Adversarial Machine Learning: A Survey on the Influence Axis

Shahad Alzahrani[1], Taghreed Almalki[1], Hatim Alsuwat[2], and Emad Alsuwat [1]

[1] Department of Computer Science, College of Computers and Information Technology, Taif University, Saudi Arabia
[2] Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia

Summary
After the everyday use of systems and applications of artificial intelligence in our world. Consequently, machine learning technologies have become characterized by exceptional capabilities and unique and distinguished performance in many areas. However, these applications and systems are vulnerable to adversaries who can be a reason to confer the wrong classification by introducing distorted samples. Precisely, it has been perceived that adversarial examples designed throughout the training and test phases can include industrious Ruin the performance of the machine learning. This paper provides a comprehensive review of the recent research on adversarial machine learning. It's also worth noting that the paper only examines recent techniques that were released between 2018 and 2021. The diverse systems models have been investigated and discussed regarding the type of attacks, and some possible security suggestions for these attacks to highlight the risks of adversarial machine learning.

Keywords:
Machine Learning; Adversarial Machine Learning; Influence attack; Evasion attack; Data poisoning attack.

1. Introduction

In many aspects of our daily life, AI has been omnipresent in the past few decades [1]. In particular, ML has become a highly-topic topic; it attracts excellent attention both from academia and industry [2], including medical diagnostics, trade, finance, and vehicles completely autonomous [3]. The central concept of machine learning is that vast quantities of data are used to train a model to generalize samples well for unseen tests [2]. Even without the specific programming, without human intervention, creates forecasts automatically [4].

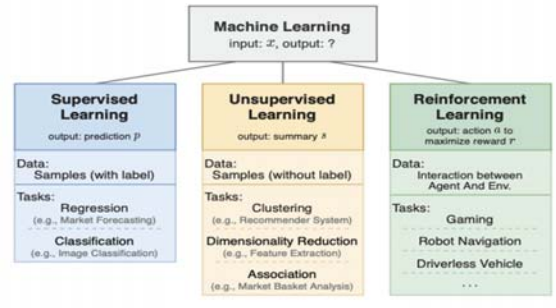

Fig. 1 Taxonomy of Machine Learning.[5]

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.28

When we define machine learning systems depending on entered data and outturn samples, we have three simple learning machine models, as seen in fig. 1 [5].

i. Learning Supervised [6]:

It is one of the two main machine learning branches, making it easy to better predict outcomes after training based on previous data, using input/output pairs, or labeling data, training the model for function generation, which, when applied, is sufficiently estimated to predict new inputs.

ii. Unsupervised Learning [7]:

Unlike supervised learning, unsupervised learning merely includes inputs but no relevant labels. Therefore, unattended education aims to understand how the data are distributed and how the data points differ. The clustering trouble of discovering groups of data, such as classifying people by their behavior, is a typical example of unsupervised learning.

iii. Strengthening Learning:

It is entirely separate from monitored and unmonitored learning. Labeling input/output pairs and explicit correction of suboptimal options are unnecessary to train a reinforcement agent [5]. An algorithm is instead used to learn and test an independent agent which attempts to solve a problem through automated learning [9].

Deep learning has emerged as a solid and efficient framework that can be applied to a broad spectrum of complex learning problems which were difficult to address utilizing the conventional AI strategies before. Over the most recent couple of years, deep learning has advanced radically to outperform human-level execution on various assignments. Consequently, deep understanding is by and large broadly utilized in the majority of the new everyday applications. Nonetheless, deep learning frameworks are powerless against created antagonistic models, which might be intangible to the natural eye, but can lead the model to misclassify the output. Lately, various sorts of adversaries dependent on their danger model influence these vulnerabilities to compromise a deep learning framework where adversaries have high incentives. Consequently, it is

exceptionally critical to give vigor to deep learning calculations against these adversaries [62].

Applications for using RL can teach neural networks to play games such as Go [10], robots for specific tasks [11], or intelligent transport systems [12]. RL is commonly implemented to satisfy the Markov property as a Markov Decision Process (MDP): the following condition depends on Markov property RL typically implemented in the Markov Decision Process (MDP). The following state can only rely on the present state and the agent's act, not the previous states [13, 8].

Algorithmic decision-making offers well-defined benefits; unlike humans, computers are not exhausted or bored [14, 15], and more variables can be considered on orders of magnitude. Algorithms are, however, as humans, subject to prejudices that "unfairly" make their choices [16, 17, 18]. The problem for a machine learning algorithm is that it can be vulnerable to an opponent who attempts to insert malicious information into the learning algorithm to making the algorithm fail to detect the attack. An intruder may use various tactics to initiate targeted assaults to bypass the detection system [19]. researchers have attempted to develop rational devices that can generalize and determine on their own since the advent of Artificial Intelligence. In this process, they initially trust and expand upon the data they get or the environment they are in [20]. But there is a problem with reliability; what about not counting the information? What if a competitor tries to change our choice or reveal our algorithm? Are the confidences safe? These fundamental questions are the basis for the concept of "Adversarial Machine Learning" in the presence of an enemy [21].

The security of any machine learning model is measured concerning the ill-disposed objectives and abilities. In this segment, we taxonomize the threat models in machine learning frameworks remembering the adversary's strength. We start with identifying the threat surface of frameworks based on machine learning models to distinguish where and how an adversary may attempt to sabotage the framework enduring an onslaught [63].

The tech industry is now undergoing Adversarial Machine Learning; Google [22], Microsoft [23], and IBM [25] have, for instance, indicated efforts for stable ML systems apart from their dedication to secure their conventional software systems. The first study on adversarial machine learning, Gartner's leading business analyst, was released in Feb 2019 [25], indicating: "System leaders need to predict and plan for future data corruption risks, product theft or adversarial samples [26].

The time for hostile attacks to occur is classified into two parts, as they often appear in two phases, the first in the training phase and the second in the testing phase [27]:

i. Adversarial Attack During Training Phase:

Certain opponents try to undermine the ML paradigm by targeting the initial information in the machine learning training process. A poisoning attack is a common form of attack attempting to alter the training data set's mathematical properties. A poisoning attack is seen as a cause of the ML model breaking its integrity and availability. The initial ML device dataset is primarily private and cannot be edited by attackers quickly. However, the ML model is typically expected to be retrained in certain apps (for example, biometric face recognition, malware detection, and spam email filtering). This may allow the training data set to degenerate while changing the setting. This allows the intruder the potential to exploit the ML training results.

ii. Adversarial Attacks During the Prediction/Test Period:

A method that yields a conclusion based on the qualified model is the predicted/test phase. Please notice that the data used to predict are not the same as in the testing dataset.

The adversarial model of machine learning includes the following elements in the table. 1 [30].

Table 1: Description of the adversarial ML model elements.[30]

Elements	***Discerption***
Adversarial Goal	The adversarial objectives can be described in two ways: the degree of damage incurred by the opponent and the precise nature of the attack. In the former, the opponent impacts ML model secrecy, integrity, availability, and privacy [28.29]. The above creates racist and non-discriminatory threats. For example, in a discriminate and reputation attack, the opponent will increase the likelihood of misclassification for an ML device and gain private guest information through an aimed attack.
Adversarial knowledge	The adversarial knowledge consists of imperfect information and perfect information dependent on the basic restrictions of the ML model. The fundamental ML limitations are training sets, model factors, knowledge on input.
Adversarial Capability	The word capability denotes the conduct of the attacks existing. This activity is based on the potential danger surface attack vectors. The ranges specify the level of protection the opponent will enter in the ML system. The adversary will analyze the actions of the model in the prediction/test process to extract device vulnerabilities (also referred to as a black box) or capture information about the inner model (also referred to as a white box). During the training level, the opponent can make the

Elements	*Discerption*
	training samples corrupt by using his/her read/write access.
Attack Strategy	The attack plan applies to how the opponents will change the training and test data collection to improve their attacks.

This paper contributes to a comprehensive survey of the recent evasion attacks and data poisoning attacks made by multiple researchers in this domain. Thus, the report focuses only on the recent papers that have been published during 2018-2021. This paper is divided into four main sections. Section 2 introduces some machine learning threat models. Section 3 presents recently published documents, which we have carefully selected that focused only on the Influence Axis in adversarial Machine Learning. Section 4 discusses. It discusses the different fields in which machine learning has been applied and the impact of hostile attacks on them, specifically (evasion attacks and data poisoning attacks).Finally, section 5 concludes the article.

2. Taxonomy of Adversarial Machine Learning:

The machine learning threat model can be categorized in three ways, as shown in fig. 2 [31].

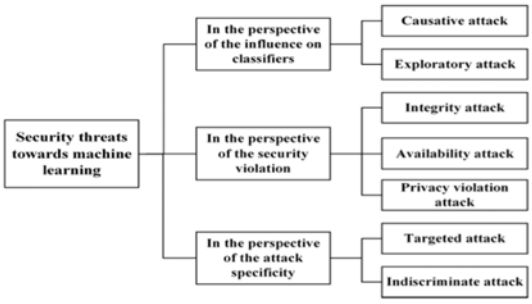

fig. 2 Taxonomy of ML Threat Model [31]

i. Security Violation [31]:

Machine learning risks may be divided into three categories:

1) Integrity attack. By classifying harmful samples, it attempts to maximize the false negatives of known classifiers.

2) Availability attack. Such an attack would lead to a rise in the fake positives of the samples of benign classifiers.

3) Privacy violation attack. This means that opponents can access classified and sensitive data from learning models.

ii. The Attacks Specificity [32]:

Determines whether the assaults alter or affect an entire model based on many attack vectors or whether a particular attack vector is used to target the model. Specific assaults can be categorized as:

1) Targeted: The emphasis is on a particular or limited the number of targets in a coordinated attack.

2) Indiscriminate: a more versatile purpose is for an indiscriminate enemy, such as misclassifying.

iii. The Influence of the Attacker [33]:

This Axis determines how the assailant controls the deep learning models process. In Xiao's view, [34] an assailant can carry out two styles of attack, keeping its effect on the classification model:

1) Causative or Data Poisoning Attacks: mainly during the training stage of the attacker, the deep learning paradigm is affected during causative attacks. For this form of attack, training samples or the training set are contaminated with opposite examples. A classification model is generated that is inconsistent with the original data distribution. The purpose of data poisoning attacks, inserting malicious samples in training data to affect the model's accuracy [35]. Fig.3 shows an explanation of the data poisoning attack process [35].

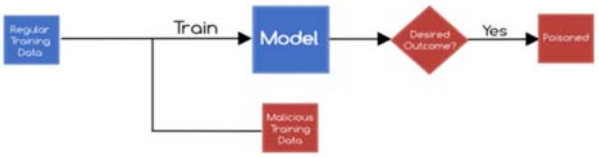

Fig. 3 Data Poisoning Attack Process [35]

For example, of an attack of data poisoning on a network irregularity detection device (IDS), an attack payload can be avoided so that the data can be decoded at the destination. Still, the IDS does not lead to potential errors. The attacker may therefore abuse the target device identified by the IDS. Another aim of the intruder may be to force the machine to retrain the principle and thus considerably degrade its performance [36].

2) Evasion or Exploratory Attacks: in comparison to causative attacks, at the inference or checking stage, an attacker affects the deep learning models. The most common type of attack is evasive attacks, where the attackers build antagonists' instances, usually with a strong faith in prediction, which contributes to misclassifying the machine learning models. Mysterious attacks can also be exploratory in design to collect information on the target model, like its specifications, architectures, price functions, etc. The I/O attack is the most common evasion attack. In which the enemy delivers adversarial images to the intended model, then the adversary analyses the model's performance and attempts to replicate the replacement model to resemble the desired model. This type of attack focuses on crafting input samples that implemented a particular job and bypassing the detection (Forcing the Model to mark it

benign. That is, misclassifying it) [37]. Fig. 4 shows an explanation of the evasion attack process [35].

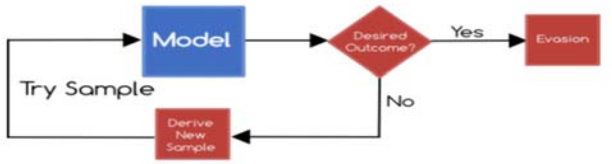

Fig. 4 Evasion Attack Process [35]

Adversarial Machine Learning in Image Classification is an active research path that is liable for the more significant part of the work in the area, with novel papers created practically day by day. Nonetheless, there is neither a known efficient answer for getting Deep Leaning models nor any completely acknowledged clarifications for the presence of negative images yet. Considering the dynamism and importance of this research area, it is vital to be accessible in literature comprehensive furthermore, state-of-the-art review papers to position and orientate their readers about the actual situation. Even though there are now some extensive studies, they have effectively gotten outdated because of the unusual activity in the area. Besides, they draw out an overall overview of the Adversarial Machine Learning field, which, thus, adds to these papers neither have zeroed in enough in works that have proposed defenses against adversarial attacks nor have given legitimate direction to the individuals who wish to put resources into novel countermeasures [64].

Example evasion attack, the enemy can construct an anomalous network-layer protocol behavior dataset and use a labeled attack-data set to train an abnormal intrusion detection device as the base truth of the attack. As a result, Network layer protocol cyber-attacks, which threaten the protection of the underlining system, cannot be detected by the detector. This attack can also significantly affect the consistency of a signature-backed intrusion detection system that detects malware that infects a system or infrastructure [36].

3. Related work

Adversarial machine learning is an environment in which a class of attacks is researched, which hinders classification performance on particular tasks. Adversary attacks usually can be categorized chiefly, such as poisoning attacks where the perpetrator impacts training data or their tags. The model is inadequately performed in implementation or evasion attacks. During deployment, the assailant manipulates data to trick previously qualified categorization devices [38, 39].

This section focuses on influence attacks because the lack of published papers on this type of attack spread significantly in machine learning. Due to the capabilities the attacker possesses in manipulating the intended target. Therefore, numerous studies have been conducted by several researchers to discover how the influence attacks happen by using various approaches detailed below:

I. Evasion Attack (Exploratory):

1) on Power Systems State Estimation:

Sayghe et al. [40] analyzed the influence of hostile machine learning systems and algorithms used in state calculations to discover FDIC. Mainly, they show the effect against poisoning and evasion adverse events on Support Vector Machinery (SVM) and Multilayer Perceptrons (MLP). The download data collected from the independent New York system operator (NYISO) were used to test IEEE 14 bus systems algorithms. Two separate SVM and MLP assault examples have been explored: the adversarial mark flipped and the TFGSM. These attacks have demonstrated that the recognition accuracy of machine learning algorithms is dramatically reduced.

2) on Convolution Neural Networks:

Qian et al. [41] suggested a faulty attack on CNN classifiers that adds preset disturbance to unique license plate regions, simulating some kind of naturally shaped locations (such as sludge, etc.). The problem is thus modeled as a perturbation search procedure. They use the proposed algorithm to produce numerous opponent examples as rectangles, triangles, elliptical ellipses, and spots. Experimental findings indicate that the human eyes ignore these adverse examples, but HyperLPR is more than 93 percent effective in attacks. Therefore, they sense that this type of spot-evasion attack will cause danger. Therefore, their opinion is that the current license plate recognition network (LPR) and the security community need more investigation.

3) on Black Box Classifiers:

Sethi et al. [42] proposed that the enemy's perspective was portrayed based on ranking. Based on a formal opposition model, the SEE Paradigm to simulate data-driven generation and reverse engineering of classifier attacks is presented. Experimental assessment, conveying the inherent weakness of the classifier and encouraging evasion on ten real-world datasets and depend on the Google Cloud Prediction Tool to be used, reveals no precise details about the classifier sort, training data, or implementation region. The architecture, algorithms, and scientific assessment suggested is intended to set light on the weaknesses and facilitates the growth of stable machine learning systems.

4) on Instruction Deduction System:

Ayub et al. [43] showed how the evasion Attack is used against the functioning of the IDS. This paper presented two sections: (1) their use of the ML way for intrusion detection

and show the efficiency of this model with two various separate network-based IDS datasets; and (2) They executed an evasion attack on a multilayer perceptron network at an intrusion detection system using an aggressive machine learning technique known as the Jacobian-based Saliency Map Attack. Their experimental findings reveal that the evasion attack can considerably minimize the IDS's accuracy, including the identification of hostile transportation as benign. Their results confirm that IDS based on neural networks is liable to evasion attacks. And attackers can depend on this strategy to quickly get away from intrusion detection systems. The scheme below explains to the evasion attack model as shown in Fig. 5.

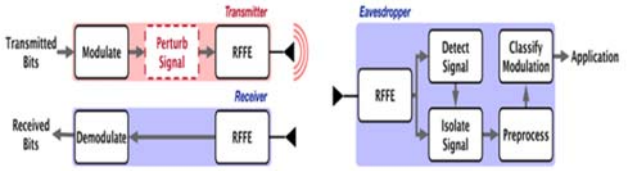

Fig. 5 scheme of evasion attack model versus a model of trained machine learning during the testing stage [43]

Alhajjar et al. [44] examined the contentious issue of the network intrusion detection systems (NIDS). They concentrate on the viewpoint of attack, which involves methods to produce exploratory samples that can evade an assortment of machine learning models. In particular, they are investigating the use of evolutionary computation (the optimization of the particulate swarm and the genetic algorithm) as well as profound learning (opposing generative network). They apply this to two openly accessible sets of data, i.e., NSL-KDD and UNSW-NB15, to evaluate the success of these algorithms by avoiding a NIDS and comparing them to the baseline disorder method: the simulation of Monta Carlo. The findings demonstrate that their adversarial example generation techniques in eleven separate machine learning models and a vote classification contribute to a high misclassification rate. Their analysis underlines the weakness in the face of detrimental disruptions of NIDS dependent on machine learning.

5) on Deep Neural Network:

HYUN et al. [45] Proposed a multi-target example of the adversary that targets many templates with a single modified image in each target class. A transformation is done to increase the likelihood of several models for various target groups. For their experiment, they used MNIST and TensorFlow datasets. The experimental results showed a 100 percent attack success rate with the proposed scheme for producing a multi-target opponent example.

6) on Deep Learning:

J. Dinal et al. [46] Discover deep learning anomaly detection models' adverse robustness on distributed system logs. They suggest the LAM (Log Anomaly Mask) technique for a real-time attack to interfere with online streaming logs with minor modifications to avoid attacks by outlier detection even by the state-of-the-art deep learning models. LAM models the disruptor as a reinforcing student to surmount the search space complexity challenge in a partially observable setting to predict the best disruption acts. They assessed the effectiveness of LAM on two log-based systems for outlier detection for distributed systems: DeepLog and AutoEncoder. Their test results reveal that, while accomplishing attack imperceptibility or real-time reactivity, LAM helps reduce the actual rate of such models.

7) on Graph Neural Network

Xixun et al. [48] developed a new exploratory attack (called EpoAtk) to improve gradient-based graph disturbances. EpoAtk experimental strategy comprises three stages, generation, evaluation, and recombination, to avoid any misinformation provided by the maximum gradient. EpoAtk is tested in experiments on benchmark data sets in various attack environments for the semi-controlled classification of nodes. Experimental findings reveal that the suggested approach beat the latest attacks using the same budgets of the attacks.

8) on Wireless Communications:

Bryce et al. [49] developed a technique to measure adversarial performance in wireless communications, in which a bit error and not the human interpretation is the key criterion of concern, as with image identification. This technique is used to know the classification of automatic modulation, which depends on raw intelligence, and showed that RFML is vulnerable to many damaging events, also in OTA attacks, by using a well-known Fast Gradient Sign Metering system. However, RFML domain-specific receiver results that could arise in an OTA attack could lead to severe impairments of adversarial evasion. Fig. 6 shows the evasion attack on an RFML system.

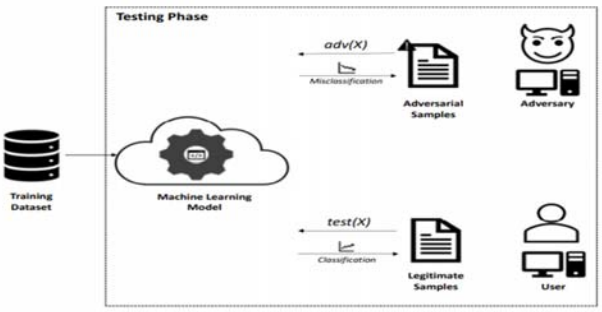

Fig. 6 system framework of a physical evasion attack on an RFML system performing AMC [49]

Brian et al. [47] showed how to use several antennas on the target to maximize the opponent's efficiency (evasion). The power distribution among the antennas and the use of channel diversity are two significant points considered when the opponents' multiple antennas are abused. First, they demonstrate that many individual opponents each have a specific antenna. Compared with one adversary with several antennas, using similar power cannot increase the attack efficiency. They then consider different forms of assigning power between the many antennas of one competitor, for example, giving control to just a single antenna and symmetric to the gain or reversed. Using various networks, they initiated an assault to convey adverse disruption across a channel with the most significant symbolic advantage. They found that in channel variation and channel correlation across antennas, this attack significantly decreases the classifying accuracy compared with other attacks with various channel conditions. They also demonstrated that the attack's effectiveness dramatically increases when there are antennas in the competitor, who will use the diversity of the channel to make adverse attacks more successful.

9) on Discrete Data:

Yutong et al. [50] is characterized by merging attack ability calculation and targeted classifier regularity with sensitive data in an evasion attack. Based on the study of the attack ability, they propose a computer-efficient orthogonal assembly system for directed attacks on discrete data. It provides computer productivity and accuracy of attacks. The experimental findings on the real-world datasets confirm the proposed requirements of attack ability and the efficacy of the proposed method of attack.

10) on Show and Tell model:

Dongseop et al. [51] produce an example using a forward-backward splitting method, which misclassifies the model. In addition, the evasion attack on the show and tell model was conducted and analyzed by using an adversarial example. Experimental results verified the effectiveness of the attack.

II. Data Poisoning Attack (Causative):

1) on Federated Learning Systems:

To begin et al. [52] researched targeted attacks against FL networks to intoxicate the global model by submitting model notifications from mistakenly labeled data in a malicious sub-set of participants. This shows that data poisoning attacks could cause significant declines in the precision of classification and reminder, even in a limited number of hateful members. Furthermore, they demonstrate that attacks are selective, i.e., only groups that are attacked have a significant adverse effect. They have also reviewed early/late-round training on attack durability, the impact of malicious involvement, and ties between both.

Xingchen et al. [61] conduct systemic surveys for federated learning risks and proposes a new model-driven poisoning attack based on optimization. Unlike current approaches, they mainly concentrate on attacks' efficiency, continuity, and steadiness. Numerical research shows that the suggested approach will achieve high attack effectiveness and sufficiently stealthy to circumvent two present protection approaches, as shown in fig.7

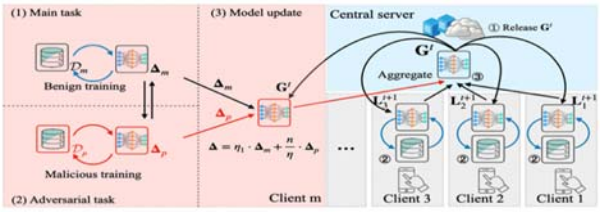

Fig. 7 deep model poisoning attack as the pipeline [52]

2) on Image Classifiers:

Truong et al. [53] developed research for ML image classification focused on backdoor data poisoning, systematic evaluation of some experiments including types of trigger models, the durability of retraining trigger trends, poisoning methods, design (ResNet-50, NasNet, NasNet-Mobile), data sets (Flowers, CIFAR-10) as well as possible defensive regulatory techniques such as Contrastive Loss, Logit Squeezing, Manifold Mixup, and Soft-Nearest-Neighbors Loss. Experiments provide four main results. Firstly, the success rate of backdoor poisoning attacks varies considerably based on many factors such as model design, the pattern of activation, and the technique of regularization. Secondly, it is impossible to identify poisoned models by the performance inspection alone. Third, regularization usually decreases the effective rate of backdoors, but it may have little effect or even marginally improve it depending on their regularization processes. After just several years of extra training on a limited collection of clean and clear data without compromising the output of the model, backdoors introduced with data poisoning will eventually become ineffective. Fig. 8 shows an example of data poisoning in a flower picture.

Fig 8. image from the Flowers dataset applied by Trigger patterns.[53]

3) on Bayesian network:

Emad et al. [54] stressed the importance of creating and using data privacy applications as a core component of machine learning software. For this reason, they analyze how an opponent might build a desired PC algorithm network. Centered on the Bayesian B1 network and the DB1 database generated with the B1 and B2, which is the same as B1, they investigate and evaluate the minimum number of modifications, such as inclusion, deletion, and replacement for the B1 database, leading to DB2 as an input in the PC algorithm, results in B2.

4) on Show and Tell model:

Lee et al. [55] created opponent data, which adjusts the feature values to different targets by processing the least number of RPG values to identify images through a single model, poisoning attacks on the show and tell model. Results showed success in filtering adverse data using a deeper neural network, autoencoding, and protecting against poisoning attacks.

5) on Graph Neural Networks:

Daniel et al. [56] Presented a review of adverse attacks on graphs that are credited, in particular models that abuse graphic convolution ideas. They also discuss the most complicated type of toxic/causative attacks that concentrate on the training stage of a machine learning model, in addition to attacks at test periods. Adversarial disruptions are generated aiming at the node and the graph structure features, taking account of dependencies between instances. Furthermore, by maintaining essential data resources, they guarantee that the disruptions stay unnoticed. They suggest an effective Nettack algorithm that allows gradual computations feasible to cope with the underlying discrete domain. Their experimental research reveals that even in a few perturbations, the precision of the node classification decreases dramatically. Moreover, their attacks can be transferred: Trained seizures are generalized to additional modern node classification models and unregulated methods and even efficient even if there's only minimal graphic information.

Dineen et al. [59] studied the data Poisoning issue using reinforcement training agents in neural networks for classification. The agent is conditioned to develop an optimum strategy under Reinforcement Learning (RL) principles so that the graphs or functions under Black Box Attack are injected, modified, or deleted. The analyses show that the process gives an additional picture of the vulnerability of particular graphing structures by a random, brute force search of graph space.

6) on Deep Learning:

Yi et al. [57] implemented an adversary learning method for initiating a spectrum information poisoning attack. In other words, an attacker knows the driving actions of the transmitter and manages to false the spectrum sensing data in the air by transmitting the input for the transmitter's decision-making process for a brief period while the channel is idle. Compared to interfering with data transfers, this assault is much more energy-efficient and harder to locate. Results demonstrate that this attack is much successful and dramatically decreases the transmitter's performance.

7) on Generative Adversarial Nets:

Luis et al. [58] created a new generative model of poisoning attacks on machine grading, which yields examples of adverse training, i.e., samples that look like real data points but compromise the classifier's accuracy in a training phase. The Generative Adversarial Net is suggested to have three components: generator, discriminator, and goal classifier. Allows them to model detectability constraints that have been expected in functional assault and classify regions more likely to be toxic to data set from the underlying distribution of data. Their experimental assessment indicates that their attacks on machine organizing, like deep networking, are booming.

8) on Data Complexity Measures:

Patrick et al. [60] formulate causative attack identification s a second-order classification problem where it represents a data set quantified by measurements of data complexity. Documenting data's geometrical properties means a sampling problem. As a consequence of a causal attack, the geometric nature of a dataset modifies, data complexity measures offer valuable knowledge for identifying causative assaults. A two-stage stable classification model is also proposed to show how the suggested cause attack identification enhances the robustness of learning. Experimental findings demonstrate that data complexity precisely measures unchanged data sets from those attacked and affirm the positive success in terms of precision and power of the proposed methods.

Table 2 shows a summary of all previous studies with the type of the adversarial attack that have been used.

Table 2: Related work summary

Author	***Year***	***Contribution***	***Type of Attack***
Sayghe et al.[40]	2020	Studied poisoning attacks and their effect on support vector machines and multilayer receptors.	Evasion
Qian et al.[41]	2020	Proposed an evasion attack on CNN classifiers in the context of License Plate Recognition (LPR).	Evasion
Sethi et al.[42]		Introduced framework used to stimulate the production	Evasion

Author	*Year*	*Contribution*	*Type of Attack*
	2018	of data-based attacks in classifications and reverse engineering. Where used in that (Google Cloud Prediction Platform).	
Ayub et al.[43]	2020	Developed a machine learning method to detect intrusion in multilayer perceptron networks. Then, executed an evasion attack on multilayer perceptron networks at an intrusion detection system using an aggressive machine learning technique known as the Jacobian-based Saliency Map Attack.	Evasion
Alhajjar et al.[44]	2020	Studied the form of the adversarial problem that exists in intrusion detection systems. Using evolutionary computation by optimizing particle swarm and genetic algorithm and used deep learning to generate negative examples.	Evasion
To begin et al.[52]	2020	Studied the poisoning attacks in the federated learning systems, where a group of malicious participants seeks to poison the basic model by sending false data forms .	Data Poisoning
Truong et al.[53]	2020	Evaluated various testing procedures, including trigger pattern sort, the durability of retraining trigger patterns, poisoning tactics, frameworks, data sets, and protective strategies.	Data Poisoning
Emad et al.[54]	2020	Studied how an opponent might create a desired Bayesian network using the PC structure learning algorithm .	Data Poisoning
Lee et al.[55]	2020	Produced adverse data to modulate selected features to various goals with fewer RGB values.	Data Poisoning
Daniel et al.[56]	2020	Produced adversarial disruptions to the features and structure of the node. In addition, an effective algorithm has been proposed Nettack uses incremental calculations.	Data Poisoning
Yi et al.[57]	2018	Proposed an attack depends on a deep neural network to establish an air data sensing poisoning attack. By exploiting the transmitter's input data	Data Poisoning

Author	*Year*	*Contribution*	*Type of Attack*
		during runtime and leading to incorrect transmittal decisions.	
Luis et al.[58]	2019	Proposed the Generator, Discriminatory, and Classifier Generative Adversarial Net to perform data poisoning attacks.	Data Poisoning
Dineen et al.[59]	2021	Studied the new Data Poisoning attack (training time) problem on neural graph-grading networks using reinforcement learning.	Data Poisoning
Patrick et al.[60]	2021	Proposed a two-stage safe classification model to show how the suggested identification of causative attacks increases learning robustness.	Data Poisoning
Xingchen et al.[61]	2021	Proposed a novel optimization-based model poisoning attack.	Data Poisoning
HYUN et al.[45]	2018	Proposed a multi-targeted adversary example targeting several templates with a single updated image within each target class.	Evasion
J. Dinal et al.[46]	2021	Suggested the LAM (Log Anomaly Mask) approach to disrupt streaming logs with minor changes. Real-time attack proposed approach.	Evasion
Brian et al.[47]	2018	Showed that the planned EMCG attack exceeds other attacks and efficiently uses multiple antennas to induce recipient misclassification of the channel diversity.	Evasion
Xixun et al.[48]	2020	Created new exploratory antagonistic attack (named EPoAtk) to increase gradient-based graph disturbances.	Evasion
Bryce et al.[49]	2019	Developed a technique for measuring adversarial performance in wireless communications, where bit error and no human perception is the critical parameter of concern, as is picture detection.	Evasion
Yutong et al.[50]	2020	Proposed a practical computerized orthogonal pursuit-guided attack	Evasion

Author	*Year*	*Contribution*	*Type of Attack*
		approach to attack discrete data by evasion.	
Dongseop et al.[51]	2020	Created an adversarial case using the iterative method for forwarding backward splitting.	Evasion

4. Discussion

In this section, we will discuss the different attacks that occurred in Influence Axis for adversarial machine learning, which we have limited to two attacks: data poisoning and an evasion attack; as we noted in the previous studies that we mentioned, the possibility of the two episodes occurring in many different models and systems. We will talk about some of them.

For example, in federated learning systems, in which the training data is characterized by being decentralized between many devices such as computers and mobile devices and characterized by the presence of more than one customer involved in the data training process, where each client will keep his data locally and cannot access the data of other customers and share model updates only with a central server. Here the risk of adversarial attack begins for one of those agents participating in the learning process to be an adversary agent and may benefit from the available information in carrying out the data poisoning attack and the white box attack, given that he has sufficient data of the learning model and also the learning algorithm used. The adversary client will also make sure to remain anonymous among the many clients.

In one of the studies that we presented, they overcame two types of defenses and achieved a high attack rate by designing an attack strategy that poisoned the model instead of the data. Even the central server cannot evaluate the accuracy of model updates and detect the opponent's attack. The malicious client trains a small group of a clean sample and a sample with poisoning in it alternately in the form to maintain the high performance of the model and so that the central server does not reject the outlier forms. As a solution for their attack, they suggested that the central server exclude the client who did not send the form update on time, as it would be assumed to be a malicious client. On the other hand, in another study of the same attack and also in federated learning systems, the data was poisoned without the model. They depended on the large number of malicious clients involved in learning the model, which would facilitate their work without being discovered by the central server. The damage of the attack would be more serious. And they provided a solution for this attack, a mechanism to compare parameters sent from malicious clients with parameters sent from honest clients.

Among other areas that have used machine learning, intrusion detection systems, which monitor unusual network activity, and are mainly considered a wide area for attacks to occur, where we rely on a human analyst to identify abnormal behaviors and therefore may fail to notice and discover some types of intrusion, which led to thinking of integrating Machine learning models in network detection systems. Researchers have used many machine learning classifiers in intrusion detection systems such as artificial neural networks, decision trees, and supporting vector machines to increase their performance and strength. Here comes the question: What if the attacker could influence this learning model? Of course, this will affect it negatively on the performance of the intrusion detection system.

In one of the research papers we mentioned, the intrusion detection system was integrated with machine learning through a multilayer model (MLP) to perform binary classification on benign and adversarial traffic and then apply an evasion attack against this model. And they were able to influence the accuracy of the model's performance to detect intrusion, making the model classify false attack records as benign. In another study, they used evolutionary computation and deep learning to generate adversarial examples to evade intrusion detection systems. There are many scientific papers in this field using different and varied models. Still, they all meet one goal: to make the intrusion detection systems classify hostile attacks as benign and to reduce the performance of the intrusion detection model without the opponent being detected.

The fields of machine learning use did not end there, and they reached wireless communications. It is used to send signals to the receiver with other modifications; here comes the role of machine learning, specifically deep understanding. The classifier in the receiver based on machine learning classifies the types of alterations for future signals. Consequently, the evasion attack appeared, where one of the papers mentioned that the opponent could send hostile disturbances on the original inputs of the deep neural network by using multiple antennas to deceive the classifier and make him misclassify the future signals. The opponent classifies the hostile disturbances for each antenna with high accuracy to increase the attack strength. Also, the opponent used the white-box attack, which assumes that the opponent knows the structure of the receiver's classifier and the input on the receiver. The article stated that the number of opponents with one antenna will not give a good result and will not increase the attack performance. On the other hand, with another paper in the same field and for the same attack, the antennas were not considered. Where the opponent tries to eavesdrop and detect and insulate the signal in spectrum in time and frequency then classifying the modification on it.

Acknowledgments

Insert acknowledgment, if any.

5. Conclusion

Machine Learning has been one of the highly talked-about subjects at this time. Machine learning has been pervasively used in a broad domain of applications, networking, computer systems, cloud, and even hardware, and It Shows excellent success in handling several complex issues. However, machine-learning algorithms are Prone to adversarial attacks. There are a significant number of research studies on hostile attacks. We focused in our research paper only on evasion and data poisoning attacks, where this paper introduced a comprehensive survey that highlighted adversarial machine learning in the cyber area. Moreover, Imperious to motion that this work focuses only on the newest documents that have been published during 2018-2021. On the other hand, more investigations are kept for future work involving strong countermeasures for adversarial attacks.

References

[1] Puiutta, E., & Veith, E. M. (2020, August). Explainable reinforcement learning: A survey. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 77-95). Springer, Cham.

[2] Gu, R., Niu, C., Wu, F., Chen, G., Hu, C., Lyu, C., & Wu, Z. (2021). From Server-Based to Client-Based Machine Learning: A Comprehensive Survey. ACM Computing Surveys (CSUR), 54(1), 1-36.

[3] Faria, J. M. (2018, February). Machine learning safety: An overview. In Proceedings of the 26th Safety-Critical Systems Symposium, York, UK (pp. 6-8).

[4] Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modeling and graphics (pp. 99-111). Springer, Singapore.

[5] Gu, R., Niu, C., Wu, F., Chen, G., Hu, C., Lyu, C., & Wu, Z. (2021). From Server-Based to Client-Based Machine Learning: A Comprehensive Survey. ACM Computing Surveys (CSUR), 54(1), 1-36.

[6] Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modeling and graphics (pp. 99-111). Springer, Singapore.

[7] Gu, R., Niu, C., Wu, F., Chen, G., Hu, C., Lyu, C., & Wu, Z. (2021). From Server-Based to Client-Based Machine Learning: A Comprehensive Survey. ACM Computing Surveys (CSUR), 54(1), 1-36.

[8] Puiutta, E., & Veith, E. M. (2020, August). Explainable reinforcement learning: A survey. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 77-95). Springer, Cham.

[9] Sequeira, P., Gervasio, M.: Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations (2019)

[10] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. nature 550(7676), 354–359 (2017)

[11] Kober, J., Bagnell, J.A., Peters, J.: Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32(11), 1238–1274 (2013)

[12] Arel, I., Liu, C., Urbanik, T., Kohls, A.: Reinforcement learning-based multi-agent system for network traffic signal control. IET Intelligent Transport Systems 4(2), 128 (2010)

[13] Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey (1996)

[14] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences 108, 17 (2011), 6889–6892.

[15] Anne O'Keeffe and Michael McCarthy. 2010. The Routledge handbook of corpus linguistics. Routledge.

[16] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In Ethics of Data and Analytics (pp. 254-264). Auerbach Publications.

[17] Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY, USA.

[18] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

[19] Pacheco, Y., & Sun, W. (2021). Adversarial Machine Learning: A Comparative Study on Contemporary Intrusion Detection Datasets.

[20] Laskov, P. and Lippmann, R., 2010. Machine learning in adversarial environments.

[21] Shirazi, S. H. A. A Survey on Adversarial Machine Learning.

[22]] "Responsible AI Practices." [Online]. Available: https://ai.google/ responsibilities/responsible- ai-practices/?category=security

[23] "Securing the Future of AI and ML at Mi- Microsoft." [Online]. Available: https://docs.microsoft.com/en-us/security/ securing-artificial- intelligence- machine- learning

[24] "Adversarial Machine Learning," Jul 2016. [Online]. Available: https://ibm.co/36fhajg

[25] S. A. Gartner Inc, "Anticipate Data Manipulation Security Risks to AI Pipelines." [Online]. Available: https://www.gartner.com/doc/3899783

[26] Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., ... & Xia, S. (2020, May). Adversarial machine learning-industry perspectives. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 69-75). IEEE.

[27] Wang, X., Li, J., Kuang, X., Tan, Y. A., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing, 130, 12-23.

[28] L. Qin, Y. Guo, W. Jie, G. Wang, Effective query grouping strategy in clouds, J. Comput. Sci. Technol. 32 (6) (2017) 1231–1249.

[29] L. Qin, G. Wang, L. Feng, S. Yang, W. Jie, Preserving privacy with probabilistic indistinguishability in weighted social networks, IEEE Trans. Parallel Distrib. Syst. 28 (5) (2017) 1417–1429.

[30] Wang, X., Li, J., Kuang, X., Tan, Y. A., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing, 130, 12-23.

[31] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A survey on security threats and defensive techniques of machine learning: A data-driven view. IEEE Access, 6, 12103-12117.

[32] Duddu, V. (2018). A survey of adversarial machine learning in cyber warfare. Defense Science Journal, 68(4), 356.

[33] Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., & Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, 100199.

[34] Huang Xiao. 2017. Adversarial and Secure Machine Learning. Ph.D. Dissertation. Universität München. https: //mediatum.ub.tum.de/1335448 04 de fevereiro de 2019.

[35] Shirazi, S. H. A. A Survey on Adversarial Machine Learning.

[36] N. Dalvi, P. Domingos, S. Sanghai, D. Verma, et al., Adversarial classification, in Proceedings of the Tenth ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 99–108.
[37] Khasawneh, K.N., Abu-Ghazaleh, N., Ponomarev, D. and Yu, L., 2017, October. RHMD: evasion-resilient hardware malware detectors. In Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 315-327).
[38] L Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in Proc. 4th ACM Workshop Secure. Artif. Intell., 2011, pp. 43–58.
[39] Martins, N., Cruz, J. M., Cruz, T., & Abreu, P. H. (2020). Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access, 8, 35403-35419.
[40] Sayghe, A., Zhao, J., & Konstantinou, C. (2020, August). Evasion attacks with deep adversarial learning against power system state estimation. In 2020 IEEE Power & Energy Society General Meeting (PESGM) (pp. 1-5). IEEE.
[41] Qian, Y., Ma, D., Wang, B., Pan, J., Wang, J., Gu, Z., ... & Lei, J. (2020). Spot evasion attacks: Adversarial examples for license plate recognition systems with convolutional neural networks. Computers & Security, 95, 101826.
[42] Sethi, T. S., & Kantardzic, M. (2018). Data-driven exploratory attacks on black-box classifiers in adversarial domains. Neurocomputing, 289, 129-143.
[43] Ayub, M. A., Johnson, W. A., Talbert, D. A., & Siraj, A. (2020, March). Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning. In 2020 54th Annual Conference on Information Sciences and Systems (CISS) (pp. 1-6). IEEE.
[44] Alhajjar, E., Maxwell, P., & Bastian, N. D. (2020). Adversarial Machine Learning in Network Intrusion Detection Systems. arXiv preprint arXiv:2004.11898.
[45] Kwon, H., Kim, Y., Park, K. W., Yoon, H., & Choi, D. (2018). Multi-targeted adversarial example in evasion attack on deep neural network. IEEE Access, 6, 46084-46096.
[46] Herath, J. D., Yang, P., & Yan, G. (2021). Real-Time Evasion Attacks against Deep Learning-Based Anomaly Detection from Distributed System Logs.
[47] Kim, B., Sagduyu, Y. E., Erpek, T., Davaslioglu, K., & Ulukus, S. (2020). Adversarial attacks with multiple antennas against deep learning-based modulation classifiers. arXiv preprint arXiv:2007.16204.
[48] Lin, X., Zhou, C., Yang, H., Wu, J., Wang, H., Cao, Y., & Wang, B. (2020, November). Exploratory Adversarial Attacks on Graph Neural Networks. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 1136-1141). IEEE.
[49] Flowers, B., Buehrer, R. M., & Headley, W. C. (2019). Evaluating adversarial evasion attacks in the context of wireless communications. IEEE Transactions on Information Forensics and Security, 15, 1102-1113.
[50] Wang, Y., Han, Y., Bao, H., Shen, Y., Ma, F., Li, J., & Zhang, X. (2020, August). Attack ability Characterization of Adversarial Evasion Attack on Discrete Data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1415-1425).
[51] Lee, D., Kim, H., & Ryou, J. (2020, February). Evasion Attack in Show and Tell Model. In 2020 22nd International Conference on Advanced Communication Technology (ICACT) (pp. 181-184). IEEE.
[52] Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020, September). Data poisoning attacks against federated learning systems. In European Symposium on Research in Computer Security (pp. 480-501). Springer, Cham.
[53] Truong, L., Jones, C., Hutchinson, B., August, A., Praggastis, B., Jasper, R., ... & Tuor, A. (2020). Systematic evaluation of backdoor data poisoning attacks on image classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 788-789).
[54] Alsuwat, E., Alsuwat, H., Valtorta, M., & Farkas, C. (2020). Adversarial data poisoning attacks against the pc learning algorithm. International Journal of General Systems, 49(1), 3-31.
[55] Lee, D., Kim, H., & Ryou, J. (2020, February). Poisoning Attack on Show and Tell Model and Defense Using Autoencoder in Electric Factory. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 538-541). IEEE.
[56] Zügner, D., Borchert, O., Akbarnejad, A., & Guennemann, S. (2020). Adversarial attacks on graph neural networks: Perturbations and their patterns. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(5), 1-31.
[57] Shi, Y., Erpek, T., Sagduyu, Y. E., & Li, J. H. (2018, October). Spectrum data poisoning with deep adversarial learning. In MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM) (pp. 407-412). IEEE.
[58] Muñoz-González, L., Pfitzner, B., Russo, M., Carnerero-Cano, J., & Lupu, E. C. (2019). Poisoning attacks with generative adversarial nets. arXiv preprint arXiv:1906.07773.
[59] Dineen, J., Haque, A. S. M., & Bielskas, M. (2021). Reinforcement Learning For Data Poisoning on Graph Neural Networks. arXiv preprint arXiv:2102.06800.
[60] Chan, P. P., He, Z., Hu, X., Tsang, E. C., Yeung, D. S., & Ng, W. W. (2021). Causative label flip attack detection with data complexity measures. International Journal of Machine Learning and Cybernetics, 12(1), 103-116.
[61] Zhou, X., Xu, M., Wu, Y., & Zheng, N. (2021). Deep Model Poisoning Attack on Federated Learning. Future Internet, 13(3), 73.
[62] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature Squeezing: Detecting adversarial examples in deep neural networks. Network and Distributed Systems Security Symposium (NDSS) 2018 (2018). https://doi.org/10.14722/ndss. 2018.23198 [196]
[63] Ziang Yan, Yiwen Guo, and Changshui Zhang. 2018. Deep Defense: Training DNNs with improved adversarial robustness. In Advances in Neural Information Processing Systems. 419–428. [197]
[64] Yuzhe Tang, Guo Zhang, Dina Katabi, and Zhi Xu. 2019. ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation. arXiv preprint arXiv:1905.11971 (2019)