AIShield
Powered by Bosch

Poisoning Analysis

Image Classification

Attack Type
Poisoning

Date
2024-08-05

Author
AIShield

Job Id
gAAAAABm...x51G_mRA==

Data Poisoning Analysis:

Executive Summary:

Poisoning Detected in Dataset: Yes

Percentage of Data Poisoned: 48%

Alert:Critical

Our assessment uncovers Critical-severity vulnerabilities, indicating that dataset contains many poisoned samples.This vulnerability poses a significant risk to businesses as poisoning attacks can compromise sensitive data and lead to financial, legal, and reputational consequences.In order to mitigate these risks, users are advised not to use the dataset for analysis.

This assessment aligns with the following OWASP Machine Learning Security Top-10 Vulnerabilities:

•ML02:2023 Data Poisoning Attack
•ML04:2023 Membership Inference Attack

The proposed measures align with these CVE/CWE Entries:

•CWE-20: Improper Input Validation
•CWE-212: Improper Removal of Sensitive Information Before Storage or Transfer
•CWE-226: Sensitive Information in Resource Not Removed Before Reuse
•CWE-501: Trust Boundary Violation
•CWE-707: Improper Neutralization
•CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
•CWE-1357: Reliance on Insufficiently Trustworthy Component

1. Simulation Report:

Type	No of Samples
Clean Data	1414
Poisoned Data	1318
Total Dataset Length	2732

The above table represents the amount of samples in the universal dataset identified as Clean and Poisoned.

In the context of this scenario, the percentage of samples in the universal dataset that are poisoned is 48%

2. Actions Recommended:

2.1 Dataset Reconstruction:

Reconstruct or reconstruct parts of the dataset using verified or untainted data sources to replace poisoned data or augment clean data to generate additional training samples from clean data points.

2.2 Diversity in Dataset:

Use a dataset which is more diverse which covers various sources and contexts. Including images from multiple sources helps the model learn general patterns, making it more resilient to poisoning strategies and ensuring its ability to accurately classify images.

2.3 Monitoring the Dataset during Training:

Monitoring data for suspicious patterns and content to detect anomalous patterns or shifts in the dataset distribution which might indicate potential poisoned samples.

2.4 Ensemble Learning:

Employ ensemble learning techniques to train multiple models on different subsets of the dataset and combine their predictions to mitigate the impact of individual poisoning attacks.

2.5 Cross-Validation and Model Evaluation:

Perform cross-validation and model evaluation to assess model performance across diverse subsets of the dataset. Measure the performance of model under different poisoning strategies.

3. Appendix:

This section offers supplementary data on our methodology, the vulnerabilities that AIShield was able to effectively detect in the dataset

3.1 OWASP Vulnerabilities:

ML02:2023 Data Poisoning Attack
The attack involves manipulating the training data, which can be related to the inadequate detection or handling of adversarial inputs.

ML04:2023 Membership Inference Attack
This weakness occurs when an automated mechanism,doesn't properly detect or handle inputs that have been modified or constructed in a way that causes the mechanism to detect an incorrect concept. This is directly related to the Membership Inference Attack where the attacker manipulates the model's training data.

For an expanded understanding of the OWASP Top - 10 ML Vulnerabilities, please visit
https://owasp.org/www-project-machine-learning-security-top-10/

3.2 CVE(Common Vulnerabilities and Exposures)/CWE(Common Weakness Enumeration):

CWE-20: Improper Input Validation
A weakness where the product receives input or data but does not validate or incorrectly validates that the input has the properties required to process the data safely and correctly. This can lead to various security issues, such as altered control flow, arbitrary control of a resource, or arbitrary code execution.

CWE-212: Improper Removal of Sensitive Information Before Storage or Transfer
This weakness occurs when a product stores, transfers, or shares a resource that contains sensitive information, but it does not properly remove that information before the product makes the resource available to unauthorized actors.

CWE-226: Sensitive Information in Resource Not Removed Before Reuse
This weakness occurs when a product releases a resource such as memory or a file so that it can be made available for reuse, but it does not clear the information contained in the resource before the product performs a critical state transition or makes the resource available for reuse by other entities.

CWE-501: Trust Boundary Violation
A base level weakness that occurs when a product mixes trusted and untrusted data in the same data structure or structured message. A trust boundary violation occurs when a program blurs the line between what is trusted and what is untrusted. By combining trusted and untrusted data in the same data structure, it becomes easier for programmers to mistakenly trust unvalidated data.

CWE-707: Improper Neutralization
The product does not ensure or incorrectly ensures that structured messages or data are well-formed and that certain security properties are met before being read from an upstream component or sent to a downstream component.

CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
This weakness occurs when a product uses an automated mechanism, to recognize complex data inputs, but it does not properly detect or handle inputs that have been modified or constructed such that it causes the mechanism to detect an incorrect concept.

CWE-1357: Reliance on Insufficiently Trustworthy Component
This weakness occurs when a product is built from multiple separate components, but it uses a component that is not sufficiently trusted to meet expectations for security, reliability, updateability, and maintainability.

For comprehensive data related to CVE/CWE list,please visit
https://cwe.mitre.org/

The digest value of the pdf file is 25b23a053c1309e25ee742d906a1a7081b8583041521ea5fb53d551a7592aafd