# Effectiveness Evaluation of Black-Box Data Poisoning Attack on Machine Learning Models

Junjing Zhan[1,*], Zhongxing Zhang[2], Ke Zhou[3]

[1] School of Computing, Beijing Institute of Technology, Zhuhai, Guangdong, 519088, China
[2] Software College, Taiyuan University of Technology, Jinzhong, Shanxi, 030600, China
[3] School of Software, Tianjin Chengjian University, Xiqing District, Tianjin, 300192, China
*Corresponding Author. Email: 210021101610@bitzh.edu.cn

**Abstract.** With machine learning has been widely used in face recognition, natural speech processing, automatic driving and medical systems, attacks against machine learning are also accompanied, which may bring serious safety risks to biometric certification systems or automobiles. Incorrect classification of malicious parking signs. Therefore, the security and privacy of machine learning become more and more prominent with its application. Data poisoning attack targets on machine learning models, which contaminates the data and makes machine learning get wrong results, thus bringing potential safety hazards. In this paper, a poison attack strategy for black box machine learning model is adopted to carry out black box attack. In the experiment, the data poisoning attack on the machine learning model is successfully carried out. Indicates that an attacker has successfully resulted in targeted misclassification of samples. The purpose of this paper is to explore the security threats that may exist in existing machine learning algorithms, and further study the defense measures to improve the security of algorithms and prevent malicious users and attackers from tampering with the training data and input samples of the model or stealing the model parameters, resulting in the damage to the confidentiality, usability and integrity of the model.

**Keywords:** Poisoning Attack, Machine Learning, Black Box Attack.

## 1 Introduction

Machine learning has been widely used in face recognition, natural speech processing, automatic driving and medical system by imitating the learning behavior of human class and obtaining new knowledge with the strong computing ability of computers [1]. However, the attack against machine learning is also accompanied, and the security and privacy of machine learning are increasingly prominent with the popularization of its application [2]. Attacks against machine learning can be divided into three categories: data poisoning, counteracting input, and model stealing. Data poisoning occurs during data training stage. Attackers attack the data by modifying the original training dataset or injecting contaminated data into the original training dataset through certain strategies, which makes the classification boundary of machine learning classifier shift

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_68

or change, so that machine learning produces wrong output results and causes potential safety hazards.

Data Poisoning attack has not only received much attention in academia, but also brought serious harm in industry [3,4]. If machine learning models learns from data of potentially implausible sources (e.g., Twitter, etc.), attackers can easily manipulate the training data distribution by appending targeted modified samples to dataset for reducing model performance and changing model behavior. Such as the Microsoft Tay, which is a chatbot for Twitter user communication, was closed only 16 hours later because it began to make racism-related comments after being attacked by poisoning [5]. This attack forced us to rethink the security of machine learning models. In conclusion, the attacker affects the model learning process by tampering with the training data or adding malicious data, and finally reduces its performance during inference stage.

The purpose and significance of this research is to explore the security threats that existing machine learning algorithms may have, and further study the defense measures to improve the security of machine learning algorithms and prevent attackers from malicious tampering with the training data and input samples of the model or stealing the model parameters, resulting in the damage to the confidentiality, usability and integrity of the model.

## 2 Related Work

In previous work, poison samples are constructed based on machine algorithms, and a poison strategy based on black box machine learning model is further proposed. In this way, the security threat and influence of machine learning algorithm are studied. This paper also proposes a sample validity evaluation method to improve the robustness of machine learning algorithms against poison attacks. In another article, the detection of data poisoning attack algorithm for campus face recognition was studied, and the article pointed out the harm brought by data poisoning, such as the possibility of making the face recognition system misjudge the identity, making the illegal identity personnel enter the campus, which brings security risks. In order to solve data poisoning attacks, Kang Fei et al from Central South University comprehensively applied two indicators of data cleaning and improving algorithm robustness to solve the problem of data poisoning attacks, and proposed a poison data detection method based on data complexity, which can effectively detect poison data, but the data model of this method is complex and inefficient [6].

Another work takes the data security of deep reinforcement learning model as the starting point, and deeply discusses its security and privacy issues, especially the potential threats to the model when the data is maliciously tampered with. According to the characteristics of reinforcement learning, a method of state poisoning is proposed to destroy the model by switching neighboring states. In order to verify the effectiveness of the state poisoning method, the adversarial sample method was first used to poison the original state diagram, aiming at disrupting the model learning. However, the experimental results show that this method of poisoning has limited effect on the model without changing the core features of the state diagram. Then, referring to the previous reward poisoning method, this paper improves the anti-sample poisoning strategy to swap the adjacent states of the model. Experimental results show that this switching state poisoning method can cause large damage to the target model at a lower rate of poisoning. Finally, to solve the problem of state data poisoning, a Deep Q-Network (DQN) model is proposed, which can automatically identify and clean the tampered state data [7]. By comparing the performance of original DQN and dc-DQN under different poisoning methods and poisoning rates, it is found that dc-DQN model has stronger robustness to state poisoning. In summary, this paper deeply studies the security of the deep reinforcement learning model, and puts forward an effective poison defense strategy, which provides a strong support for the security application of reinforcement learning [8,9].

Moreover, with the wide application of automatic driving, access control security and face payment, the security problem of machine learning has gradually become a new research hotspot. Attacks on machine learning approaches could be roughly separated into poison attacks and adversarial attacks based on the attack phase, the difference is that the former attack occurs in the training phase, the latter attack occurs in the test phase. This article reviews the methods of poisoning attacks in machine learning, reviews the poisoning attacks in deep learning, analyzes the possibility of such attacks, and studies the existing defenses against these attacks. Finally, the future research direction of poisoning attack is discussed. The main research is the data backdoor poisoning attack in federation learning. A large number of experiments in this paper show that under data poisoning, even if the percentage of malicious clients is small, the attack effect is obvious, and with the increase of the percentage of malicious participants, the overall model test accuracy decreases. Even if m is small, the work observes a decrease in model accuracy compared to the non-poisoned model, and an even greater decrease in source-class accuracy. An attacker who controls even a small portion of a client has the ability to have a significant impact on global model effects. If the attacker continuously participates in the attack, it has a great impact on the next round. On the contrary, if the attacker does not attack for several rounds, the model may recover significantly. Defense results: plot results show that malicious actors' updates belong to a significantly different cluster than honest actors' updates form their own cluster, that malicious models are clearly identified even if there are fewer attackers, and that the defense is not affected by the "gradient drift" problem [10].

# 3 Method

## 3.1 Data Poisoning Attack

Poisoning attacks are when malicious actors attack training data sets to manipulate the predictions of machine learning models during training or retraining. There are two common methods of data poisoning: one is to tamper with the data in the existing data set, and the other is to subtly inject destructive toxic data into the data set. If a machine learning learns from data collected from untrusted sources (such as Twitter, etc.), an attacker can disturb the distribution of training data by adding poisoned samples for training to degrade model performance and change model behavior.

### 3.2 Multi-Layer Perceptron (MLP)

The working principle of MLP is based on a fully connected feedforward neural model. The MLP has an input layer, one or more hidden layers, and an output layer. These components are made up of multiple neurons. Data sequentially goes through these operations, where the input layer usually does not use activation functions, while the hidden layer and the output layer use activation functions, such as Sigmoid, rectified linear unit (ReLU), or Tanh, which introduce nonlinearities into the network, allowing the MLP to learn and process complex nonlinear relationships. Each neuron of the hidden and the output layer is connected to the previous layer by weight and bias, and the output is usually processed by the softmax function to accommodate multi-class classification problems.

The training process for MLP typically uses a backpropagation algorithm that calculates the gradient of the loss function to the network parameters, and updates the parameters according to those gradients to minimize the loss function, a process that involves randomly initializing all parameters (weights and biases) and then optimizing them through a process of iterative training, calculating the gradients, and updating the parameters.

## 4 Experiments and Results

### 4.1 Experimental Setup

The first is sample generation, which generates 200 samples, 2 features are followed by. The first 100 samples will be used to train the model for the 2 features, and the last 100 samples will be used to visually express whether the model is well trained. Next, data poisoning is carried out, that is, malicious data is added to affect the model training process, and ultimately degrade its performance during prediction stage. This work simulates an attacker to dynamically and incrementally attack the target model. During learning, existing classifier incrementally updates by fitting part of the model to the new 5 points, and finally by repeating continuously, the decision boundary is offset.

### 4.2 Experimental Results

The contour map of the decision function of the MLP classifier is shown in Fig. 1, It could be observed the classifier basically fits the data set. The red dots are basically in the red-shaded part, and the blue dots are basically in the blue-shaded part

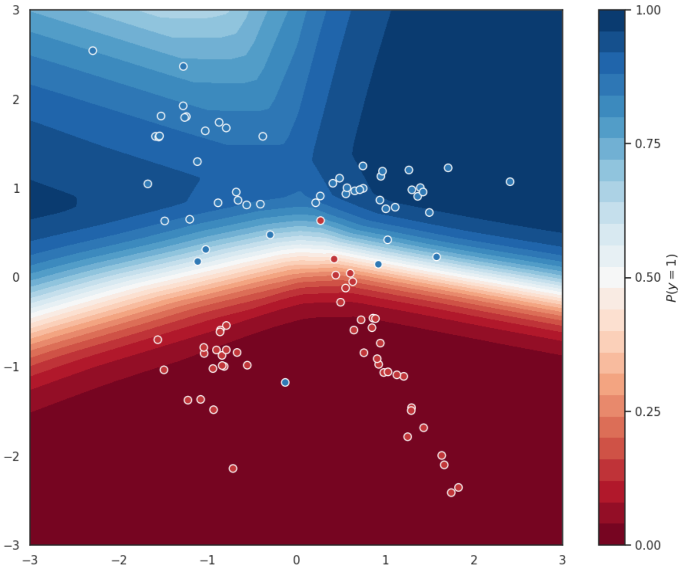

**Fig. 1.** Decision boundary of MLP classifier.

Here this work uses 0.5 as the confidence threshold to form the decision boundary, i.e., if the classifier predicts P(y=1)>0.5, then predict y=1; otherwise, predict y=0. It could be observed from Fig. 2 that the 5 points added at the moment (represented by a five-pointed star), in the space of y=1(y=1is represented by a hollow circle, and y=0 is represented by a solid circle).

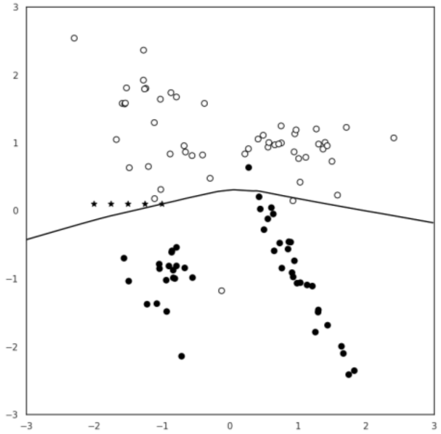

**Fig. 2.** Classification Results with 5 poisoning data.

In order to present a dynamic effect, attack means is reused iteratively, with more and more new boundary offsets. Results are shown in Fig. 3. Repeat 1 time, 5 times, 10 times, and 15 times respectively to get the four pictures. Through this offset, the parts that should have been classified as y=0 that were originally located between the two decision boundaries y=0 will now be classified as y=1. Note that hollow points near coordinates (1,0) should be classified as y=1 under the original decision boundary, but under the current decision boundary, they are classified as y=0. In this way, this work successfully attacked the machine learning model through data poisoning. It shows that the attacker has successfully caused the sample to be misclassified in a targeted manner.

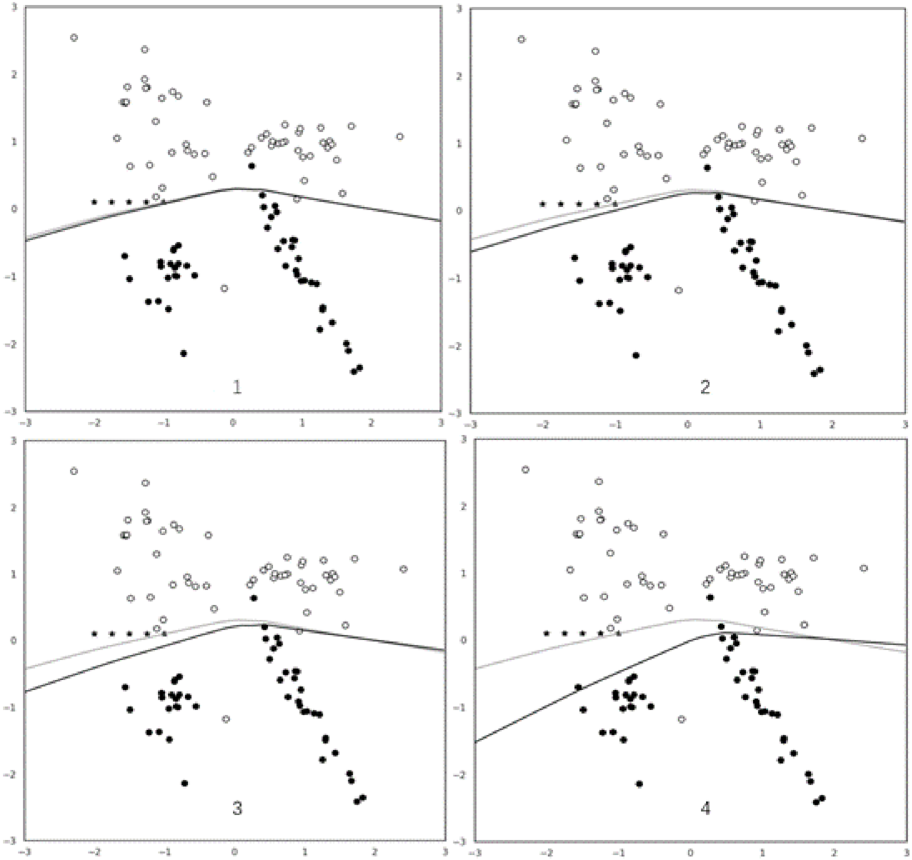

**Fig. 3.** Poisoning attack results with 1, 5, 10, 15 times respectively.

Through the comparison before and after the attack, the changes in the results before and after the attack can be quantified, as shown in Table 1.

**Table 1.** Performance under different times of attack.

| Number of attacks | Attack result |
|---|---|
| 1 time | 0% |
| 5 times | 0.21% |
| 10 times | 0.73% |
| 15 times | 30% |

# 5 Discussion

## 5.1 The Effectiveness of Poisoning Attacks.

The effectiveness of machine learning poisoning attacks can be considered from the aspect of the success rate of the attack. The purpose of poisoning attacks is to make the model produce inaccurate outcome in practical applications. Attackers can modify the training data in a targeted manner to make the model misjudge under specific circumstances. According to the above experiments, the number of attacks is positively correlated with the success rate of the attack result, which shows that the poisoning attack was successfully implemented in this experiment, causing the system to misjudge.

## 5.2 Applicable Occasions

Poisoning attack could be applied in the following occasions. Firstly, in the field of network security: poisoning attacks can be used in network attacks Domain Name System (DNS) poisoning, that is, by tampering the cache of the DNS server to redirect user requests to malicious websites, thereby carrying out illegal activities such as phishing and the spread of malicious software. Secondly, in Internet of Things: Poisoning attacks can be used to interfere with the normal operation of Internet of Things devices. For example, an attacker can tamper with sensor data to cause the monitoring system to misjudge or fail, resulting in security risks. Thirdly, in the financial field: poisoning attacks can be used for financial fraud. Attackers tamper with transaction data or account information to commit illegal transfers, steal funds, etc. Fourthly, in the field of social media: Poisoning attacks can be used for the dissemination of false information, such as by manipulating the algorithms of social media platforms to push false information to users, influence public opinion, or obtain illegal benefits.

## 5.3 Defend Against Poisoning Attacks

Several measurements could be applied to defend poisoning attacks. Firstly, data preprocessing and cleaning: Before using machine learning algorithms, the training data is preprocessed and cleaned to remove outlier values, noise, and inconsistent data. This can reduce the tampering and manipulation of data by attackers. Secondly, data verification and monitoring: establish a data verification and monitoring mechanism to monitor and verify the input data in real time. An anomaly detection algorithm or a rule engine could be used to detect abnormal data and take appropriate measures in a timely manner. Thirdly, multi-model integration: Use multiple independent models for integration. By voting or weighting the results of multiple models, the impact of a single model being attacked can be reduced. An attacker needs to tamper with the training data of multiple models at the same time to successfully deceive the entire system.

### 5.4 Future Prospects for Poisoning Attacks

The further development of adversarial attack technology: With the widespread application of machine learning models, adversarial attack technology will continue to evolve and improve. Attackers may develop more concealed and efficient attack methods to evade current defense mechanisms.

Improvement of laws and policies: As the threat of poisoning attacks becomes more and more serious, relevant laws and policies will continue to be improved. Governments and organizations will strengthen the supervision of machine learning systems and formulate corresponding norms and standards to protect users and society from poisoning attacks.

## 6 Conclusion

Data poisoning is an attack method for machine learning. It intervenes deep learning training data sets, such as inserting or modifying some training samples, so as to degrade the performance of the model or realize the directional or undirectional output of specific inputs, thus bringing serious potential safety hazards. Through the poisoning attack strategy for black box machine learning model, the purpose of poisoning attack is to make the model produce wrong prediction results in practical application. Attackers can modify the training data in a targeted manner, so that the model produces misjudgment under specific circumstances. In this paper, many samples are generated by experimental design, and the samples are poisoned, that is, malicious data are added to affect the model training process. By simulating attacker dynamics, incremental attack machine learning model, many repetitions. Eventually causing the decision boundary to shift. This proves that it has succeeded in causing samples to be misclassified.

## Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

## References

1. Mahesh, B.: Machine learning algorithms-a review. International Journal of Science and Research. **9**(1), 381-386 (2020).
2. Rigaki, M., & Garcia, S.: A survey of privacy attacks in machine learning. ACM Computing Surveys, **56**(4), 1-34 (2023).
3. Ahmed, I. M., & Kashmoola, M. Y.: Threats on machine learning technique by data poisoning attack: A survey. In Advances in Cyber Security: Third International Conference, pp. 586-600 Springer Singapore (2021).
4. Yerlikaya, F. A., & Bahtiyar, Ş.: Data poisoning attacks against machine learning algorithms. Expert Systems with Applications, 208, 118101 (2022).
5. Wolf, M. J., Miller, K., & Grodzinsky, F. S.: Why we should have seen that coming: comments on Microsoft's tay experiment, and wider implications. Acm Sigcas Computers and Society, **47**(3), 54-64 (2017).
6. Kang F., Li J.: Method for detecting poisoning data based on data complexity. Application Research of Computers, **37**(7), 2140-2143 (2020).
7. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D.: Human-level control through deep reinforcement learning. nature, **518**(7540), 529-533 (2015).
8. Wang, L., Javed, Z., Wu, X., Guo, W., Xing, X., & Song, D.: Backdoorl: Backdoor attack against competitive reinforcement learning. arXiv preprint arXiv:2105.00579 (2021).
9. Li, M., Sun, Y., Lu, H., Maharjan, S., & Tian, Z.: Deep reinforcement learning for partially observable data poisoning attack in crowdsensing systems. IEEE Internet of Things Journal, **7**(7), 6266-6278 (2019).
10. Chen, J., Zou, J., Su, M., & Zhang, L: Poisoning Attack and Defense on Deep learning Model: A Survey. Journal of Cyber Security (Chinese), **5**(4), 14-29 (2020).