completely inaccessible.

Feature Set:

It seeks to understand the features of the ML models used to carry out its detection. It might take the shape of complete, limited, or no knowledge.

Moreover, it is worth noting that the size of an ML model's feature set can be abused as an attack surface. The fact that an enemy can alter any feature analyzed by a model is a significant challenge [9].

The authors of [31] pointed out that large feature sets have more features, giving an adversary more opportunities to manipulate them. As a result, larger feature sets may be perturbed more quickly than smaller feature sets, which have fewer modifiable features and require more perturbation.

Detection Model:

The trained ML model included in the IDS and utilized to carry out the detection was described in detail. There may be zero, some, or all of this knowledge.

However, the detection model (ML-NIDS) demands high administrative rights. In other words, it can be allowed for a small number of carefully chosen devices [32]. Therefore, assuming that an infected host will grant the attacker access to the NIDS that holds its detection model is unreasonable [10,29]. Therefore, it is difficult for an attacker to access the detection models.

Oracle:

This component indicates the potential for receiving feedback from an attacker's input to the ML output. This input may be small, limitless, or nonexistent.

Depth Manipulation:

It represents adversary manipulation that may change the traffic volume or one or more features in the examined feature space.

As a result, these capabilities clarify that implementing the black-box attack is within reach. It is expected that it will not have the same impact as the white-box attack. Indeed, it is considered a weak attack [33].

4.2.2. Adversaries Challenges

The authors of [34] mentioned three challenges faced by creating adversarial instances:

a. Generalizable Method

Some adversarial attacks are only suited for specific ML or DL models, which means they do not fit other models.

b. Regulating the Size of the Perturbation

The adversary's size should not be too small or too large since this would impact their actual purpose.

c. Maintaining Adversarial Stability in Real-World Systems

Certain adversarial instances cannot be transformed, such as blurring.

4.2.3. Potential Threats

a. Security Threats

The following security threats in ML can be classified as adversarial attacks based on their intent to attack the ML/DL model [21,35]:

b. Influence Attack

There are two sorts of influence attacks: (1) causative, which seeks to gain control over training data, and (2) exploratory, which exploits the ML model's misclassification without interfering with the model's training.

c. Adversaries' Goals in Network Security

In this situation, the CIA triad and privacy were used because they are more appropriate for hostile categorization of the enemy's aims in the network security sector [22].

i. Confidentiality

This attack aims to acquire confidential information shared by two parties, A and B, by intercepting communication between them. This occurs in the context of AML, in which network security tasks are performed using ML algorithms.

ii. Integrity

This attack aims to affect the ML model and lead it to misclassification by implementing malicious activities without interfering with regular system functions, but the attacker chooses the model's output, increasing the false-negative rate. This includes a poisoning attack that affects the training data.

iii. Availability

During operations, the adversary compromises the system's functioning to deny service to users. One method is to increase the misclassification or dramatically alter the model's output to decrease its performance and cause it to crash, increasing the false-positive rate.

iv. Privacy

The adversary attempts to acquire sensitive user data and essential knowledge about the model architecture from the ML model. For example, equation-solving attacks [36] may be used against cloud services that offer ML through APIs and models such as multilayer perceptrons, binary logistic regression, and multiclass logistic regression. The attacker should be able to learn about the model and its architecture as a result of these attempts. Moreover, privacy attacks can be categorized as "model inversion attacks" and "membership inference attacks." Furthermore, inversion attacks are divided into two categories. The first attack uses the person's unique label generated by a facial recognition system to rebuild a face image. The second assault can obtain the victim's identity by extracting a clean image from a blurred image by attacking the system [37,38].

The membership inference attack has access to the model as a black-box attack (this attack will be presented in the following section). It is determined only if a data point is part of the training data for a learning model. Furthermore, in [39], the authors were able to create membership inference attacks against the Google prediction API and Amazon ML to identify whether a data point belonged to the training set.

d. Attack Specificity

An attack's specificity may be described in two ways. First: targeted assault, this attack is aimed at a single input sample or a group of samples, and an adversary attempts to impersonate an authenticated person in a facial recognition/biometric system [40]. Second: not-targeted attacks, the ML model in this attack fails randomly. Additionally, non-targeted assaults are more straightforward to execute than targeted attacks because they offer more options and room to reroute the output [35]. Moreover, Figure 5 clarifies a categorization of the potential threats against ML.

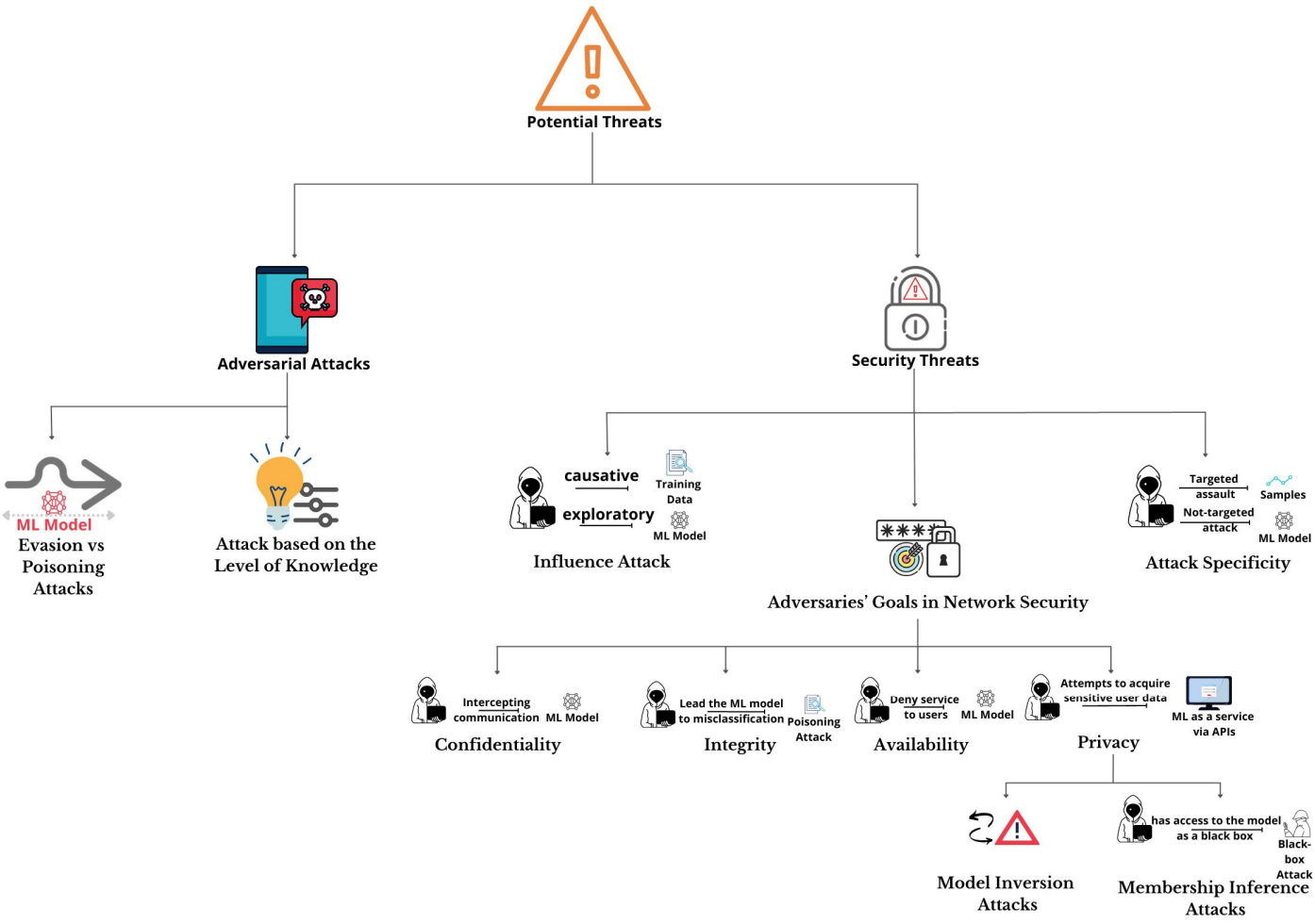

Figure 5. Categorization of security threats against ML.

e. Adversarial Attacks

Attacks based on the Level of Knowledge

Adversarial samples can be created using a variety of approaches. These approaches vary in complexity, speed of creation, and performance. Manual perturbation of the input data points is a crude method of creating such samples. On the other hand, manual perturbations of massive datasets are time-consuming to create and may be less precise. Automatically assessing and finding characteristics that best differentiate between target values is one of the more complex ways. These characteristics are discretely disturbed to reflect values comparable to those representing target values different from their own [28]. Moreover, adversaries may fully understand the ML system or have a limited understanding.

i. White-Box Attack

This is frequently the case when the ML model is open source, and everyone has access. Thus, in this attack, there is a thorough understanding of the network architecture

and the parameters that resulted from training. Furthermore, four of the most well-known white-box assaults for autonomously creating perturbed samples are [37]:

- *Fast Gradient Sign Method (FGSM)*

The fast gradient sign method (FGSM) was proposed by [41] as a fast method for producing adversarial samples. At each pixel, they perform a one-step gradient update in the direction of the gradient sign. However, this attack includes changing the value of each feature in the input concerning the neural networks. Its focus is on rapidly generating adversarial samples; therefore, it is not regarded as a powerful assault [42].

- *Jacobian-Based Saliency Map Attack (JSMA)*

The authors in [43] devised a Jacobian-based saliency map attack (JSMA), which is an excellent saliency adversarial map under L0 distance. The most influential characteristics are used when modest input variances cause substantial output changes.

- *CW Attack*

Carlini and Wagner [40] devised a tailored approach to avoid defensive distillation. Most hostile detection defenses are vulnerable to CW attacks. The details of this attack can be found in [37,44].

- *DeepFool*

Moosavi-Dezfooli suggested DeepFool in [45] to discover the shortest distance between the original input and the judgment boundary of adversarial cases.

- *Basic Iterative Method (BIM)*

This method is responsible for carrying out gradient calculations repeatedly in small steps; it expands the FGSM. To prevent significant changes in traffic characteristics, the value of the perturbation is trimmed [46].

Furthermore, an experimental study presented these methods utilized in crafting adversaries in detail; it can be found in [47]. Generally, when deciding on an adversarial attack, the authors in [48] stated there is a trade-off. JSMA, for example, uses more computing resources than FGSM but alters fewer features. In addition, DeepFool-based techniques can be considered potent adversaries [49].

ii. Black-Box Attack

In this attack, assume no prior knowledge of the paradigm and analyze the paradigm's vulnerability using information from the settings or previous inputs.

Furthermore, two ways are utilized to learn more about the classification algorithm. Firstly, the attacker might alter the malicious samples several times until they are misclassified to identify the model's parameters to differentiate between malware and benign samples. The attacker can also create a substitute model of the detection system and then use the transferability aspect of ML to create adversarial samples that fool both the substitute classifier and the actual detector [30,50].

Furthermore, the authors in [51] implement a black box attack against an ML model by developing a different model to take the place of the target ML model. Thus, the substitute model crafts adversarial samples based on understanding the substitute model and the migration of adversarial samples. Moreover, black-box attacks include [52,53]:

- *Zeroth-Order Optimization (ZOO)*

ZOO does not compute the gradient directly. Instead, ZOO used the symmetric difference quotient approach to estimate the gradient, which resulted in a higher computing cost. To estimate the gradient, knowledge of the structure of the DNN network is not necessary [37].

- *OnePixel*

This attack deceives a DNN without understanding its network topology by modifying the value of only one pixel of a clear picture. DNN is vulnerable to very-low-dimension attacks with minimal information [37].

iii. Gray-Box Attack

To grow from the black box to the white box, the adversary undergoes an iterative learning process that uses inference processes to gather additional understanding of the model. Thus, it may have partial knowledge of the model.

When knowledge is restricted, such as in gray-box and black-box scenarios, privacy attacks might be performed to learn more about the targeted ML classifier [54]. Figure 6 demonstrates adversarial attacks that depend on the level of knowledge.

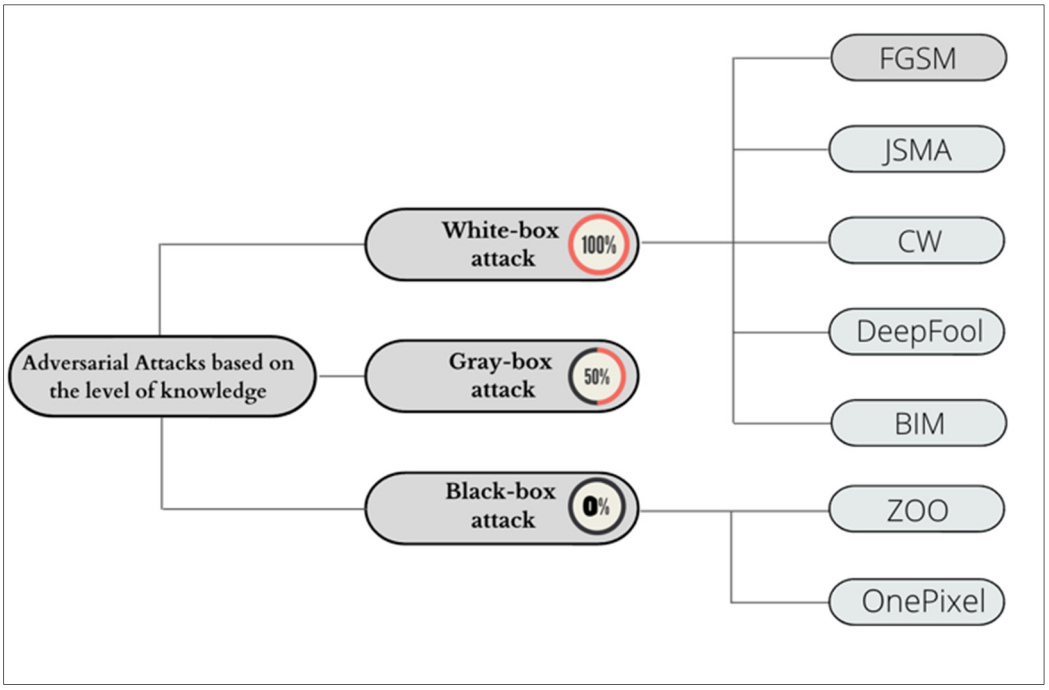

Figure 6. Adversarial attacks depend on the level of knowledge.

f. Evasion vs. Poisoning Attacks

- *Evasion Attack*

Avoid the system by injecting adversarial samples, which do not affect the training data [17]. Thus, the objective is to misclassify malware samples as benign while the model operates [30,55]. Furthermore, evasion attacks can be classified as (1) error-generic evasion attacks, in which the attacker is interested in deceiving classification regardless of what the classifier predicts as the output class. (2) error-specific evasion attacks; the attacker seeks to deceive classification but misclassifies the adversarial samples as a specific class [54].

- *Poisoning Attack*

In cybersecurity, adversarial assaults are created using a thorough grasp of computer systems and their security rules. One of the six types of attacks against intrusion detection systems is poisoning [56]. Thus, in this attack, the enemy seeks to contaminate the training data by introducing precisely planned samples, ultimately jeopardizing the learning process [17]. Furthermore, poisoning attacks can be classified as (1) error-generic poisoning attacks, in which the attacker attempts to cause a denial of service by causing as many classification errors as possible. (2) Error-specific poisoning attacks: in this situation, the attacker's goal is to produce certain misclassifications. Figure 7 illustrates these two attacks and their ways of affecting ML [54].

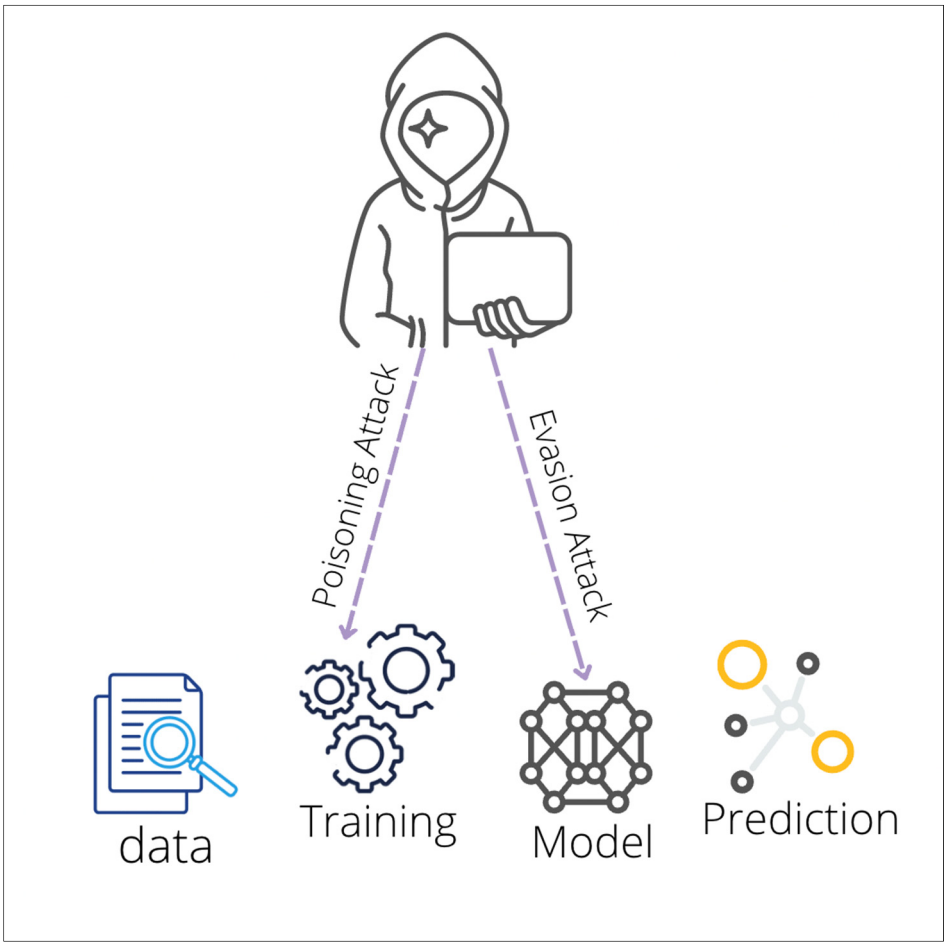

Figure 7. Evasion and poisoning attacks.

5. Machine Learning Adversaries Against IDS

Accordingly, we concluded that ML could also be fooled, necessitating some protection mechanisms. Thus, the research on AML is divided into two categories. One category is the continuous development of new attacks to counter existing ML algorithms and systems. On the other side, the second category strives to dramatically increase ML techniques' ability to withstand adversarial attacks. Therefore, this section focuses on the first category. Thus, related works on adversarial attacks against IDS are presented in this section. In addition, we discuss the various types of perturbations and how they affect IDS.

5.1. White-Box Attacks against IDS

5.1.1. A White-Box Attack against MLP

In [14], the authors presented an application of an evasion attack in a white-box setting by using a Jacobian-based saliency map attack (JSMA) against an MLP (multilayer perceptron) model, which is considered an IDS-ML by using two different datasets: CICIDS [57] and TRAbID [58] to classify network traffic. Furthermore, the adversary creates hostile samples with minor differences from the actual testing samples to deceive the MLP model during testing and to prove that the attackers can exploit the vulnerabilities to escape intrusion detection systems and misclassification. Moreover, the MLP model achieved an accuracy of almost 99.5% and 99.8% in detecting malware intrusions. Despite this success, the precision was down by about 22.52% and 29.87% for CICIDS and TRAbID, respectively, after applying an evasion attack.

5.1.2. A White-Box Attack against DNN

The research in [59] crafted adversarial attacks against DNN-based intrusion detection systems to evaluate the robustness of the DNN against these attacks. Furthermore, the authors used SDL-KDD datasets containing standard and Five-categorized attack samples. Moreover, comparing the DNN performance in classification with SVM resulted in similar performances. As a result, the authors concluded that the attacks designed by FGSM and projected gradient descent (PGD) could notably affect the DNN model.

5.1.3. A White-Box Attack against IDS in Industrial Controlling Systems (ICS)

The authors of [60] presented experimental research about adversarial attacks against IDS in industrial control systems (ICS). First, they crafted adversarial samples by using a Jacobian-based saliency map. Second, they evaluated the IDS (Random Forest, J48) after exposure to these adversarial samples. Finally, they suggested some solutions for enhancing the robustness of IDSs against adversarial attacks in the practical module of adversarial attacks against IDS methods. Furthermore, the authors' dataset used in this research was initiated based on a power system.

Additionally, this attack was crafted by insiders, for instance, administrators. As a result, the attacker was already aware of the classifier system. According to the authors, the random forest and J48 performance had decreased. In addition, the J48 achieved a lower level of robustness than the random forest model, which achieved a high level of robustness in facing these adversarial attacks. The authors applied these adversarial attacks on two IDSs, which may not affect other MLs. The authors recommend that these attacks be used on other IDS-ML systems.

5.1.4. Monte Carlo (MC)

Researchers in [11] simulated a white-box assault named a Monte Carlo (MC) simulation for the random generation of adversarial samples and compared their samples to several machine-learning models to clarify performance across a wide range of platforms and detect the vulnerability in NIDS, which assists the organizations in protecting their networks. Moreover, this research used three adversarial attack methods, which are: particle swarm optimization (PSO), genetic algorithm (GA), and generative adversarial network (GAN). In addition, the researchers used the NSL-KDD and UNSW-NB15 datasets for evaluation. Then, they confirmed that these two techniques could deceive 11 ML models. Based on their findings, the MLP model has the best accuracy under adversarial attacks with an 83.27% classification rate, followed by the BAG at 80.20% and the LDA at 79.68% with the NSL-KDD dataset. In particular, attackers with knowledge of a target network NIDS may use the most efficient perturbation process to attack that network.

5.2. Black-Box Attacks against IDS

5.2.1. FGSM, PGD, and CW Attacks against GAN

The research in [61] presented the contribution of generated adversarial network (GAN) attacks against black-box IDS. GAN's main contribution is misclassifying between actual and adversarial samples. Furthermore, this research compared the impact of the GAN attack with the fast gradient sign method (FGSM), project gradient descent (PGD), and the CW attack (CW). According to the authors, a GAN attack achieved a high rate of compromise and misclassification on IDS. Moreover, they evaluated the research using the NSL-KDD [62] dataset. Moreover, the experiments resulted in a higher rate of GAN attacks: about 87.18% against NB compared to other attack algorithms.

5.2.2. Deceiving GAN by Using FSGM

A research paper [63] proposed crafting adversarial attacks to deceive network intrusion detection systems. They trained the GAN classifier and made it robust against adversarial attacks. This research used GAN for two reasons. First, to generate adversarial

samples. Second, to train the neural network and improve its performance by increasing accuracy.

Furthermore, the GAN discriminator was evaluated for distinguishing the samples generated from the generator and classifying them as "attack" or "non-attack." Then, GAN was deceived by using the fast-sign gradient method (FSGM). As a result, the GAN classifier had been deceived by adversarial attacks and misclassified the "attack" samples as "non-attack." As a limitation, the authors did not mention how to address these attacks and defend against them.

5.2.3. A Black-Box Attack Using GAN

The authors in [64] crafted black-box attacks using GAN to improve the performance of IDS in detecting adversarial attacks. This experimental research used the KDD99 dataset. Furthermore, they trained the IDS models to detect all kinds of attacks by using GAN since IDSs have difficulty facing new attacks. Moreover, they compared the IDS's performance before the attacks, during the attacks, and after the GAN training. As a result, the GAN training increased the performance of the IDS. As a limitation, this GAN training worked only for IDSs and did not evolve for other networks.

5.2.4. IDSGAN

Researchers in [65] devised a black-box attack against an IDS to evade detection. The model's objective is to provide malicious feature records of the attack traffic that can trick and evade defensive system detection and, ultimately, direct the evasion assault in real networks. The IDSGAN evaluation demonstrated its effectiveness in producing adversarial harmful traffic records of various attacks, effectively lowering the detection rates of various IDS models to near zero.

5.2.5. DIGFuPAS

The DIGFuPAS module was presented in [66] that crafted adversarial samples using a Wasserstein GAN (WGAN) attack to deceive an IDS in SDN (software-defined networks) in a black-box manner. In addition, they compared nine ML/DL algorithms using two datasets: NSL-KDD and CICIDS-2018. If the detection capability of IDS in SDN deteriorates, they propose adding DIGFuPAS. More specifically, DIGFuPAS-generated assaults were used to repeatedly train the IDS to tackle new threats in SDN-enabled networks preemptively and to evaluate the resilience of the IDS against altered attack types. DIGFuPAS might easily fool the IDS without revealing the classification models' information, according to this experimental research.

5.2.6. Anti-Intrusion Detection AutoEncoder (AIDAE)

A research work [67] presented a novel scheme named anti-intrusion detection autoencoder (AIDAE) for adversarial features to deceive an IDS by using GAN. This experimental research used three datasets: NSL-KDD, UNSW-NB15, and CICIDS-2017. Furthermore, this research evaluated the performance of the IDS facing adversarial attacks and enhanced its robustness. According to the authors, the AIDAE model crafted adversarial attacks that evaded IDSs. Furthermore, the authors did not mention a defense method against this attack.

5.2.7. DDoS Attack by Using GAN

To highlight the IDS vulnerabilities, the authors of [68] devised a DDoS attack using GAN to fool the IDS and determine the robustness of the IDS in detecting DDoS attacks. Additionally, they improved the training of the IDS for defense. Thus, the authors conducted their experiment in three stages. First, they deceived the black-box IDS by generating adversarial data. Then, they trained the IDS with the adversarial data. Finally, they created adversarial data in order to deceive the IDS.

Moreover, this research was evaluated using the CICIDS2017 dataset [57]. According to the experiment results, transmitting attack data without being detected by an IDS was quite successful. As a limitation, it must be used in various attacks to fool any IDS.

Table 2 demonstrates the previous studies in nine columns: reference, year, adversarial generating method, objectives, (ML/DL) technique, dataset, evaluation metrics, limitations, and results. Moreover, most of these studies focused on applying GAN to create some adversarial attacks against an IDS, then evaluating the classification accuracy of the IDS in detecting cyberattacks. Additionally, almost all these works used the known evaluation metrics in an IDS, which are accuracy (ACC), precision rate (PR), recall rate (RR), and F1-score. These studies provided a decent overview of the subject by describing various well-known adversarial attacks and their effectiveness against an IDS.

Table 2. Summary of Adversarial attacks against IDS.

Ref.	year	Adversarial Generating Method	Objectives	(ML/DL) Technique	Dataset	Evaluation Metrics	Limitations	Results
[67]	2019	AIDAE	It evaluated the performance of the IDS facing adversarial attacks and enhanced its robustness.	- Logistic regression (LR). - K-nearest neighbor. - Decision tree. - Random forest.	- NSL-KDD [69] - UNSW-NB15 - CICIDS2017	- Detection rate (DR). - Evasion increase rate (EIR).	The authors did not mention a defense method against this attack.	The AIDAE model crafted adversarial attacks that evade IDSs.
[64]	2019	GAN	It made the performance of the IDS more robust in detecting adversarial attacks.	- Logistic regression (LR). - Support vector machine (SVM). - K-nearest neighbor (KNN). - Naïve Bayes (NB). - Random forest (RF). - Decision trees (Dt). - Gradient boosting (GB).	- KDD99	- Accuracy. - Precision. - Recall. - F1 score.	This GAN training worked only for IDSs and did not evolve for other networks.	The GAN training increased the performance of IDSs.
[63]	2020	FSGM	Training the GAN classifier and making it robust against adversarial attacks.	- GAN	BigData 2019 Cup: Suspicious Network Event Recognition challenge[70].	- Precision. - Recall. - F1 score.	The authors did not mention how to address and defend against these attacks.	The GAN classifier had been deceived by adversarial attacks and misclassified the "attack" samples as "non-attack."
[14]	2020	Jacobian-based saliency map	It proved that the attackers could exploit the	- MLP	- CICIDS2017 [57] - TRAbID [58]	- Precision. - Recall.	There were no experiments on defense	The accuracy of the IDS classifier dropped to 22.52% and 29.87% for

		attack (JSMA).	vulnerabilities to escape from intrusion detection systems.			- F1 score.	methods implemented by the researchers.	CICIDS [57] and TRAbID [58] datasets.
[68]	2020	GAN	To highlight the IDS vulnerabilities.	- Decision tree (DT). - Random forest (RF). - Naive Bayes (NB). - Logistic regression (LR).	CICIDS2017 [57]	- Precision. - Recall. - F1 score.	It is required to be used in a variety of attacks in order to fool any IDS.	Transmitting attack data without being detected by the IDS was quite successful.
[59]	2020	- FGSM - BGD	Evaluating the robustness of DNN against adversarial attacks.	- DNN	- SDL-KDD[69]	- Accuracy (ACC).	It lacked extract comprehensive information.	The attacks could notably affect the DNN model.
[11]	2021	- PSO. - GA. - GAN.	Detecting the vulnerability in NIDS which assists organizations in protecting their networks.	- NIDS.	-NSL-KDD [69] -UNSW-NB15	- Accuracy (ACC).	In the NIDS scenarios, it was unclear why some were more resilient than others.	The MLP model had the best accuracy under adversarial attacks with an 83.27% classification rate.
[61]	2021	-GAN.	Proving that using GAN to attack IDS can achieve a higher rate of compromission and misclassification.	- Support vector machine (SVM). - Decision tree (DT). - Random forest (RF). - Naive Bayes (NB). - Deep neural network (DNN).	-NSL-KDD [69]	- Detection Accuracy. - Attack success Rate. - Evad e increase rate.	It lacked defense mechanisms.	The experiments resulted in a higher rate of GAN attacks: about 87.18%.
[66]	2021	-DIGFuPAS	Evaluating the resilience of the IDS against altered attack types.	- Support vector machine (SVM) - Naive Bayes (NB) - Multilayer perceptron (MLP) - 50 logistic regression (LR) - Decision tree (DT) - Random forest (RF) - K-nearest neighbor (KNN)	- NSL-KDD - CI-CIDS2018	- Detection rate (DR). - Evasion increase rate (EIR).	The IDS robustness issues were not addressed.	DIGFuPAS might easily fool the IDS without revealing the classification models' information, according to this experimental research.

				- Convolutional neural networks (CNN) - Recurrent neural networks (RNN)				
[60]	2021	(JSMA)	Presenting the practical module of adversarial attacks against IDS methods.	- Random forest (RF) - J48	- Power system.	- Precision rate (PR). - Recall rate (RR). - F1-score.	The authors applied these adversarial attacks on two IDSs, but it might not affect other ML.	The random forest and J48 performance decreased.
[65]	2022	GAN	Providing malicious feature records of the attack traffic that can trick and evade defensive system detection.	- Support vector machine (SVM) - Naive Bayes (NB) - Multilayer perceptron (MLP) - Logistic regression (LR) - Decision tree (DT) - Random forest (RF) - K-nearest neighbor (KNN)	- NSLKD	- Detection rate (DR). - Evasion increase rate (EIR).	–	The assessment of IDSGAN demonstrated its efficiency in producing adversarial harmful traffic records of various assaults, bringing down the detection rates of various IDS models to almost 0%.

6. Benchmark Datasets

This section clarifies the datasets that were frequently used in previous studies. Since the distribution, quality, quantity, and complexity of dataset training samples impact the trust and quality of a model, it is essential to think about the dataset on which models are trained [71]. In virtually all previous studies, the NSL-KDD dataset [69], the UNSW-NB15 dataset [72], and the CICIDS2017 dataset [57] were used to evaluate IDS models. Furthermore, NSL-KDD contains 148,517 samples, UNSW-NB15 has 2,540,044 samples, and CICIDS2017 has 2,827,829 samples. Table 3 demonstrates these datasets, including their features and classes.

Table 3 below demonstrates the IDS's datasets that contain an imbalanced number of records in each class. In contrast, this imbalance may impact how the ML-based IDS model classifies all classes since the model's accuracy may have reduced after training on the imbalanced dataset. This restriction might be overcome through adversarial training by using GANs to increase the number of cyberattacks [73].

Table 3. Datasets used for IDS studies [74–76]

Dataset	Features	Classes	Number of Records
NSL-KDD [69]	- Basic features of network connections. - Content-related traffic. - Time-related traffic. - Host-based traffic.	- Normal. - Denial of service (DoS). - Probe. - User to root (U2R). - Remote to local (R2L).	- 77,054 - 53,385 - 14,077 - 3749 - 252
UNSW-NB15 [72]	- Basic features of network connections. - Content-related features. - Time-related features. - General-purpose features. - Connection-based features.	- Normal - Fuzzers - Analysis - Backdoors - DoS - Exploits - Generic - Reconnaissance - Shellcode - Worms	- 2,218,761 - 24,246 - 2677 - 2329 - 16,353 - 44,525 - 215,481 - 13,987 - 1511 - 174
CICIDS2017 [57]	- Basic features of network connections. - Features of network packets. - Features of network flow. - Statistic of network flows. - Content-related traffic features. - Features of network sub-flows. - General-purpose traffic features.	- Normal - DoS Hulk. - PortScan. - DDoS. - DoS GoldenEye. - FTP-Patator. - SSH-Patator. - DoS slow loris. - DoS slowhttptest. - Bot. - Web attack—brute force. - Web attack—XSS. - Infiltration. - Web attack—SQL injection. - Heartbleed.	- 2359087 - 231072 - 158930 - 41835 - 10293 - 7938 - 5897 - 5796 - 5499 - 1966 - 1507 - 652 - 36 - 21 - 11

7. Defense Strategies

Several research papers [14,65,67] have mentioned many types of adversarial attacks against IDSs. On the one hand, the authors have presented such attacks and their impact on IDS-ML, which meets the first category of AML research. On the other hand, some researchers have mentioned some methods for hardening the ML-based IDS against these attacks, but there are no experiments on these defenses in the adversarial attacks against the IDS section. Therefore, this section tackles the second category, which is defense methods.

In general, these studies provided a great perspective on IDS cybersecurity, which is a critical topic. To sum up, all the studies in the adversarial attacks against IDS section focused on generating some attacks and then clarifying the impact of adversarial attacks on IDS accuracy. Thus, this section clarifies the most state-of-the-art defense strategies

used to protect the ML/DL algorithms from adversaries. The defense strategies can be divided into many primary categories, and here we present some of them in detail [18].

7.1. Changing the Training Procedure and Input Data

Continuously inputting various types of hostile data and undertaking adversarial training improve the robustness of a deep network [77].

7.1.1. Adversarial Training

The fundamental goal of adversarial training is to increase the regularity and robustness of a DNN [37]. Moreover, in training, adversarial samples are used, and fresh adversarial samples are generated at every stage of the process [41,78,79]. More precisely, the adversarial training that is accomplished on some models can improve the accuracy of pre-trained models.

The researchers in [80] suggested a solution for AML detection. This paper measured the performance of intrusion-detecting algorithms after being exposed to four different attack methods: the fast gradient sign, the primary iterative method, the Carlini and Wagner attacks, and the projected gradient descent created by the researchers by putting five different ML classifiers under the test. Then, they implemented a method for detecting such attacks as a new way of dealing with adversarial attacks on artificial neural networks (ANN). As a result, this study recalled 0.99 for adversarial attacks using random forest and the nearest neighbor classifier. Nevertheless, significant reductions in the false positive rate are critical for the method's future development.

Moreover, we have listed a few defensive techniques that fall under the adversarial training scope as follows:

1) ZK-GanDef

The authors in [33] proposed a defense strategy called zero-knowledge adversarial training defense (ZK-GanDef) to defend against adversarial attacks in neural networks (NN). Additionally, this approach enhanced the accuracy by 49.17% against adversarial attacks compared to other attacks.

2) AFR

To assess the resilience of ML-based NIDS, the authors in [81] performed the first comprehensive investigation of gray-/black-box traffic-space adversarial assaults. Moreover, they implemented an attack on NIDS and suggested an adversarial feature reduction (AFR) method, which reduced the attack's efficacy by reducing adversarial feature development. This study also demonstrated the need to consider an attacker's capacity to mutate traffic. To sum up, the attackers can affect NIDS even if they do not have a precise understanding of the characteristics utilized by them. The findings of this experimental research clarified that the creation of adversarial features could be reduced via adversarial feature development (AFR). Moreover, the attack achieved a rate of more than 97% in half of the cases, and the proposed defense technique might successfully minimize such attacks. In addition, AFR could not prevent attackers from exploiting the vulnerable feature during traffic mutation. According to the authors, this attack technique was intended to evade NIDS without paying attention to the payload. Therefore, it was ineffective for systems that use payload-based detection. Additionally, this attack is now unavailable online.

3) APE-GAN

A study [82] presented a new idea for defense against eliminating adversarial perturbations in deep neural networks (DNNs) named APE-GAN. Furthermore, there were two ways to defend against adversaries: first, training the data to strengthen the model, and second, replacing the learning strategies. Thus, this research focused on training using GAN, which includes a generator and discriminator. First, they generated adversarial samples and then used the discriminator to discriminate those samples. The main goal of this research was to use a trained network to remove the adversarial perturbation before

feeding the processed sample to classification networks. Moreover, the researchers inferred that the APE-GAN has many applications because it works despite no understanding of the model on which it is based.

There is no mechanism to prevent the model from generating confident judgments. Thus, the first defense approach was to enrich the training set with samples altered using Gaussian noise to diminish the confidence of doubtful regions. Additionally, inserting random scaling of training photos can lower the severity of assaults, according to [83].

7.1.2. Preprocessing

Carefully planned preprocessing processes were also developed to limit the influence of adversarial perturbations. In this regard, a study [84] presented feature squeezing by spatial smoothing or pixel color bit depth reduction. In addition, picture modifications, such as total variance reduction and image quilting, were revealed to assist in removing adversarial perturbations, according to [85]. Additionally, in [86], the authors recommended that adversarial samples be denoised before being fed into a classifier using a GAN.

Furthermore, we have also included a list of defensive tactics that fall under the preprocessing umbrella as follows:

1) ME-Net

The research in [87] presented a defense technique named matrix estimation (ME-Net) to deal with the adversarial samples in deep neural networks (DNNs). This was achieved by taking incomplete or damaged images and eliminating the noise from these images to eliminate the adversarial examples from the original pictures affecting the classification performance. Thus, there were two stages for the image before it was processed: first, arbitrary pixels were discarded from the picture, and then the picture was rebuilt using ME. According to the authors, the results showed that the ME-Net had made the deep neural networks more robust against adversarial attacks than other methods.

2) DIPDefend

A research work in [88] presented a defense technique named "deep image prior driven defense" (DIPDefend) to remove adversarial examples from the image before passing the image into the classifier. Furthermore, this method was distinct for its adaptability to different types of attacks. Thus, it examined the internal prior of the image and then divided it into two steps: robust feature learning and non-robust feature learning. Additionally, it reconstructed the image by beginning with robust features and then non-robust features to make them stronger against adversarial attacks. According to the authors, the DIPDefend strategy yielded better visual results by eliminating adversarial disturbance while preserving picture information. It is worth noting that the DIPDefend technique can be applied without pretraining, making it useful in various situations.

3) Stochastic Transformation-based Defenses

The authors in [89] proposed an improved method based on transformation to extract the features of the clean images. Moreover, they employed two transformation-based approaches that are already in use: pixel deflection (PD) [90] and the image random resize and pad (RRP) [91]. Furthermore, they investigated the impact of random image alterations on clean pictures to understand better how accuracy deteriorates. They trained a unique classifier to identify distinguishing characteristics in the distributions of softmax outputs of converted clean pictures and predict the class label. Additionally, untargeted assaults on CNN have been studied, and it would be interesting to compare their distribution classifier approach with targeted attacks.

7.2. Adding an Extra Network

This defense idea utilizes specific external models as network add-ons while identifying samples that have not been shown yet [77].

Researchers in [92] developed a methodology for defending against adversarial assaults utilizing universal perturbations. The basic concept behind this strategy is to combine the original model with a second trained network to create a solution that does not require adjustment measures and cannot impact the sample.

7.2.1. Detection

Many detection methods have been proposed to detect adversarial attacks. Consequently, in [93], the authors advocated employing a subnetwork as a detector. In contrast, the authors in [94] used a confidence score to identify antagonistic and out-of-class data. In addition, to find adversarial samples that differed from the clean picture distributions, the authors in [95] applied statistical hypothesis testing. In addition, here, we mention a few defensive techniques that rely on an additional model's detection as follows:

1) Def-IDS

A study in [96] proposed Def-IDS, which includes two models, multiclass generative adversarial network (MGAN) and multisource adversarial retraining (MAT). It is a defense strategy against known and unknown adversarial attacks against NIDS to enhance the robustness of NIDS accuracy through training. Moreover, they used CSE-CIC-IDS2018 datasets to evaluate the effectiveness of the frameworks that had been proposed. Furthermore, this research used four methods to generate the adversarial attacks: the fast gradient sign method (FGSM), the basic iterative method (BIM), DeepFool, and the Jacobian-based saliency map attack (JSMA). The experiments showed that the Def-IDS could increase the robustness of NIDS by enhancing the accuracy of detecting adversarial attacks.

2) ASD

The research outcome in [97] presented the adversarial sample detector (ASD) module, which is considered a defense algorithm based on the bidirectional generative adversarial network (BiGAN) to classify the adversarial samples of NIDS-ML. It successfully reduced adversarial attacks and influenced NIDS performance. Moreover, the researchers used attack methods such as the fast gradient sign method (FGSM), projected gradient descent (PGD), and momentum iterative–fast gradient sign method (MI- FGSM). Furthermore, the generative adversarial network (GAN) framework and NSL-KDD dataset were used for evaluation. As a result, ASD discovered the adversarial samples before the samples were input into the NIDS.

Additionally, the accuracy improved by 26.46% in the PGD adversarial environment and by 11.85% in the FGSM adversarial environment. However, the influence of ASD on MI-FGSM is not apparent, necessitating more research. Notwithstanding, by using ASD, normal data were stripped of adversarial samples, and the remaining normal samples were fed into the classification model.

3) APE-GAN++

According to the authors in [98], the defense method that had been presented in [82] had some shortcomings, which were: (1) its training procedure was insecure and suffered from a vanishing gradient issue; (2) it could boost its efficiency even further. Thus, they proposed an improved method named APE-GAN++. In comparison to the APE-GAN, the APE-GAN++ has a generator, a discriminator, and a recently added third-party classifier in its design. As a result, APE-GAN++ achieved a better performance than other defenses, including APE-GAN.

4) Dropout

Most adversarial attack implementations rely on knowledge of the model's architecture. Consequently, dropout is a random process that perturbs the model's architecture [99]. The researchers in [100] used dropout with neural networks. As a result, it makes neural networks resistant to various inputs [101]. Dropout can, thus, be used to detect adversarial samples. Adversarial samples tend to be transcribed as wrong or garbled sentences when inference is performed with dropout turned on. This study focused on the

ability to apply CW attacks in this field as well as the ability to detect them. Additionally, this defense can detect adversarial examples effectively [99].

5) Adversary Detection Network

The research in [49] suggested training a binary detector network to distinguish between samples from the original dataset and adversarial instances. Furthermore, DeepFool adversary-specific detectors perform admirably compared with all other adversarial attacks. As a result, transferability is not perfect for the detectors. It typically works between comparable opponents and from a stronger to a weaker adversary.

6) GAN-Based Defense

The study [44] presented a model of a defensive approach to improve the IDS robustness against the CW attack. This defense is based on GAN; thus, it aims to classify the data that the IDS receives as an attack or normal. Furthermore, they used the CSE-CIC-IDS2018 dataset to evaluate their model. As a result, the IDS's performance and accuracy improved.

Figure 8 illustrates the defense strategies presented in this paper. Moreover, Table 4 illustrates a summary of these strategies.

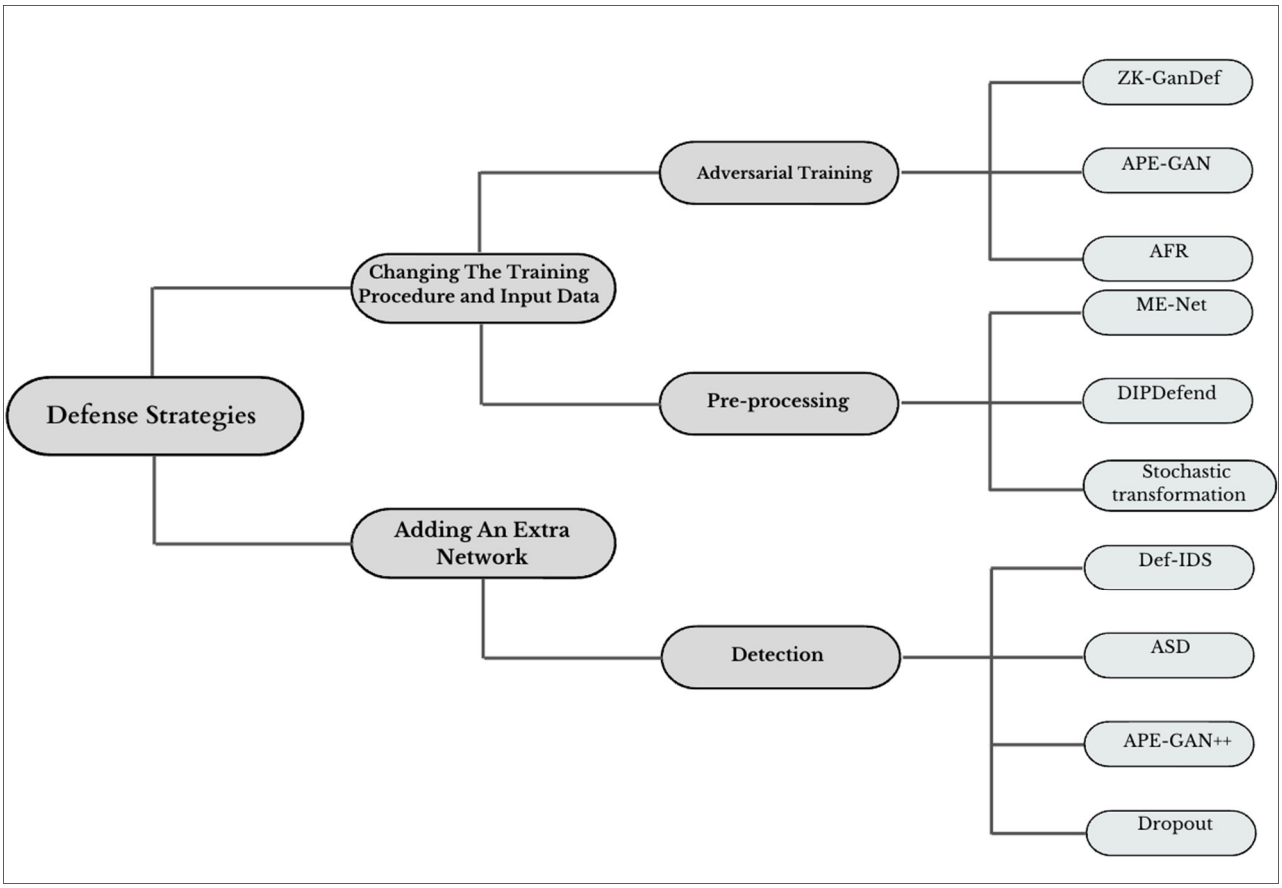

Figure 8. Defense Strategies.

Table 4. Summary Of Defense Methods.

	Year	Method	Description	Result
[49]	2017	Adversary Detection Network	The authors suggested training a binary detector network to distinguish between samples from the original dataset and adversarial instances.	DeepFool adversary-specific detectors performed admirably compared to all other adversarial attacks.

[82]	2017	APE-GAN	This method was based on training the model to remove the adversarial perturbation before feeding the processed sample to classification networks. Then, they generated adversarial samples and used the discriminator to discriminate those samples.	The researchers might infer that the APE-GAN has a wide range of applications because it works despite no understanding of the model on which it is based.
[100]	2017	Dropout	Dropout is a random process that perturbs the model's architecture [99]. Furthermore, this study focused on the ability to apply CW attacks in this field as well as the ability to detect them.	This defense detected adversarial examples effectively [99].
[87]	2019	ME-Net	This method takes incomplete or damaged images and eliminates noise from these images. Furthermore, there are two stages for the image before being processed; first, arbitrary pixels are discarded from the picture, and then the picture is rebuilt using ME.	The ME-Net made deep neural networks more robust against adversarial attacks than other methods.
[88]	2020	Stochastic transformation-based defenses	In the first place, the researchers investigated the impact of random image alterations on clean pictures to understand better how accuracy deteriorates. They trained a unique classifier to identify distinguishing characteristics in the distributions of softmax outputs of converted clean pictures and predict the class label.	Untargeted assaults on CNN have been studied, and it would be interesting to compare their distribution classifier approach with targeted attacks.
[81]	2021	AFR	This method implemented an attack on NIDS and then suggested adversarial feature reduction (AFR), which decreased the attack's efficacy by reducing adversarial feature development.	The implemented attack achieved more than a 97% rate in half cases, and the proposed defense technique (AFR) successfully minimized such attacks.
[88]	2021	DIPDefend	This method examined the internal prior of the image and then divided them into two steps: robust feature learning and non-robust feature learning. It reconstructed the image by beginning with a robust feature and then a non-robust feature to make them stronger against adversarial attacks.	It can be applied without pretraining, making it useful in various situations.

8. Challenges and Future Directions

Generally, the process of creating an adversarial example entails adding the necessary amount of perturbation to the model's direction. Thus, the defense strategies section

has presented many studies that can protect the models against adversaries in the two significant areas of computer vision and IDS by detecting or eliminating the adversaries. Equally important the following questions: Are these solutions effective in addressing our problem? Which of these defenses is the best fit for our problem? Thus, in light of this, this section lists some of the gaps and challenges in this domain.

8.1. Key Research Challenges and Gaps

- In adversarial situations, the competition between attacks and defenses becomes an "arms race": suggested defenses against one assault were later shown to be vulnerable to another, and vice versa [102,103].
- The adversarial examples have transferability properties, indicating that adversarial examples created for one model will most likely work for other models [104,105]. This can be utilized as the basis for various black-box attacks in which a substitute model generates adversarial instances that are then presented to the target model [9].
- The successful attacks in one circumstance could fail in another; for example, the attacks that had success in the computer vision domain may fail or have fewer effects when implemented on IDS [40,106].
- Some defenses demonstrated their ability to repel a particular attack but later fell victim to a minor modification of the attack [100,107].
- Defenses are sometimes tailored to a specific assault strategy and are less suitable as a generic defense [34]. For example, the authors in [80] suggested a method to detect adversarial attacks even though it is not compared to other techniques. Because it is considered a relatively new topic, it is not easy to evaluate this research. However, in our opinion, it is considered helpful research since it covers a new topic with solutions based on experiments and presents the results.
- Each domain has unique features; therefore, it is more challenging to spot disturbances when modifications are performed on the network traffic data [30].
- A critical component of defensive tactics is their ability to withstand all attacks. Nevertheless, most defense techniques are ineffective against black-box attacks [97] or need more experimentation, as in [44]. In addition, some of the strategies are ineffective, such as adversarial training, which has flaws and may be evaded [78].
- The research [9] praised the dropout as a perfect defense technique. However, even with this defense, the adversary can defeat it if they know the dropout rate and try to break it by training with dropouts but with a meager success rate [99].
- The AML term is widespread in image classification, but it is relatively new and shallow in the cybersecurity area, especially in IDSs. Thus, some defense methods ensured their effectiveness in protecting IDS specifically. On the other hand, the rest had successfully applied defense strategies in the computer vision field, such as APE-GAN++.
- In detection strategies, in the worst situation, it is possible to attack the detector that the ML and DL models employ to identify their adversaries [49].
- Some of the defense ideas are repeated, such as using GAN in various research forms, demonstrating its efficacy to the reader. Unfortunately, using GAN is not always the best choice; for example, in [48], the authors mentioned that it might lead the model to misclassification.
- To address white-box attacks, the defender can impede the transferability of adversarial examples. However, a comprehensive defense method could not be used for all ML/DL applications [35].
- Most studies demonstrated how to improve a model's accuracy rather than its resilience and robustness [44].
- The datasets must reflect current traits because network traffic behavior patterns change over time. Unfortunately, the majority of publicly accessible IDS datasets lack modern traffic characteristics. According to the authors in [44], there is a shortcoming in the IDS's datasets; thus, the IDS lacks a dataset that can include all types of network attacks. However, using the GAN-IDS will offer a high volume of attacks in training

since it can generate more attack types. Then, we can use the discriminators to distinguish new attacks [27]. Furthermore, in [108], the authors also presented research on handling this shortage.

8.2. Future Directions

- There is no way to evaluate something without experimentation, but we may draw some conclusions from the experiment's owners. These defense strategies, for example, had been used against white-box attacks, but what about black-box attacks? Thus, there is a need for techniques to counter the black-box attack in the future. In contrast, in [105], the authors presented an approach to address transferable adversarial attacks. We believe it to be a promising defense approach with excellent efficiency against black-box attacks, although it has been examined using a white-box attack.
- In the future, there will be a demand for a solution that handles all types of adversaries that affect the robustness of an IDS.
- Various models may necessitate several defenses [9]. Thus, they need to measure their effectiveness in protecting ML and DL based on IDS.
- Some researchers have stated that their technique may be used in other ML/DL models or is available online for experimentation. Therefore, we suggest increasing the effectiveness of the dropout strategy to make it more reliable and suitable for additional domains such as IDS.
- In this paper, we focus on IDSs; generally, we think that protecting ML/DL-based IDSs is easier to preserve since it is difficult to deceive IDSs because the features contain discrete and non-continuous values [109]. Therefore, we believe that enhancing the GAN defense strategies such as APE-GAN++ will make them more reliable for IDSs, which will be a valuable technique for handling adversaries in the future. Moreover, Table 5 demonstrates a comparison between these strategies.

Table 5. Defense Strategies Using GAN.

Ref.	Year	Defense Approach	Attack Type	Dataset	ML/DL Model	Can Address New Types of Attacks?	Defense Category	Result
[32]	2019	ZK-GanDef	White-box	- MNIST - Fashion-MNIST - CIFAR10	NN	Yes	Changing the training procedure and input data	ZK-GanDef enhanced the accuracy by 49.17% against adversarial attacks compared to other attacks.
[97]	2020	ASD	White-box	- NSL-KDD	DNN	Not apparent	Adding an extra network	ASD improved the accuracy by 26.46% in the PGD adversarial environment and 11.85% in the FGSM adversarial environment, but the influence of ASD on some attacks was not apparent.
[96]	2021	Def-IDS	White-box	- CSE-CIC-IDS2018	DNN	Yes	Adding an extra network	The experiments showed that the Def-IDS could increase the robustness of NIDS by enhancing the accuracy of detecting adversarial attacks.

[98]	2021	APE-GAN++	White-box	- T - R10	MNIS CIFA	CNN	Yes	Adding an extra network	APE-GAN++ achieved an outstanding performance than other defenses, including the APE-GAN.
[44]	2022	GAN-based defense	White-box	- CIC-IDS2018	CSE-	- - -	DT RF SVM	Not apparent	Adding an extra network	The IDS performance improved, and its accuracy increased.

Despite the threats that face the ML and DL when using them as an engine for IDS, it is a powerful technique that has served cybersecurity in general and IDSs in particular. Therefore, this paper highlighted various attacks and defense techniques to improve the precision of ML-based IDSs and the IDSs' robustness.

9. Conclusions

In cybersecurity, using ML algorithms takes much attention, especially in intrusion detection systems (IDS). Therefore, significant research has been conducted to improve the speed, accuracy, precision, and other essential metrics of ML-based IDS. Moreover, adversarial attacks can have a significant impact on ML algorithms. Hence, the ML-based IDS is vulnerable to adversarial attacks that spark security concerns.

For example, the IDS classification accuracy is affected by identifying a "malicious" input as "benign" or vice versa. In this situation, the IDS will be unreliable in defense, posing a severe threat to our networks. Thus, this paper presented a general overview of the ML methods in IDSs to improve their performance. Furthermore, it clarifies the various types of adversarial attacks that can affect the IDS based on ML to evaluate its robustness. In addition, we mentioned the benchmark datasets for IDSs and some state-of-the-art defense strategies that improved IDS accuracy. Finally, we discussed the open issues facing implementing defensive methods to improve ML-based IDSs.

Author Contributions: Conceptualization, M.A.R. and A.A.; methodology, A.A.; software, A.A.; validation, M.A.R. and A.A.; formal analysis, A.A.; investigation, A.A.; resources, A.A.; data curation, A.A.; writing—original draft preparation, A.A.; writing—review and editing, M.A.R.; visualization, A.A.; supervision, M.A.R.; project administration, M.A.R.; funding acquisition, M.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not Applicable.

Acknowledgments: The author(s) gratefully acknowledge Qassim University, represented by the Deanship of "Scientific Research, on the financial support for this research under the number (COC-2022-1-1-J- 24954) during the academic year 1444 AH/2022 AD".

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

References

1. Ford, V.; Siraj, A. Applications of machine learning in cyber security. In Proceedings of the 27th International Conference on Computer Applications in Industry and Engineering, New Orleans, LA, USA, 13–15 October 2014; Volume 118.
2. Denning, D.E. An intrusion-detection model. *IEEE Trans. Softw. Eng.* **1987**, *2*, 222–232.
3. Liao, H.-J.; Lin, C.-H.R.; Lin, Y.-C.; Tung, K.-Y. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **2013**, *36*, 16–24.
4. Aldweesh, A.; Derhab, A.; Emam, A.Z. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Syst.* **2020**, *189*, 105124.
5. Liu, H.; Lang, B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Appl. Sci.* **2019**, *9*, 4396. https://doi.org/10.3390/app9204396.
6. Pervez, M.S.; Farid, D.M. Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs. In Proceedings of the The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, Bangladesh, 18–20 December 2014; pp. 1–6.

7. Gu, X.; Easwaran, A. Towards safe machine learning for cps: Infer uncertainty from training data. In Proceedings of the Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems, Montreal, QC, Canada, 16–18 April 2019; pp. 249–258.
8. Ghafouri, A.; Vorobeychik, Y.; Koutsoukos, X. Adversarial regression for detecting attacks in cyber-physical systems. *arXiv* **2018**, arXiv:1804.11022.
9. McCarthy, A.; Ghadafi, E.; Andriotis, P.; Legg, P. Functionality-Preserving Adversarial Machine Learning for Robust Classification in Cybersecurity and Intrusion Detection Domains: A Survey. *J. Cybersecurity Priv.* **2022**, *2*, 154–190. https://doi.org/10.3390/jcp2010010.
10. Yang, K.; Liu, J.; Zhang, C.; Fang, Y. Adversarial examples against the deep learning based network intrusion detection systems. In Proceedings of the MILCOM 2018–2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, USA, 29–31 October 2018; pp. 559–564.
11. Alhajjar, E.; Maxwell, P.; Bastian, N. Adversarial machine learning in Network Intrusion Detection Systems. *Expert Syst. Appl.* **2021**, *186*, 115782. https://doi.org/10.1016/j.eswa.2021.115782.
12. Dalvi, N.; Domingos, P.; Mausam; Sanghai, S.; Verma, D. Adversarial classification. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Miningk, New York, NY, USA, 22 August 2004; pp. 99–108. https://doi.org/10.1145/1014052.1014066.
13. Matsumoto, T.; Matsumoto, H.; Yamada, K.; Hoshino, S. Impact of artificial" gummy" fingers on fingerprint systems. In Optical *Security and Counterfeit Deterrence Techniques IV*; International Society for Optics and Photonics: San Jose, CA, USA, 2002; Volume 4677, pp. 275–289.
14. Ayub, M.A.; Johnson, W.A.; Talbert, D.A.; Siraj, A. Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning. In Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 18–20 March 2020. https://doi.org/10.1109/CISS48834.2020.1570617116.
15. Suo, H.; Wan, J.; Zou, C.; Liu, J. Security in the internet of things: A review. In Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering, Hangzhou, China, 23–25 March 2012; Volume 3, pp. 648–651.
16. Wang, X.; Li, J.; Kuang, X.; Tan, Y. an; Li, J. The security of machine learning in an adversarial setting: A survey. *J. Parallel Distrib. Comput.* **2019**, *130*, 12–23. https://doi.org/10.1016/j.jpdc.2019.03.003.
17. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial Attacks and Defences: A Survey. *arXiv* **2018**, arXiv:1810.00069. Available online: http://arxiv.org/abs/1810.00069 (accessed on 18/October/2022).
18. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430.
19. Zhou, Y.; Kantarcioglu, M.; Xi, B. A survey of game theoretic approach for adversarial machine learning. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1259. https://doi.org/10.1002/widm.1259.
20. Dasgupta, B, and Collins, J. A survey of game theory methods for adversarial machine learning in cybersecurity tasks. *Amnesty Int. J.* 40, 31–43. https://doi.org/10.1609/aimag.v40i2.2847.
21. Duddu, V. A survey of adversarial machine learning in cyber warfare. *Def. Sci. J.* **2018**, *68*, 356.
22. Ibitoye, O.; Abou-Khamis, R.; Matrawy, A.; Shafiq, M.O. The Threat of Adversarial Attacks on Machine Learning in Network Security—A Survey. *arXiv* **2019**, arXiv:1911.02621. Available online: http://arxiv.org/abs/1911.02621 (accessed on 22 December 2022).
23. Qayyum, A.; Qadir, J.; Bilal, M.; Al-Fuqaha, A. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Rev. Biomed. Eng.* **2021**, *14*, 156–180. https://doi.org/10.1109/RBME.2020.3013489.
24. Homoliak, I.; Teknos, M.; Ochoa, M.; Breitenbacher, D.; Hosseini, S.; Hanacek, P. Improving network intrusion detection classifiers by non-payload-based exploit-independent obfuscations: An adversarial approach. *arXiv* **2018**, arXiv:1805.02684.
25. Khamis, R.A.; Shafiq, M.O.; Matrawy, A. Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Dublin, Ireland , 7–11 June 2020. https://doi.org/10.1109/ICC40277.2020.9149117.
26. Yuan, X. Phd forum: Deep learning-based real-time malware detection with multi-stage analysis. In Proceedings of the 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 29–31 May 2017, 2017; pp. 1–2.
27. Shahriar, M.H.; Haque, N.I.; Rahman, M.A.; Alonso, M. G-IDS: Generative Adversarial Networks Assisted Intrusion Detection System. In Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 13–17 July 2020; pp. 376–385. https://doi.org/10.1109/COMPSAC48688.2020.0-218.
28. Huang, L.; Joseph, A.D.; Nelson, B.; Rubinstein, B.I.P.; Tygar, J.D. Adversarial machine learning. In Proceedings of the 4th ACM Workshop on SECURITY and Artificial Intelligence, Chicago, IL, USA, 21 October 2011; pp. 43–58.
29. Shetty, S.; Ray, I.; Ceilk, N.; Mesham, M.; Bastian, N.; Zhu, Q. Simulation for Cyber Risk Management–Where are we, and Where do we Want to Go? In Proceedings of the 2019 Winter Simulation Conference (WSC), National Harbor, MD, USA, 8–11 December 2019; pp. 726–737.
30. Apruzzese, G.; Andreolini, M.; Ferretti, L.; Marchetti, M.; Colajanni, M. Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems. *Digit. Threat. Res. Pract.* **2021**, *3*, 1–19. https://doi.org/10.1145/3469659.

31. Sarker, I.H.; Abushark, Y.B.; Alsolami, F.; Khan, A.I. Intrudtree: A machine learning based cyber security intrusion detection model. *Symmetry* **2020**, *12*, 754.
32. Khalil, K.; Qian, Z.; Yu, P.; Krishnamurthy, S.; Swami, A. Optimal monitor placement for detection of persistent threats. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washinton, DC, USA, 4–8 December 2016; pp. 1–6.
33. Liu, G.; Khalil, I.; Khreishah, A. ZK-GanDef: A GAN Based Zero Knowledge Adversarial Training Defense for Neural Networks. In Proceedings of the 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Portland, OR, USA, 24–27 June 2019; pp. 64–75. https://doi.org/10.1109/DSN.2019.00021.
34. Zhang, J.; Li, C. Adversarial examples: Opportunities and challenges. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 2578–2593.
35. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. neural networks Learn. Syst.* **2019**, *30*, 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017.
36. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing machine learning models via prediction apis. In Proceedings of the 25th {USENIX} Security Symposium ({USENIX} Security 16); 2016; pp. 601–618.
37. Xi, B. Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges. *Wiley Interdiscip. Rev. Comput. Stat.* **2020**, *12*. https://doi.org/10.1002/wics.1511.
38. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, 12–18 October 2015; pp. 1322–1333.
39. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 3–18.
40. Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M.K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 Acm Sigsac Conference on Computer and Communications Security, New York, NY, USA, 24–28 October 2016; pp. 1528–1540.
41. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
42. Zizzo, G.; Hankin, C.; Maffeis, S.; Jones, K. INVITED: Adversarial machine learning beyond the image domain. In Proceedings of the 56th Annual Design Automation Conference 2019, New York, NY, USA, 2 June 2019. https://doi.org/10.1145/3316781.3323470.
43. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; pp. 372–387.
44. Pujari, M.; Cherukuri, B.P.; Javaid, A.Y.; Sun, W. An Approach to Improve the Robustness of Machine Learning based Intrusion Detection System Models Against the Carlini-Wagner Attack. In Proceedings of the 2022 IEEE International Conference on Cyber Security and Resilience (CSR), Rhodes, Greece, 27–29 July 2022; pp. 62–67. https://doi.org/10.1109/csr54599.2022.9850306.
45. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; pp. 2574–2582.
46. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: London, UK, 2018; pp. 99–112 ISBN 1351251384.
47. Wang, Z. Deep Learning-Based Intrusion Detection with Adversaries. *IEEE Access* **2018**, *6*, 38367–38384. https://doi.org/10.1109/ACCESS.2018.2854599.
48. Martins, N.; Cruz, J.M.; Cruz, T.; Henriques Abreu, P. Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. *IEEE Access* **2020**, *8*, 35403–35419. https://doi.org/10.1109/ACCESS.2020.2974752.
49. Metzen, J.H.; Genewein, T.; Fischer, V.; Bischoff, B. On detecting adversarial perturbations. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017; pp. 1–12.
50. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, New York, NY, USA, 2–6 April 2017; pp. 506–519.
51. Guo, S.; Zhao, J.; Li, X.; Duan, J.; Mu, D.; Jing, X. A Black-Box Attack Method against Machine-Learning-Based Anomaly Network Flow Detection Models. *Secur. Commun. Netw.* **2021**, *2021*, 5578335. https://doi.org/10.1155/2021/5578335.
52. Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, New York, NY, USA, 3 November 2017; pp. 15–26.
53. Su, J.; Vargas, D.V.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841.
54. Biggio, B.; Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.* **2018**, *84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023.
55. Laskov, P. Practical evasion of a learning-based classifier: A case study. In Proceedings of the 2014 IEEE Symposium on Security and Privacy, San Jose, CA, USA 18–21 May 2014; pp. 197–211.
56. Corona, I.; Giacinto, G.; Roli, F. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. *Inf. Sci.* **2013**, *239*, 201–225.

57. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy—ICISSP, Madeira, Portugal, 22–24 January 2018, pp. 108–116. https://doi.org/10.5220/0006639801080116.
58. Viegas, E.K.; Santin, A.O.; Oliveira, L.S. Toward a reliable anomaly-based intrusion detection in real-world environments. *Comput. Netw.* **2017**, *127*, 200–216. https://doi.org/10.1016/j.comnet.2017.08.013.
59. Zhang, X.; Zheng, X.; Wu, D.D. Attacking Attacking DNN-based DNN-based Intrusion Intrusion Detection Detection Models Models Attacking Intrusion Detection Models Models Attacking Intrusion Detection Attacking DNN-based Intrusion Detection Models. *IFAC Pap.* **2020**, *53*, 415–419. https://doi.org/10.1016/j.ifacol.2021.04.118.
60. Anthi, E.; Williams, L.; Rhode, M.; Burnap, P.; Wedgbury, A. Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. *J. Inf. Secur. Appl.* **2021**, *58*, 102717. https://doi.org/10.1016/j.jisa.2020.102717.
61. Zhao, S.; Li, J.; Wang, J.; Zhang, Z.; Zhu, L.; Zhang, Y. AttackGAN: Adversarial Attack against Black-box IDS using Generative Adversarial Networks. *Procedia Comput. Sci.* **2021**, *187*, 128–133. https://doi.org/10.1016/j.procs.2021.04.118.
62. M. Tavallaee, E. Bagheri, W. Lu, e A. A. Ghorbani, A detailed analysis of the KDD CUP 99 data set. In In Proceedings of the 2009, Ottawa, ON, Canada, 8–10 July 2009. doi: 10.1109/CISDA.2009.5356528..
63. Piplai, A.; Sree, S.; Chukkapalli, L.; Joshi, A. NAttack ! Adversarial Attacks to Bypass a GAN Based Classifier Trained to Detect Network Intrusion. Available online: https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00020 (accessed on 21 December 2022).
64. Usama, M.; Asim, M.; Latif, S.; Qadir, J.; Ala-Al-Fuqaha Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 78–83. https://doi.org/10.1109/IWCMC.2019.8766353.
65. Lin, Z.; Shi, Y.; Xue, Z. IDSGAN: Generative Adversarial Networks for Attack Generation Against Intrusion Detection. In *Lecture Notes in Computer Science. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*; Springer: Cham, Germany, 2022; 13282 LNAI, pp. 79–91. https://doi.org/10.1007/978-3-031-05981-0_7.
66. Duy, P.T.; Tien, L.K.; Khoa, N.H.; Hien, D.T.T.; Nguyen, A.G.T.; Pham, V.H. DIGFuPAS: Deceive IDS with GAN and function-preserving on adversarial samples in SDN-enabled networks. *Comput. Secur.* **2021**, *109*, 102367. https://doi.org/10.1016/j.cose.2021.102367.
67. Chen, J.; Wu, D.; Zhao, Y.; Sharma, N.; Blumenstein, M.; Yu, S. Fooling intrusion detection systems using adversarially autoencoder. *Digit. Commun. Netw.* **2021**, *7*, 453–460. https://doi.org/10.1016/j.dcan.2020.11.001.
68. Chauhan, R.; Shah Heydari, S. Polymorphic Adversarial DDoS attack on IDS using GAN. In Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 20–22 October 2020; pp. 1–6. https://doi.org/10.1109/ISNCC49221.2020.9297264.
69. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6. https://doi.org/10.1109/CISDA.2009.5356528.
70. Janusz, A.; Kałuza, D.; Chądzyńska-Krasowska, A.; Konarski, B.; Holland, J.; Ślęzak, D. IEEE BigData 2019 cup: Suspicious network event recognition. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 5881–5887.
71. Gonzalez-Cuautle, D.; Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, L.K.; Portillo-Portillo, J.; Olivares-Mercado, J.; Perez-Meana, H.M.; Sandoval-Orozco, A.L. Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets. *Appl. Sci.* **2020**, *10*, 794.
72. Jatti, S.A. V.; Kishor Sontif, V.J.K. Intrusion detection systems. *Int. J. Recent Technol. Eng.* **2019**, *8*, 3976–3983. https://doi.org/10.35940/ijrte.B1540.0982S1119.
73. Yilmaz, I.; Masum, R.; Siraj, A. Addressing Imbalanced Data Problem with Generative Adversarial Network for Intrusion Detection. In Proceedings of the 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 11–13 August 2020; pp. 25–30. https://doi.org/10.1109/IRI49571.2020.00012.
74. Huang, S.; Lei, K. IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks. *Ad Hoc Netw.* **2020**, *105*, 102177. https://doi.org/10.1016/j.adhoc.2020.102177.
75. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10–12 November 2015. https://doi.org/10.1109/MilCIS.2015.7348942.
76. Panigrahi, R.; Borah, S. A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems. *Int. J. Eng. Technol.* **2018**, *7*, 479–482.
77. Li, J. hua Cyber security meets artificial intelligence: A survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 1462–1474. https://doi.org/10.1631/FITEE.1800573.
78. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv* **2017**, arXiv:1705.07204.

79. Song, D.; Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Tramer, F.; Prakash, A.; Kohno, T. Physical adversarial examples for object detectors. In Proceedings of the 12th USENIX Workshop on Offensive Technologies (WOOT 18), Baltimore, MD, USA, 13–14 August 2018.
80. Pawlicki, M.; Choraś, M.; Kozik, R. Defending network intrusion detection systems against adversarial evasion attacks. *Futur. Gener. Comput. Syst.* **2020**, *110*, 148–154. https://doi.org/10.1016/j.future.2020.04.013.
81. Han, D.; Wang, Z.; Zhong, Y.; Chen, W.; Yang, J.; Lu, S.; Shi, X.; Yin, X. Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2632–2647. https://doi.org/10.1109/JSAC.2021.3087242.
82. Jin, G.; Shen, S.; Zhang, D.; Dai, F.; Zhang, Y. APE-GAN: Adversarial Perturbation Elimination with GAN. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3842–3846. https://doi.org/10.1109/ICASSP.2019.8683044.
83. Zantedeschi, V., Nicolae, M. I., & Rawat, A. Efficient defenses against adversarial attacks. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017.Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial examples for semantic segmentation and object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–19 October 2017; pp. 1369–1378.
84. Xu, W.; Evans, D.; Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv* **2017**, arXiv:1704.01155.
85. Guo, C.; Rana, M.; Cisse, M.; Van Der Maaten, L. Countering adversarial images using input transformations.(2018). *arXiv* **2018**, arXiv:1711.00117.
86. Samangouei, P.; Kabkab, M.; Chellappa, R. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv* **2018**, arXiv:1805.06605.
87. Yang, Y.; Zhang, G.; Katabi, D.; Xu, Z. ME-Net: Towards effective adversarial robustness with matrix estimation. In Proceedings of the 36th International Conference on Machine Learning, June, 10–15 June 2019; pp. 12152–12173.
88. Dai, T.; Feng, Y.; Chen, B.; Lu, J.; Xia, S.T. Deep image prior based defense against adversarial examples. *Pattern Recognit.* **2022**, *122*, 108249. https://doi.org/10.1016/j.patcog.2021.108249.
89. *Enhancing Transfomation Based Defenses Against Adversarial Attacks With A Distribution Classifier*. https://openreview.net/pdf?id=BkgWahEFvr (accessed on 21 December 2022).
90. Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; Storer, J. Deflecting adversarial attacks with pixel deflection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8571–8580.
91. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. Mitigating adversarial effects through randomization. *arXiv* **2017**, arXiv:1711.01991.
92. Akhtar, N.; Liu, J.; Mian, A. Defense against universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3389–3398.
93. Metzen, J.H.; Genewein, T.; Fischer, V.; Bischoff, B. On detecting adversarial perturbations. *arXiv* **2017**, arXiv:1702.04267.
94. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
95. Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, *arXiv* **2018**, arXiv:1710.10766.
96. Wang, J.; Pan, J.; Alqerm, I.; Liu, Y. Def-IDS: An Ensemble Defense Mechanism against Adversarial Attacks for Deep Learning-based Network Intrusion Detection. In Proceedings of the 2021 International Conference on Computer Communications and Networks (ICCCN), Athens, Greece, 19–22 July 2021; pp. 1–9. https://doi.org/10.1109/ICCCN52240.2021.9522215.
97. Peng, Y.; Fu, G.; Luo, Y.; Hu, J.; Li, B.; Yan, Q. Detecting Adversarial Examples for Network Intrusion Detection System with GAN. In Proceedings of the 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS); IEEE, 2020; pp. 6–10.
98. Yang, R.; Chen, X.Q.; Cao, T.J. APE-GAN++: An Improved APE-GAN to Eliminate Adversarial Perturbations. *IAENG Int. J. Comput. Sci.* **2021**, *48*, 1–18.
99. Jayashankar, T.; Le Roux, J.; Moulin, P. Detecting audio attacks on ASR systems with dropout uncertainty. 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 4671–4675. https://doi.org/10.21437/Interspeech.2020-1846. 15 Sep 2020.
100. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting adversarial samples from artifacts. *arXiv* **2017**, arXiv:1703.00410.
101. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
102. Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv* **2017**, arXiv:1702.06280.
103. Carlini, N.; Katz, G.; Barrett, C.; Dill, D.L. Provably minimally-distorted adversarial examples. *arXiv* **2017**, arXiv:1709.10207.
104. Papernot, N.; McDaniel, P.; Goodfellow, I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv* **2016**, arXiv:1605.07277.

105. Chhabra, A.; Mohapatra, P. Moving Target Defense against Adversarial Machine Learning. In Proceedings of the MTD 2021—Proceedings of the 8th ACM Workshop on Moving Target Defense, co-located with CCS 2021, New York, NY, USA,,15 November 2021; pp. 29–30. https://doi.org/10.1145/3474370.3485662.
106. Hashemi, M.J.; Cusack, G.; Keller, E. Towards evaluation of nidss in adversarial setting. In Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, New York, NY, USA, 9 December 2019; pp. 14–21.
107. Bhagoji, A.N.; Cullina, D.; Sitawarin, C.; Mittal, P. Enhancing robustness of machine learning systems via data transformations. In Proceedings of the 2018 52nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA 21–23 March 2018; pp. 1–5.
108. Thakkar, A.; Lohiya, R. A review of the advancement in intrusion detection datasets. *Procedia Comput. Sci.* **2020**, *167*, 636–645.
109. Labaca-Castro, R.; Biggio, B.; Dreo Rodosek, G. Poster: Attacking malware classifiers by crafting gradient-attacks that preserve functionality. In Proceedings of the Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, 11–15 November 2019; pp. 2565–2567.