Adversaries on ML Models: The Dark side of Learning

[1*]Sahithi Godavarthi, [2]Dr. G.Venkateswara Rao

[1a]Research Scholar, Dept. of CSE, GITAM School of Technology, GITAM(Deemed to be University), Visakhapatnam
[1b]Assistant Professor, Department of Emerging Technologies, CVR College of Engineering,Hyderabad
*sahithi.godavarthi@gmail.com

[2]Professor, Dept. of CSE, GITAM School of Technology, GITAM(Deemed to be University), Visakhapatnam
venkateswararao.gurrala@gitam.edu

Abstract. Today's technological trends are advancing to new levels and showing a diverse array of uses. One of these that has recently grown in prominence is machine learning. The ability of ML to analyze data, learn, make decisions and predictions made it the outstanding technology to be used in plentiful of gadgets. Conversely, adversaries also affect ML models in different phases. One challenge for ML users is therefore to make the models robust before using them in applications. The focus of this work is on the several hostile scenarios that machine learning models encounter and the countermeasures that can be taken to lessen the opponents' influence. There is a need for study that concentrates on creating stronger defenses against assaults on ML models. This paper can provide a full overview of machine learning (ML) and its history. It also outlines future research possibilities for securing ML models.

Key words: Adversary, Resilient, Mitigate, Counter measures, Safe guarding ML

1 Introduction

The advent of the Internet of Things, Cyber Physical Systems, and other offerings that can network data, resources, materials, and the public could not have occurred without the Web surrounding us in every way. It created a linked and smart environment. The recent advancements of technological environment is leading to the development of trending devices like IoT, CPS and other smart appliances. These devices are intended to bring more comfort to human life and make our tasks easier and accurate. Industry 4.0's future depends on the integration of several emerging industries, including Intelligent Manufacturing Systems (IMS), Cyber Physical Systems (CPS), and the Internet of Things (IoT). By combining and computerizing the device and its surroundings quickly, these technologies all operate together [19].

Employed in critical domains such as farming, medical, transportation, defense, domestic automation, and electrical structures, IoT and CPS systems produce a significant amount of data because all these devices are always connected and running. This resulted in the Big-Data implications for IoT and CPS [14].

The difficulty with these gadgets at this time is their security, which needs to be addressed. Cyber attacks have increased in frequency over the past few years, and stopping them is getting increasingly challenging. Attacks shown in a planned manner may commit financial fraud, which is becoming common these days. The majority of online assaults target a certain item or objective that may

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_124

be highly exploited. To defeat these attacks, it is imperative that new countermeasures be investigated [12].

A recent development in artificial intelligence technology is machine learning. A wide number of industries, including data analytics, cyber security, and numerous other ones, have made extensive use of machine learning (ML). When it comes to developing and maintaining network security and protocols that make it simpler for computers to train and learn, ML is fastidious. ML also assists to handle huge data in less time. Machine learning computations are used for malware inspection and categorization in both static and dynamic ways. ML can also be used in cyber attack detection to identify whether a system is secure or being attacked [15]. To identify network-based dangers and to assist stop them, supervised and unsupervised ML models can be deployed [28]. ML has also been successful in resolving the problem of zero-day malware detection, which traditional sign-based techniques are unable to handle [9].

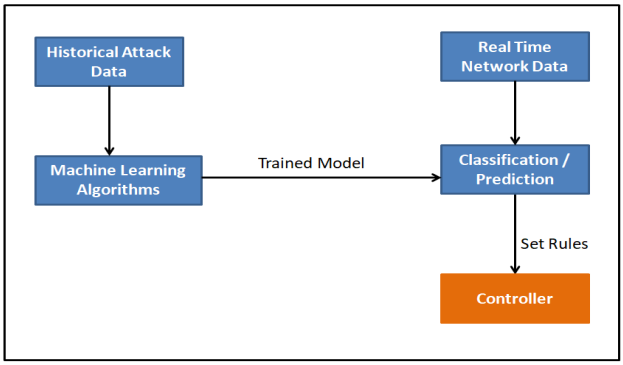

Fig.1.Training ML models with attack data

Fig.1 shows how ML models are conditioned using attack data. To make the machine learning model opponent aware, it is trained using historical attack data. Presently, the trained model creates rules for the controller and uses real-time network data to complete classification and prediction tasks. Before using ML models for securing systems, the resiliency of the algorithms should be assured. The adversaries may compromise the models in multiple aspects and various kinds of attacks are taking place on the ML models. For ML models to be robust and useful for system security, researchers must come up with ways to counter this [14]. Machine learning is currently available as a service, and the owners are providing this service to their clients, making ML technology available to any user. Privacy breaches may happen in this context also [6].

2 Adversaries on ML models

The recent trends in ML has shown incredible progress. These technologies are being used widely and in a variety of contexts. The victory of ML can be affected by the adversaries which pose a serious challenge to the security of ML models [8]. Generally ML models can be subjected to adversaries by inducing

little perturbations to the clear datasets which leads to the misclassification by the ML model [24]. Adversaries on Machine Learning framework can be understood based upon numerous factors, including the adversary's aim, the attack phase, and the attacker's understanding of the targeted model [14].

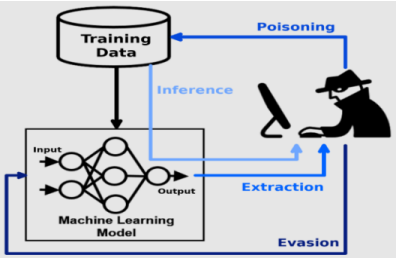

Fig.2-Adversarial attempts on ML model

Table-1: Attacks and proposed techniques to overcome by various writers

Authors	Attack Addressed	Proposed Technique	Operation
Reza Shokri et al.[6]	Membership-inference attack	Shadow-training technique	Teaches an attack model how to distinguish between training dataset members and non-members based on the target model's output.
Derui Wang et al.[20]	Examples of DNN classifiers that are adversarial	Malicious-VAE Decoder (MVD)	Create hostile examples having high success rates that misclassify the DNN classifiers.
Peva Blanchard et al. [7]	Byzantine failures	Krum	a combination rule that meets the resilience requirement
Pieter Delobelleet al.[16]	Unfair representations stemming from the training datasets	Structure that helps with relieving unreasonable portrayals originating from the preparation datasets	Three very much examined datasets and demonstrated the way which can convey elevated practicality to frameworks with fulfilling decency standards
Krishna Yadav et al. [21]	Foes in a united AI based IoT climate	Gradient-filtration protocol	Proficient in identifying foes and killing them
Elan Rosenfeld et al.[5]	Poisoning attacks	Model-driven strategy	Cross attack technique is used to protect cross models
N. Papernotet al.[3]	Adversarial attacks	Distillation based defense	Network taught using standard, methodologies.

2.1 Inference Attacks:

The expanded approval of ML models in diversified applications has raised novel concerns pertaining to confidentiality and security. Inference attacks are of one such kind which cause privacy breach. The attacker makes the destined ML model to disclose the details regarding its training data [6]. Due to their requirement to preserve their training info, machine learning models are susceptible to a variety of assaults, including model extraction attacks [22], attribute inference attacks [17], property-inference attacks [27], and membership-inference attacks [6]. Model retreival aims to replicate the performance of the framework while the adversary attempts to create a competing algorithm constituting functionalities comparable to the destined model. The other inference adversaries, however, such as attribute-inference, property-inference, and membership-inference attacks, seek to interpret sensitive training data information. While property inference attacks aim to infer complete properties of the training data, and attribute-inference attacks aim to infer sensitive and non-sensitive attributes of the target data record, membership inference attacks seek to determine whether a specific data record was used to train the destined model. [25].

2.2 Backdoor attack

Machine learning programs are increasingly being subjected to backdoor assaults. Through this backdoor assault, the adversary can easily trespass even highly secured authentication system. When developing the machine learning model that misclassifies the input, the backdoor attack's perpetrator inserts a static initiate to a particular aimed label. To be able to launch backdoor attack, the adversary adds the trigger into the pure dataset to create the back door data and may modify the labels of the aimed label. The adversary creates the back door info with inserting the initiate to the pure dataset which may change the labelling to the targeted label in order to initiate a backdoor attack. In order for the model to understand both its intended goal and the back door nature, the attacker now trains it using both clean and back door data. This way the backdoor attacks raises severe security concerns [10]. The authors of [23] showed the attack by tainting the practice information which is useful for creating the model. Findings showed that around 99 percent of compromised inputs were incorrectly assigned the target label of the adversary.

2.3 MITM attack

Among the most dangerous assaults that compromise the security of the system is the man-in-the-middle attack. The adversary of MITM tries to modify the conversation between the two ends by distributing pernicious payload to the two ends of communication. By this, adversary can be able to achieve his adversarial intention. Adversaries may be able to exploit a pre-processing classification function that stages the input data using a generative model before submitting it to the classifier for decision-making. In this case, the MITM attacker gets in the way of the categorization task and the data collection. These assaults aim to trick the classifier by changing the generative model's input data, which could result in the production of inaccurate output. [20]

2.4 Poisoning Attack

The data poisoning attacks [2] and evasion attacks are becoming serious threats to the security of the ML models [11]. Attacks using data poisoning are a potential threat when machine learning models are being trained. Introducing harmful input into the data sets needed to train the machine learning models is the main objective of this attack. As data-driven assaults, poisoning attacks come with deportation of assaulting numerous machine

learning models. By adding phony poisoned samples to the training data set, this technique causes the model to react strangely to poisoned input. The model may become hostile as a result of using the poisoned sample into training [18]. Protection against poisoning attacks, whether intentional or accidental Create learning algorithms that can handle certain hostile input, usually by identifying and removing out-of-distribution locations [1].

2.5 Black Box Attack

Black box attack is highly useful today because they enable the adversary to query the model even while he is unfamiliar with specific model information. [24]. The internal workings of the model are barely known to the adversary. In this case, the attacker tries to initiate the attack by examining the relationships between input and output datasets. In this instance, the attacker may either directly alter the input dataset or utilize an agent learning model to design the attack, which leads to the model's misclassification [14]. Hard label black box attacks do not rely on probability vectors like soft label black box attacks do; instead, the opponent has access to merely the final anticipated label in these cases. [24].

2.6 Enhancement Attack

Attack like this might occur in bio-medical machine learning. Although machine learning is becoming common in biomedical research, the reliability of this study is sometimes overlooked. Potential threats to biomedical machine learning could come from newly created "enhancement attacks" that secretly enhance performance rather than from studies looking at how adversarial attacks could impair model performance in medical imaging. Attackers' capabilities in enhancement attacks are significantly greater than in poisoning attacks. Due to their capacity to alter the entire dataset, which in the case of most within-dataset predictions may include both training and test data, the attacker in the research setting of enhancement assaults has even more knowledge than in the conventional "white-box" configuration [28].The below figure depicts the various ways in which the attacker may attempt to create various adversarial effects on the ML model.

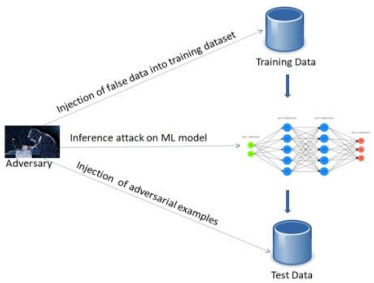

Fig.3- Creating negative impacts on the test data and training data

3 Crafting Adversarial examples

As inputs for machine learning models, adversaries created what are referred to as adversarial samples in an effort to weaken the model. Adversarial inputs include those that are given to neural networks with the intention of misclassifying them. Examples that have perturbations added to them are referred to as adversarial examples. Machine learning-based malware detection algorithms cannot be employed in real-world scenarios

if certain adversarial tactics can be swiftly defeated by them [9]. Here are few Adversary crafting methods:

Fast Gradient Sign Method (FGSM):

Adversarial perturbations were introduced into the input images using the gradient or derivative of the model's loss function with respect to the input feature vector [2].

Generative Adversarial Network(GAN):

Gradient information and custom rules are the main tools utilized in this strategy to turn original samples into adversarial instances. [9] suggests a solution based on generative neural networking that utilizes original samples as inputs and produces adversarial examples. Neural networks can produce more complex and adaptive adversarial samples to fool the target model because of their innate non-linear structure.

Adversarial Transformation Networks (ATNs):

Adversarial instances can be produced by instructing the network to disrupt the input or by utilizing an adversarial auto-encoding of the input. The method's ability to give both targeted and non-targeted attacks and execute training in a white-box or black-box manner makes it appealing [11].

Deep Fooling:

Geometrical ideas are utilized to look for the smallest disturbance required to trick a classifier into assigning something to the wrong category. Additionally, it employs the L2 minimization algorithm to look for adversarial examples [4].

Table-2: Various methods for crafting Adversarial Examples

Author	Technique	Approach	Remarks
I. J.Goodfellow et al.[2]	FGSM	The technique is quick and uses less resources, enabling adversarial training to become a reality.	Creating adversarial examples is really easy, which makes it especially popular. played a more vulnerable antagonistic attack after being initially involved to assess the effectiveness of most users.
Moosavi-Dezfooli et al. [4]	Deep Fool	Adds to the construction of more robust classifiers by making deep classifiers more resilient to adversarial perturbations than previous methods, especially in large data sets.	With the use of this technique, it is possible to demonstrate the adversarial training considerably raises robustness by computing a more ideal adversarial perturbation.
Baluja et al. [11]	Adversarial Transformation Network	Any input could become an adversarial input by using a different network that has been trained to attack the target network. Benefits include control over the kind	Unknown to most people, a relatively new method has the power to completely transform this area of study

		of misclassification, non-transferability, and quick and efficient training due to single-forward pass.	
Papernot et al. [5]	Jacobian Saliency Map Approach (JSMA)	Building adversarial saliency maps with knowledge of the target model, the adversary uses this information to determine which input features are most important for output classification. It then attacks these features with large perturbations.	It is effective since it is a targeted assault and only needs to change a small number of pixels in the input image.
Szegedy et al. [1]	Penalized optimization problems	Adversarial instances were generated by solving optimization problems, demonstrating that neural networks are vulnerable to adversarial instability.	Time-consuming and poorly scalable to large data sets was procedure

4 Mitigating Approaches

Federated Learning is the most contemporary technique to counterpart adversaries against training models. By retaining the training data on the local system where it was created, federated learning is implemented. Data decentralization removes the need for data sharing during the model's training phase and creates a secure environment against outside threats. Finding a balance between data secrecy and model achievement is necessary, though, in order to use Federated Learning systems. In addition, it might be challenging to provide security and privacy in a Federated Learning context [26].

Poisoning attacks must be addressed in order to deal with the tainted data that leads to the model's misclassification. The poisoned sample leads to adversarial situations [3]. When using conventional machine learning, the model's training phase can be used to implement defense strategies. The strategies can be done either preceding the preparation or during the preparation. Prior to training the model, during the pre-processing phase, data-driven countermeasures are implemented. These methods often operate independently of the learning algorithm. Often, model-driven preventative measures are implemented during the model's training phase. The learning algorithms can be modified to be resilient against poisoned data. So, the model driven counter measures are algorithm specific [18].

One popular method to make ML algorithms resistant to adversaries is adversarial training. This process creates adversarial samples, which are then used to teach the machine learning system how to fight against attacks. This adversarial training helps to diminish the erroneous results caused by perturbed input. The model may appeal to provide labels for new datasets. To improve the efficacy of this approach, numerous antagonistic training adjustments are developed [14].

Table-3: Various attack types on ML models and their operation

Model Category	Attack	Operation
ML	Inference Attack [6]	Identifies if a certain record was used to train the adversarial model.
	Back Door Attack [10]	After contaminating the training set, use this contaminated set to train the algorithm.
	Black Box Attack [24]	An analysis of the relationships between the input and output data sets
	Man-in-the-Middle [20]	data manipulation that is done covertly to provide inaccurate results
	Poisoning Attack[18]	Adds harmful facts to the practice set.

5 Conclusion and Future Scope

The upcoming technological advancements are showing us the updated versions of the world around us. We should also be prepared to acknowledge that every topic has advantages and disadvantages. In order to use well-liked technology, we should consider the risks to machine learning models that decrease their efficacy. In order to create ML models strong enough to withstand any vulnerabilities, methods must be researched. This will help to defend widely used devices like IoT and CPS. After different attack types on ML models have been investigated and existing defenses have been put through, this research aims to examine how to safely defend the models from various adversaries while also making them durable.

References:

1. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.
2. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014
3. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597, IEEE, 2016
4. S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in Proceedings of the IEEE conference ,2016
5. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387, IEEE, 2016
6. Reza Shokri Cornell Tech, Marco Stronati∗ INRIA, Congzheng Song Cornell, Vitaly Shmatikov Cornell Tech, “Membership Inference Attacks Against Machine Learning Models” in 2017 IEEE Symposium on Security and Privacy
7. Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent” in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

8. T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” in NIPS MLSec Workshop, 2017
9. Weiwei Hu and Ying Tan , “Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN” arXiv:1702.05983v1 [cs.LG] 20 Feb 2017
10. B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli, “Adversarial malware binaries: Evading deep learning for malware detection in executables,” in IEEE European Signal Processing Conference, 2018, pp. 533–537
11. S. Baluja and I. Fischer, “Learning to attack: Adversarial transformation networks.,” in AAAI, pp. 2687–2695, 2018
12. Masashi KADOGUCHI, Shota HAYASHI, Masaki HASHIMOTO, Akira OTSUKA, “Exploring the Dark Web for Cyber Threat Intelligence using Machine Leaning” in 978-1-7281-2504-6/19/$31.00 ©2019 IEEE
13. Diego G.S.Pivoto a,1 , Luiz F.F.de Almeida a , Rodrigo da Rosa Righi b,1 , Joel J.P.C.Rodriguesc,d , Alexandre BaratellaLugli a , Antonio M.Alberti a,∗ “Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review” in Journal of Manufacturing systems Dec 2020
14. Felix O. Olowononi, Danda B. Rawat, Chunmei Liu “Resilient Machine Learning for Networked Cyber Physical Systems: A Survey for Machine Learning Security to Securing Machine Learning for CPS” in DOI 10.1109/COMST.2020.3036778, IEEE Communications Surveys & Tutorials
15. Mamata Rath, SushrutaMishra ,“Advanced level security in Network and Real Time applications using Machine Learning Approaches” Advanced-Level Security in Network and Real-Time Applications Chapter-Sep-2020
16. Pieter Delobelle1 , Paul Temple 2 , Gilles Perrouin 2 , BenoˆıtFr´enay 2 , Patrick Heymans 2 , and Bettina Berendt 1 , 3 “Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning” in arXiv:2005.06852v2 [cs.LG] 1 Sep 2020
17. Sushant Sharma, PavolZavarsky, Sergey Butakov ,“Machine Learning based Intrusion Detection System for Web-Based Attacks” in 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)
18. Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. 2020. “Certified robustness to label-flipping attacks via randomized smoothing” In International Conference on Machine Learning. PMLR, 8230ś8241
19. Hind Bril El-Haouzi, Etienne Valette, Bettina-Johanna Krings and António Brandão Moniz “Social Dimensions in CPS & IoT Based Automated Production Systems” in Societies 2021
20. Derui (Derek) Wang ,Chaoran Li , Sheng Wen , Surya Nepal , and Yang Xiang “Man-in-the-Middle Attacks Against Machine Learning Classifiers Via Malicious Generative Models” in IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 18, NO. 5, SEPTEMBER/OCTOBER 2021
21. Krishna Yadav1 , B. B. Gupta1,* “Clustering based Rewarding Algorithm to Detect Adversaries in Federated Machine Learning based IoT Environment” in 2021 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-7281-9766-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE50685.2021.9427586
22. Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. 2021. “De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks” IEEE Trans. Inf. Forensics Secur. 16 (2021), 3412ś3425
23. K. Aryal, M. Gupta, and M. Abdelsalam, “A Survey on Adversarial Attacks for Malware Analysis,” arXiv preprint arXiv:2111.08223, 2021.
24. ZHIYI TIAN, LEI CUI, JIE LIANG, SHUI YU,“A Comprehensive Survey on Poisoning Atacks and Countermeasures in Machine Learning” in © 2022 Association for Computing Machinery
25. Ehsan Hallaji, RoozbehRazavi-Far and Mehrdad Saif ,“Federated and Transfer Learning: A Survey on Adversaries and Defense Mechanisms” in arXiv:2207.02337v1 [cs.LG] 5 Jul 2022
26. Florian Tramèr, Reza Shokri, Ayrton San Joaquin et al.,” Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets” in CCS ’22, November 7–11, 2022, Los Angeles, CA, USA

27. Kshitiz Aryal, Maanak Gupta, Mahmoud Abdelsalam, “Analysis of Label-Flip Poisoning Attack on Machine Learning Based Malware Detector” in arXiv:2301.01044v1 [cs.CR] 3 Jan 2023
28. Matthew Rosenblatt 1 , Javid Dadashkarimi 2 , and Dustin Scheinost,” Enhancement attacks in biomedical machine learning” in arXiv:2301.01885v1 [stat.ML] 5 Jan 2023