# Adversarial Machine Learning Problems: Threats and Attacks

Revathi Lavanya Baggam[1*], V. Valli Kumari[2], Gowri Shankar Wuriti[3]

[*]Research Scholar, Andhra University College of Engineering, Waltair
[2]Professor, Andhra University College of Engineering, Waltair
[3]Scientist & Technology Director, ASL, DRDO, Hyderabad

***ABSTRACT***

*Machine Learning is being extensively used in many wide range applications, as it shows cases its technical breakthroughs. It also has proved its success by handling complex scenarios and exhibited its capabilities very close to humans or even much beyond the human ability. But, the research in recent times even prove that machine learning models are being much more helpless to various kinds of attacks, which in turn are compromising its security aspects of the models and their applications. However, these types of attacks are cautious due to their unexpected behavior of deep learning methods. In this survey, a systematic analysis of the security issues of ML majorly focusing on the existing set of attacks on ML systems, and their related defenses or the secure learning approaches and the evaluation techniques. Inspite of keenly focusing on only one type of approach or attack, this paper details about the various phases of machine learning security starting from training phase to the testing phase. Initially, this paper explains about the machine learning in the existence of the adversaries present, and the reasons behind why the machine learning can be easily attacked is analyzed. Later, the aspects with respective to machine learning security are classified into majorly five sets namely: backdoors present in the training set; theft to model; poisoning the training set; sensitive training data recovery; adversarial example attacks. Also, the threat models, various approaches to attack and the defense methods are analyzed.*

***Keywords:*** *Poisoning attacks, backdoor attacks, adversarial examples, Artificial intelligence security*

## 1. INTRODUCTION

In the recent years' machine learning has its wide share in many sectors like driverless cars, natural language processing, smart healthcare, NLP and image classification. In image classification sector the speed and accuracy of the machine learning is extensively exceeding the humans. In some of the applications like malicious program detection, spam filtering for e.g., machine learning enables the security aspects and its capability.

But recent study proves that machine learning algorithms are being easily prone to threat to security. Some of the security threats like:

i. Poisoning the training data results in reduction of the model accuracy or may even lead to many other error-specific attacks.
ii. Risky circumstances for the system can also be brought on by backdoors in the training data.
iii. The adversarial example, which is a well-designed disturbance in the test input, may cause the model to act incorrectly.
iv. Some of the attacks like model stealing attack, membership inference attack and model inversion attack may result in stealing the parameters of the model or even retrieve some very sensitive training data.

However, threat to machine learning security may lead to severe consequences like safety and security of some of the typical applications like smart healthcare, autonomous driving etc.

In recent times there have been a large amount of research are going on safety of deep learning algorithms like:

[1]*Corresponding author : Email: Revathi Lavanya Baggam

Szegedy et al. [1] focused on the threat of hostile examples, whereas the security of machine learning is not a new concept [3]. The previously stated work by Dalvi et al.[4] [5] focused on the adversarial machine learning with respective to non-deep machine learning techniques in relation to PDF malware exposure, intrusion detection etc. [3]. These previously attacks are said to be as the evasion attacks, and some other as poisoning attacks.

Motivated by the above issues, this paper presents the comprehensive survey about the security of machine learning.

In 2010, Barreno et al. [6] examined the evasion techniques against non-deep learning systems and concentrated on spam filter. Akthar et. al. [7] surveyed adversaries in the field of computer vision. Yuan et al. [8] reviewed about adversaries in deep learning and they discussed the various adversary example generation methods and discussed its countermeasure too. Riazi et.al [9] discussed regarding privacy protection techniques with respective to cryptographic primitives. All the above reviews mainly focus on only adversarial attack. Biggio et.al [3] majorly focused into the evasion and poison attacks, as they reviewed the wild patterns in adversarial machine learning over a decade which includes the earlier non-deep and recent develop deep learning techniques in computer vision and cyber security sectors. Lie et al. [10] surveyed about the assessment and data security. Papernot et al. [11] have described three security aspects that are availability, honesty, and confidentiality, and also they discussed the defenses with respect to the robustness, accountability and privacy [11].

**The major difference of this paper to the previous survey papers is this paper:**

1. Properly explains about the entire machine learning security starting restraining phase to the testing phase, including entire categories of attacks.
2. Majorly focuses on five sets of attacks namely: backdoors present in the training set; theft to model; poisoning the training set; sensitive training data recovery; adversarial example attacks.
3. Direction to future research on machine learning security is presented which includes the privacy preserving machine learning methods, along with attacks in real physical conditions.

The paper is organized as follows: Section II contains the behavior of machine learning in the presence of adversary. In Section III the threat models along with the approach of attack is reviewed. Section III discusses about the evaluation techniques. The future direction on security of machine learning is discussed in Section IV. Section V concludes the paper.

## 2. Behaviour of Machine Learning Model in Presence of Adversary

### 2.1. Overview of machine learning

There are mainly two stages of a machine learning system: (1) training phase, which helps the algorithm to learn from the training data to develop a model with the parameters (2) Testing phase. The trained model is then applied to a specific assignment, like classification to provide predicted level value for the input data. The machine learning algorithms have been broadly divided into a set of two division: (a) N algorithms (b) non-NN algorithms. The abbreviation NN is used to depict the Deep Neural Networks (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Neural Network (NN) algorithms have found to be majority of contribution in recent years by increasing the performance of machine learning systems. Non-NN algorithms represent the machine learning algorithms like SVM (Support Vector Machine), Naïve-Bayes, etc. A machine learning algorithm usually consists of the

following attributes, data holder, system and the clients. But in the case of adversarial model where attackers too are present.

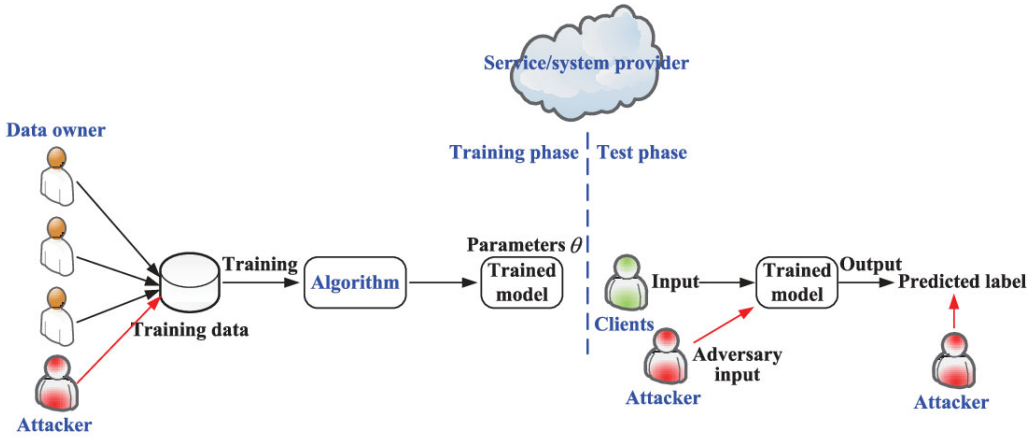


Figure 1: Overview of machine learning systems, which illustrates the two phases, the learning algorithms, and different entities

From the diagram, the data owners are the owners of the huge training data which is private. The service provider builds the algorithm, trains the model, and then uses it to carry out a job or offer a service. The clientele make advantage of the services. An attacker in this scenario may be a system user, either internal or foreign, who has a strong desire to learn more about the other entities.

### 2.2. WHY IS MACHINE LEARNING ATTACKED?

In training phase, huge training data along with its computational complexity; the training procedure is leading to: 1) to outsource the training procedure [12]; 2) the third parties lending the pre-trained models which are the intellectual properties (IP), are integrated; 3) a huge chunk of data that lack any data validations and are obtained from unreliable third parties. These are bring new security threats to system. Also the machine-earning-as a-service are used intensively , where the present model the machine learning model which is works on the server/cloud, as clients are able to analyzed through prediction of APIs. This huge data is of great commercial and confidential too, which is the major target for attackers. Good fellow et al. [2] pointed linear DNN in high dimensional space causes adversarial attack which is being vulnerable too. Studies [1], [8] incomplete training data is one of the reasons these adversarial example attacks. Yeom et al. [13] studied that overfitting is one of the reasons for the attacker to implement membership inference attacks.

## 3. ATTACKS ON ML

The security threats are broadly classified into five: 1) Backdoors present in the training set; 2) Theft to model; 3) Poisoning the training set; 4) Sensitive training data recovery; 5) adversarial example attacks

ISBN - 978-93-95038-50-8

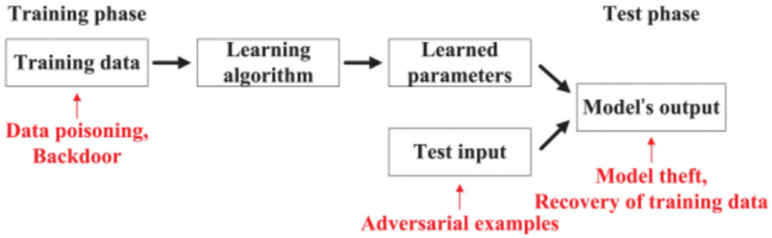


Figure 2: Attacks on Machine Learning

### 3.1. BACKDOORS PRESENT IN TRAINING SET

The approaches suggested by the previous surveys is as follows:

BadNet [12]; PLMs [32]; Attack on deep learning [33]; Attack on CNNs [34] [35]; Attack on federated learning [36]. These approaches follow a working mechanism of adding well-designed signals of backdoor to the clean data and also to manipulate the parameters of the model directly. Due to this the target model is able to misclassify the backdoor instances. The advantage of these approaches is that they are stealthy and do not affect the model's performance, but they need to participate in the process of training or they need to re-train the model which is the major drawback of this.

The below diagram explains how an attacker can hide the backdoor in the training data or in a pre-trained model.

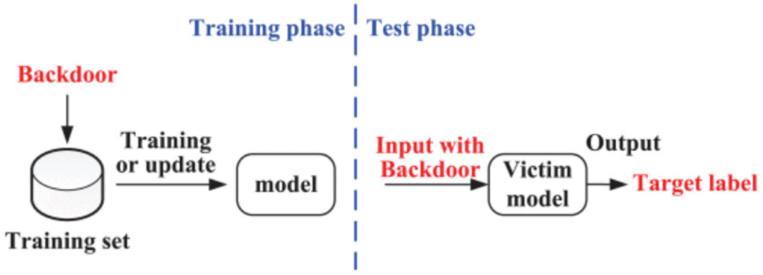


Figure 3: Backdoor attacks

### 3.2. MODEL THEFT

The recent research has proved that by observing the output label and confidence levels the adversaries can steal the machine learning models, which is also known as model extraction attack or model stealing attack.

Some of the researchers have surveyed the attacks and the approaches are as follows:

1. Tramer et al. [63]; Shi et al. [64] and Chandrasekharan et al. [65] have proposed an approach that uses the working mechanism of extraction through prediction APIs and also to train a model which is similar to that of the target model, and to use an active learning technique [65]. This is effective in generating a model which is very similar to that of the target model and so helpful in extracting the model parameters without the knowledge of the target model. But the target model needs to be queried very frequently which is a drawback of this approach.
2. Wang and Gong [66] have suggested to calculate the hyperparamters by solving the linear equations. This is effective in to steal the hyperparameters of the models and is also suitable to be applied to various machine learning algorithms, but this requires the information about the training set and the target algorithm too.

ISBN - 978-93-95038-50-8

3. Milli et al. [67] proposed that the query gradients of a model to be chosen inputs and the heuristic model reconstruction, which is effective to reconstruct the target model. This is more efficient when compared to querying the labels, but this incurs too much of computational cost and this only capable of evaluating on a two-layer NN.

### 3.3. POISONING THE TRAINING SET

Some of the researchers have proposed some of the training set poisoning approaches which have been broadly classified as below:

Some of the researchers have proposed some of the training set poisoning approaches which have been broadly classified as below:

*Targeting non-NN models*

A. Anomaly detection or security detection application: This target system or model uses the approach of poisoning PC-subspace based anomalies detection [14] as well as chronic poisoning against IDS [15]. This model follows the poisoning the training data of the model working mechanism with an effect to reduce the performance of the anomaly detection algorithms. The advantages of this target system is this is very simple and effective on several training models [15]. This target system too has some disadvantages which is that it is non-generic [14] and it is long-term slow poisoning and complicated to implement.

B. Biometric recognition systems: This target system follows the approach of poisoning PCA based face recognition systems [16], and poisoning face templates [17]. This target system follows the mechanism of submitting a set of fake faces and claiming to be the victim. The effect of this type of target system is to replace the victim's template in the system. The advantage of this model is that it is stealthily and gradually. This model requires the information about the target system or about the victim.

C. SVM: This model follows the approach of poisoning the SVM approach and uses the strategy of gradient ascent which is based about the optimal solution of SVM. This has an effect of misleading the clustering algorithms. The main advantage of this model is, it uses the optimized formulation and it can also be kernalized too, but it needs the full information/knowledge about the algorithm and also about the training too which is the major drawback of this approach.

D. Clustering algorithms: This model follows the approach of poisoning the clustering data [19] and also poisoning the behavioral of malware clustering [20]. Mainly the poisoning attack is based out of the distance between the clusters which is its working mechanism. This showcases the effect of misleading the clustering algorithms. This is very generic to the clustering algorithms which is its major advantage, of course it requires the whole sum information about the target algorithm which is this model's drawback.

E. Machine learning algorithms or processing methods: The model follows the approach of poisoning the feature selection [21], as well poisoning the collaborative filtering [22] and also poisoning the regression learning [23]. This model follows the optimized based or statistical based poisoning attack approach. It misleads the feature selection algorithms, as well the collaborative filtering system or the linear regression one, but it is very powerful and effective that adds as an advantage to this model. But this is non-generic and hugely requires the information/data about the target system.

F. Online learning or data sanitization defenses: There are two proposed approaches in this model:

- The first approach follows to poison the breaks data sanitization defenses [24], and works on optimizing the poisoning attacks. This approach is effective in bypassing the data sanitization defenses and it is strong and effective under the defenses which is an advantage of this approach. This incurs too much of overhead on computations.
- The second approach follows poisoning the online data [25] [26] and it follows the mechanism of optimizing the poisoning attacks likewise the first approach. This mainly effects the performance of online learning in reducing the same, helps to work for the online learning scenarios which is its advantage. This approach too incurs too much of overhead on computations like the first one.

*Targeting NN models:*

The below approaches have been proposed:

- Generative poisoning against NN [27] follows the working mechanism of poisoning attack based out of Generative Adversarial Networks(GAN). This approach effects the accuracy of the NN model by reducing the same. This majorly generates the poisoning data quickly, but it requires too many interactions with the model very frequently.
- The second approach is poisoning of deep learning algorithms by optimizing the back gradient [28]. This follows the mechanism of optimizing the poison attacks, which effects the healthcare systems, recommender systems as well the crowd sensing systems by misleading them. We will be able to apply the poisoning attacks towards the practical applications and scenarios which is these approaches major advantage, but this requires too much of information with respective to target system.

*Specific applications oriented:*

This majorly follows the approach of poisoning the healthcare systems [29] as well poisoning the graph based recommendation system [30] and also poisoning the crowd sensing systems [31]. The approaches follow the functionality of optimizing the poisoning attacks and effectively misleads the systems of healthcare, recommender and crowd sensing. These helpful in applying the poisoning attacks to most of the practical applications as well as scenarios, but this too needs too much of information/knowledge with related to the target system.

### 3.4. ADVERSARIAL EXAMPLE ATTACKS

Researchers have proposed few of the below approaches to the adversarial example attacks:

*Earlier evasion attacks on non-NN models:*

- Targets like game theory can be handled with the approach of classifying adversarial [4] by the mechanism of modifying the features which show a great impact on the detector, due to this approach on this type of targets the malicious instances cannot be detected by the detector. This greatly evades the detection of security related applications, but this needs the information of feature extraction algorithm.
- The approach of adversarial learning [40] through reverse engineering the classifier follows the working approach of modifying the features which are having the greatest impact on the detector, due to wchih the malicious instances cannot be detected by the detector. This avoids in detecting pf the security related applications, but this needs great knowledge about the feature extraction algorithm.
- The approach suggests by Nelson et.al [5] for spam filtering follows the mechanism of modifying the features which are having the greatest impact on the detector, because of which instances which are malicious cannot be dected by the detector. This needs too much of

ISBN - 978-93-95038-50-8

information about the feature xtraction algorithm, but helps in evading the detection of the security related applications.

- Stochastic search –based attack [44] and the Evade-HC [45] approach for malware detection(black-box) works using the heuristic search-based method as well as the stochastic search- based method. Eventhough the malicious instances cannot be detected by the detector, but this requires too many iterations and feedback from the system. This is very generic and does not requires any information about the systems.
- The approach to attacka linear classifier [43] which targets the android malware detection follows the mechanism of modifying the features that have the great impact on the detector which is effective in not detecting the malicious instances. This completely evades the concept of detecting the security related applications but requires data about the algorithm of feature extraction.

*Adversarial example attacks on NN models:* The approaches of Intriguing properties of NN[1], FGSM[2], BIM[3], JSMA[47] and One-pixel attack[48], Carlini and Wagner[49][50], Universal perturbations[51], Malware classification[52], Face recognition[53] which are targeted to first work of adversarial examples on NN, one-step/one-shot methods, take multiple small steps iteratively, perturbing only a few pixels, high-confidence adversarial examples, universal adversarial perturbations, targeting deep learning based security detection and inbiometric authentication systems follow the mechanism of optimization-based or gradient-based adversarial example attacks. This is effective in misleading the model and also this cannot be completely perceived by human-being. This do not require any information about the feature extraction algorithm of the system and also the perturbations are very small and imperceptible too. But these attacks may have a chance of failing in the real world.

In real world conditions the approaches like road sign recognition attack [54], cellphone camera attack [46], face recognition attack [46],3D objects attack [55] follow the method of simulating the physical properties of the adversarial examples generation. These are effective in attacking the physical conditions successfully and are very robust to any physical condition. But the perturbations are high and conspicuous.

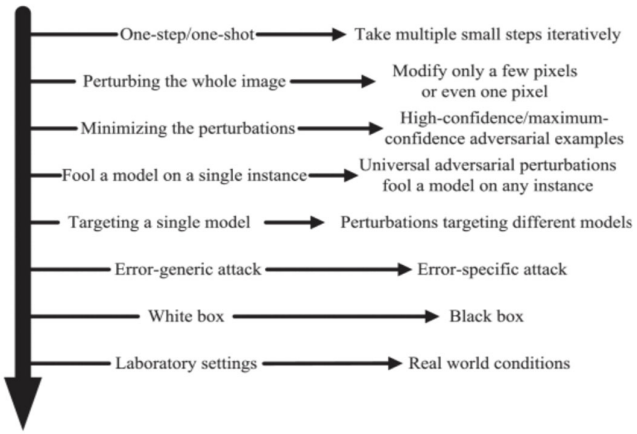


Figure 4: Pipeline of adversarial example attacks in deep learning algorithms

## 4. FUTURE DIRECTION OF SECURITY THREAT

There are have been numerous of security attacks on machine learning algorithms but the effect of these attacks under real physical conditions are the current research topics

ISBN - 978-93-95038-50-8

Also in recent years the machine learning algorithms privacy has gained much focus, so the deep learning needs to answer the issue of privacy protection, along with this the cryptographic primitive – based ML approaches efficiency too needs to be improved.

## 5. CONCLUSION

Current manuscript provides information about the proposed approaches by various researchers about the attacks on machine learning algorithms. Machine learning algorithms facing a wide variety of security threats all though its phases, as new security threats are continuously emerging. Studies prove that adversarial and poisoning examples can be generalized among various different machine learning models. This transferability can be helpful for the attacks to be successful in black-box scenarios.

## REFERENCES

1. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, ‘‘Intriguing properties of neural networks,’’ in Proc. 2nd Int. Conf. Learn. Represent., Apr. 2014, pp. 1–10.
2. I. J. Goodfellow, J. Shlens, and C. Szegedy, ‘‘Explaining and harnessing adversarial examples,’’ in Proc. Int. Conf. Learn. Representations, Mar. 2015, pp. 1–11.
3. B. Biggio and F. Roli, ‘‘Wild patterns: Ten years after the rise of adversarial machine learning,’’ Pattern Recognit., vol. 84, pp. 317–331, Dec. 2018.
4. N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, and D. Verma, ‘‘Adversarial classification,’’ in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Aug. 2004, pp. 99–108.
5. B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar, and K. Xia, ‘‘Exploiting machine learning to subvert your spam filter,’’ in Proc. USENIX Workshop Large-Scale Exploit. Emerg. Threat., Apr. 2008, pp. 1–9.
6. M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, ‘‘The security of machine learning,’’ Mach. Learn., vol. 81, no. 2, pp. 121–148, Nov. 2010.
7. N. Akhtar and A. S. Mian, ‘‘Threat of adversarial attacks on deep learning in computer vision: A survey,’’ IEEE Access, vol. 6, pp. 14410–14430, Jul. 2018.
8. X. Yuan, P. He, Q. Zhu, and X. Li, ‘‘Adversarial examples: Attacks and defenses for deep learning,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2805–2824, Sep. 2019.
9. M. S. Riazi and F. Koushanfar, ‘‘Privacy-preserving deep learning and inference,’’ in Proc. Int. Conf. Comput.-Aided Design ICCAD, Nov. 2018, pp. 1–4.
10. Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, ‘‘A survey on security threats and defensive techniques of machine learning: A data driven view,’’ IEEE Access, vol. 6, pp. 12103–12117, 2018.
11. N. Papernot, P. D. McDaniel, A. Sinha, and M. P. Wellman, ‘‘SoK: Security and privacy in machine learning,’’ in Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P), Apr. 2018, pp. 399–414.
12. T. Gu, B. Dolan-Gavitt, and S. Garg, ‘‘BadNets: Identifying vulnerabilities in the machine learning model supply chain,’’ 2017, arXiv:1708.06733. [Online]. Available: http://arxiv.org/abs/1708.06733
13. S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, ‘‘Privacy risk in machine learning: Analyzing the connection to overfitting,’’ in Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF), Jul. 2018, pp. 268–282.
14. B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S. Lau, S. Rao, N. Taft, and J. D. Tygar, ‘‘ANTIDOTE: Understanding and defending against poisoning of anomaly detectors,’’ in Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. (IMC), Nov. 2009, pp. 1–14.
15. P. Li, Q. Liu, W. Zhao, D. Wang, and S. Wang, ‘‘Chronic poisoning against machine learning based IDSs using edge pattern detection,’’ in Proc. IEEE Int. Conf. Commun. (ICC), May 2018, pp. 1–7.
16. B. Biggio, G. Fumera, F. Roli, and L. Didaci, ‘‘Poisoning adaptive bio-metric systems,’’ in Proc. Int. Workshops Stat. Tech. Pattern Recognit. Struct. Synt. Pattern Recognit., Nov. 2012, pp. 417–425.
17. B. Biggio, L. Didaci, G. Fumera, and F. Roli, ‘‘Poisoning attacks to com-promise face templates,’’ in Proc. Int. Conf. Biometrics (ICB), Jun. 2013, pp. 1–7.

18. B. Biggio, B. Nelson, and P. Laskov, ''Poisoning attacks against support vector machines,'' in Proc. 29th Int. Conf. Mach. Learn., Jun. 2012, pp. 1467–1474.
19. B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli, ''Is data clustering in adversarial settings secure?'' in Proc. ACM Workshop Artif. Intell. Secur. AISec, Nov. 2013, pp. 87–98.
20. B. Biggio, K. Rieck, D. Ariu, C. Wressnegger, I. Corona, G. Giacinto, and F. Roli, ''Poisoning behavioral malware clustering,'' in Proc. Workshop Artif. Intell. Secur. Workshop AISec, Nov. 2014, pp. 27–36.
21. H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, ''Is feature selection secure against training data poisoning,'' in Proc. 32nd Int. Conf. Mach. Learn., Jul. 2015, pp. 1689–1698.
22. B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, ''Data poisoning attacks on factorization-based collaborative filtering,'' in Proc. Annu Conf. Neural Inf. Proc. Syst., Dec. 2016, pp. 1885–1893.
23. M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, ''Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2018, pp. 19–35. P. W. Koh, J. Steinhardt, and P. Liang, ''Stronger data poisoning attacks break data sanitization defenses,'' 2018, arXiv:1811.00741. [Online]. Available: http://arxiv.org/abs/1811.00741
24. Y. Wang and K. Chaudhuri, ''Data poisoning attacks against online learning,'' 2018, arXiv:1808.08994. [Online]. Available: http://arxiv.org/abs/1808.08994
25. X. Zhang, X. Zhu, and L. Lessard, ''Online data poisoning attack,'' 2019, arXiv:1903.01666. [Online]. Available: http://arxiv.org/abs/1903.01666
26. C. Yang, Q. Wu, H. Li, and Y. Chen, ''Generative poisoning attack method against neural networks,'' 2017, arXiv:1703.01340. [Online]. Available: http://arxiv.org/abs/1703.01340
27. L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, ''Towards poisoning of deep learning algorithms with back-gradient optimization,'' in Proc. 10th ACM Workshop Artif. Intell. Secur. AISec, 2017, pp. 27–38.
28. M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, ''Systematic poisoning attacks on and defenses for machine learning in healthcare,'' IEEE J. Biomed. Health Informat., vol. 19, no. 6, pp. 1893–1905, Nov. 2015.
29. M. Fang, G. Yang, N. Z. Gong, and J. Liu, ''Poisoning attacks to graph-based recommender systems,'' in Proc. 34th Annu. Comput. Secur. Appl. Conf., Dec. 2018, pp. 381–392.
30. C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, and L. Su, ''Towards data poisoning attacks in crowd sensing systems,'' in Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. - Mobihoc, Jun. 2018, pp. 111–120.
31. Y. Ji, X. Zhang, and T. Wang, ''Backdoor attacks against learning sys-tems,'' in Proc. IEEE Conf. Commun. Netw. Secur. (CNS), Oct. 2017, pp. 1–9.
32. X. Chen, C. Liu, B. Li, K. Lu, and D. Song, ''Targeted backdoor attacks on deep learning systems using data poisoning,'' 2017, arXiv:1712.05526. [Online]. Available: http://arxiv.org/abs/1712.05526
33. C. Liao, H. Zhong, A. C. Squicciarini, S. Zhu, and D. J. Miller, ''Backdoor embedding in convolutional neural network models via invisible pertur-bation,'' 2018, arXiv:1808.10307. [Online]. Available: http://arxiv.org/abs/1808.10307 VOLUME 8, 2020 74739
34. A. Kurakin, I. J. Goodfellow, and S. Bengio, ''Adversarial examples in the physical world,'' in Proc. Int. Conf. Learn. Represent. (ICLR) Workshop, Feb. 2017, pp. 1–14.
35. N. Carlini and D. A. Wagner, ''Towards evaluating the robustness of neural networks,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2017, pp. 39–57.
36. N. Carlini and D. A. Wagner, ''Adversarial examples are not easily detected: Bypassing ten detection methods,'' in Proc. 10th ACM Workshop Artif. Intell. Secur. (AISec), 2017, pp. 3–14.
37. S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, ''Universal adversarial perturbations,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 86–94.
38. K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. D. McDaniel, ''Adversarial examples for malware detection,'' in Proc. 22nd Eur. Symp. Res. Comput. Secur., Sep. 2017, pp. 62–79.
39. M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, ''Accessorize to a crime: Real and stealthy attacks on State-of-the-Art face recognition,'' in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. CCS, 2016, pp. 1528–1540.

40. K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, ‘‘Robust physical-world attacks on deep learning visual classification,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1625–1634.
41. A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, ‘‘Synthesizing robust adversarial examples,’’ in Proc. 35th Int. Conf. Mach. Learn., Jul. 2018, pp. 284–293