Countermeasures against Data Poisoning Attacks against Machine Learning Models with Multi-Sensor Inputs

Shyam Maisuria

Murata Laboratory, Graduate School of Information Science and Technology, Osaka University

Graduate School of Information Science and Technology | MURATA LAB. | Advanced Network Architecture Research Laboratory | OSAKA UNIVERSITY

1

Research Background[1]

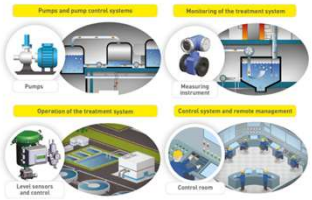
- **Popularity of Machine Learning**
  - Used in various fields such as anomaly detection in factories and farms
  - Machine learning-based applications are also of interest.
    - Smart healthcare, smart grid, smart water treatment, smart home , etc.
- **Concerned about attacks on machine learning models**

[1] Adedeji, K.B.; Hamam, Y. Cyber-Physical Systems for Water Supply Network Management: Basics, Challenges, and Roadmap. *Sustainability* **2020**, *12*, 9555. https://doi.org/10.3390/su12229555

2

Poisoning Attack Scenario

- The attacker can cause service disruption by altering the treatment process.
- Can also release untreated water to the rivers and streams
- To prevent this a machine learning model is trained to detect this attacks.
- What if attacker compromises the machine learning model itself



3

Attack risk in systems using multiple sensors

- **In the case of a system using multiple sensors, the risk of a hostile sample is higher if the sensor-based attack could also be misinterpreted by the machine learning model.**
  - Attackers may use vulnerable sensors to attack

Research Objectives

Verification of the possibility of attacks in cases where an attacker can monitor and tamper with some of the input features.
Propose countermeasures to the above attacks

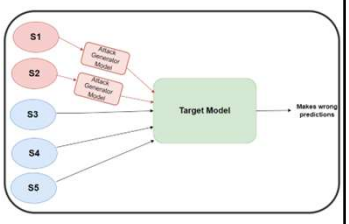
Approach

Attack: Create a machine learning model that can generate attacks by inputting some of the input features.
Countermeasures: Find the attacked features and exclude and discriminate them

4

Generation of attacks

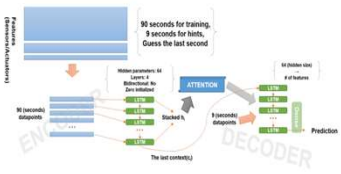
- The attacker installs an attacker generator model that changes the output of sensors.
- The attack generator model is trained in advance by using the knowledge of the attacker



5

How to Evaluate Attacks

- **Dataset:** Secure Water Treatment (Swat) dataset[4]
  - Total of 6 different water treatment processors
  - 51 sensors and actuators



Metric Used in this experiment

- Anomaly Score was used as a metric

$$D_t = \sqrt[4]{\sum_i (\hat{d}_{t,i} - d_{t,i})^4}$$

- Dt is prediction error

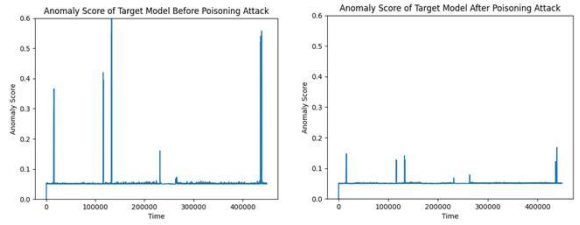
$$S = \min\left(\frac{H}{L \times R}, 1.0\right)$$

- S is anomaly score
- H and L are 90 and 20 percentile values of Dt

[4]Jonguk Kim, Jeong Han Yun, and Hyoung Chun Kim. Anomaly detection for industrial control systems using sequence-to-sequence neural networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 11980 LNCS:3–18, 2020.

6

Anomaly Scores of Target Model



7/13

7

Countermeasures -Overview

- Main idea is to identify and remove compromised sensors before the model is updated.
- How to do this? We use a Robust Model.

Robust Model:
- We use a "robust model" that doesn't heavily depend on specific features
- It allows us to calculate results without relying on certain sensor data

Detecting Compromised Sensors:
- Gradients calculated using attacker-generated values differ from those without compromised values
- Enables detection and identification of compromised sensor devices by comparing gradients.

Architecture of Robust Model

8/13

8

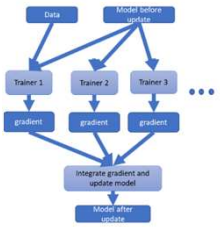
Countermeasures- Training the Robust Model

- **Training Process:**
- We create multiple trainers, each excluding features from potentially compromised sensors
- Calculated gradients are combined, outliers are removed, and the model is updated.

$$S_{A,j} = \text{cosineSimilarity}(g_{\text{all},j}, g_{\setminus A,j})$$

- Below equation determines which gradient to exclude

$$S_{A,j} > \text{Average}_i(S_{i,j}) - (\alpha \cdot \text{StdDev}_i(S_{i,j}))$$



9/13

9

Countermeasures- Detection

- **Compromised Sensor Detection:**
- We check if gradients calculated using a sensor's values are outliers.
- Each sensor gets a score, measuring how different its gradients are from the rest.

$$O_{A,j} = \max\left(\frac{Average_i(S_{i,j}) - S_{A,j}}{Stddev(S_{i,j})} - H, 0\right)$$

- Cumulative scores are tracked over time, and sensors exceeding a threshold are flagged. $\overline{O}_A = \overline{O}_A + O_A - \Delta.$

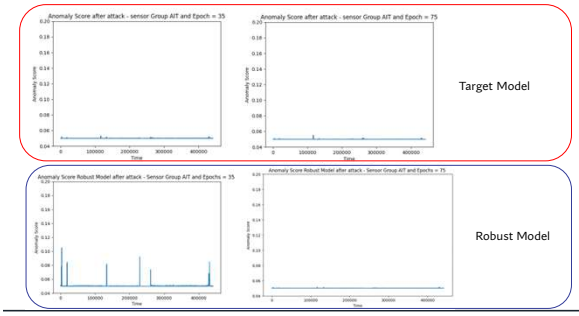
where Δ is a parameter.

If $\overline{O}_A$ exceeds the threshold, detect sensor A as the suspicious sensor.

10/13

10

Countermeasure-Results

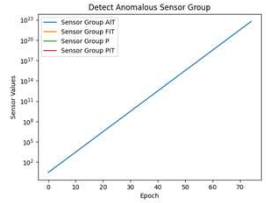


11/13

11

Countermeasure-Results

- Shows the cumulative anomaly score
- Detection threshold 10^{10}



- Show which sensor group is hacked
- Helps prevent further impact from poisoning attack on machine learning model

12/13

12

Conclusion

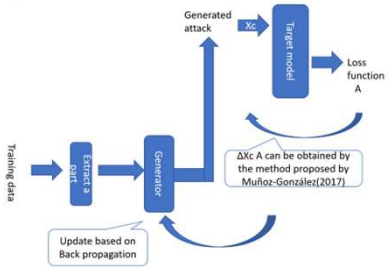
- Machine learning algorithms are **vulnerable to data poisoning** (compromising data collection), including Deep Learning systems.
- Poisoning attack is generated using part of hacked sensors.
- We propose a robust model as countermeasure and successfully detected hacked sensors.
- Robust model takes a lot of time to train
- Future task is to implement measures to shorter the training time

13/13

13

How to train the generator[2]

- **Input for Attack Generator Model**: Hacked sensor data
- **Output from Attack Generator Model**: Generated poisoned data
- **Training Objective**: Minimize loss function A to degrade target model accuracy



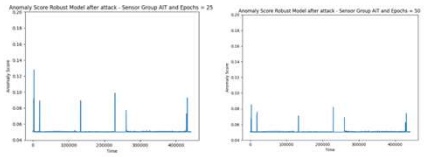
[2] Luis Mu˜noz-Gonz´alez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. AISec 2017 - Proc. 10th ACM Work. Artif. Intell. Secur. co-located with CCS 2017, pages 27–38, 2017.

14/16

14

Countermeasure-Results

- Shows poisoning attack for **robust model** with countermeasure
- Each figure trained for 25, 50 and 75 epochs respectively



15/16

15

Attack Assessment Results

- Each model is based on each process from Swat dataset
- We only focused on process 5
  - Which has total of 13 sensors and actuators
- Target model is trained using those 13 sensors and actuators
- Model accuracy: 71%
  - Found attacks: 5/7

AIT-501	Sensor	RO pH analyser; Measures HCl level.
AIT-502	Sensor	RO feed ORP analyser; Measures NaOCl level.
AIT-503	Sensor	RO feed conductivity analyser; Measures NaCl level.
AIT-504	Sensor	RO permeate conductivity analyser; Measures NaCl level.
FIT-501	Sensor	Flow meter; RO membrane inlet flow meter.
FIT-502	Sensor	Flow meter; RO Permeate flow meter.
FIT-503	Sensor	Flow meter; RO Reject flow meter.
FIT-504	Sensor	Flow meter; RO re-circulation flow meter.
P-501	Actuator	Pump; Pumps dechlorinated water to RO.
P-502 (backup)	Actuator	Pump; Pumps dechlorinated water to RO.
PIT-501	Sensor	Pressure meter; RO feed pressure.
PIT-502	Sensor	Pressure meter; RO permeate pressure.
PIT-503	Sensor	Pressure meter;RO reject pressure.

16/16

16

Research Background[1]

- **Popularity of Machine Learning**
  - Used in various fields such as anomaly detection in factories and farms
  - Machine learning-based applications are also of interest.
    - Smart healthcare, smart grid, smart water treatment, smart home , etc.
- **Concerned about attacks on machine learning models**
- Hostile samples that carry small noises that misclassify the model
  - E.g. Traffic sign “Stop” misinterpreted as “speed sign”

[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

17/16

17

Goal of Research

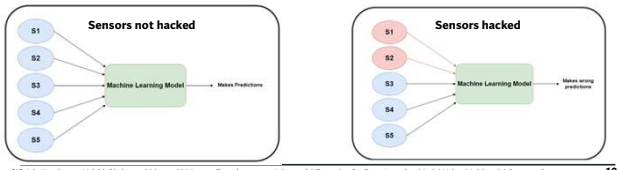
- **Objective:** Investigate data poisoning attacks and their profound impact on machine learning models.
- **Why Focus on Data Poisoning Attacks?**
  - **Significant Impact:** These attacks can severely compromise model integrity, making them a critical concern.
  - **Long-Term Consequences:** Once attacked, models may remain vulnerable over an extended period.
- **Our Research Aims To:**
  - Understand attack strategies that break machine learning models.
  - Develop effective mitigation and prevention techniques.
  - Safeguard real-world applications, ensuring their trustworthiness and safety.

18/22

18

Background- Data Poisoning Attacks[1]

- Many Machine Learning(ML) based systems uses multiple sensors
- It is possible that some sensors are hacked by the attacker
- The ML model can also be attacked from the hacked sensors
- Kurniawan et al demonstrated the possibility of this kind of attacks as an adversarial examples



[1] Ade Kurniawan, Yuichi Ohsita, and Masayuki Murata. Experiments on Adversarial Examples for Deep Learning Model Using Multimodal Sensors. Sensors, 22(22):8642, nov 2022.

19

Presentation Outline

- Background – Data Poisoning Attacks
- Poisoning Attack
- Target Model
- Results
- Conclusion
- References

20

Background

21

Poisoning Attack

22

Experiments

23

Secure Water Treatment Plant (Swat)[3]

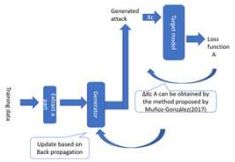
- Secure Water Treatment (Swat) is a water treatment site for cybersecurity research
- 11 days of continuous operation:
  - 7 days under normal operation
  - 4 days with attack scenarios
- 51 sensors and actuators
- Total of 6 different water treatment processors

[3]Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 10242 LNCS(October):88–99, 2017.

24

Attackers' objective

- To generate data poisoning attacks that evades detection
- Objective function = weighted sum loss function
- $\mathcal{L}_T = \mathcal{A} \cdot \alpha + (1 - \alpha) \cdot \mathcal{L}_C$
- $\mathcal{L}_C$ = generator models objective.$\mathcal{A}$ = target models objective
- Alpha(α) controls the importance of each of the objective function
- Alpha(α) = low : prioritize evading detection
- Alpha(α) = high: prioritize effectiveness of attack



25/22

25

Target Model – Features

35	AIT-501	Sensor	RO pH analyser; Measures HCl level.
36	AIT-502	Sensor	RO feed ORP analyser; Measures NaOCl level.
37	AIT-503	Sensor	RO feed conductivity analyser; Measures NaCl level.
38	AIT-504	Sensor	RO permeate conductivity analyser; Measures NaCl level.
39	FIT-501	Sensor	Flow meter; RO membrane inlet flow meter.
40	FIT-502	Sensor	Flow meter; RO Permeate flow meter.
41	FIT-503	Sensor	Flow meter; RO Reject flow meter.
42	FIT-504	Sensor	Flow meter; RO re-circulation flow meter.
43	P-501	Actuator	Pump; Pumps dechlorinated water to RO.
44	P-502 (backup)	Actuator	Pump; Pumps dechlorinated water to RO.
45	PIT-501	Sensor	Pressure meter; RO feed pressure.
46	PIT-502	Sensor	Pressure meter; RO permeate pressure.
47	PIT-503	Sensor	Pressure meter;RO reject pressure.

26/22

26

Hacked Sensors

35	AIT-501	Sensor	RO pH analyser; Measures HCl level.
36	AIT-502	Sensor	RO feed ORP analyser; Measures NaOCl level.
37	AIT-503	Sensor	RO feed conductivity analyser; Measures NaCl level.
38	AIT-504	Sensor	RO permeate conductivity analyser; Measures NaCl level.
39	FIT-501	Sensor	Flow meter; RO membrane inlet flow meter.
40	FIT-502	Sensor	Flow meter; RO Permeate flow meter.
41	FIT-503	Sensor	Flow meter; RO Reject flow meter.
42	FIT-504	Sensor	Flow meter; RO re-circulation flow meter.
43	P-501	Actuator	Pump; Pumps dechlorinated water to RO.
44	P-502 (backup)	Actuator	Pump; Pumps dechlorinated water to RO.
45	PIT-501	Sensor	Pressure meter; RO feed pressure.
46	PIT-502	Sensor	Pressure meter; RO permeate pressure.
47	PIT-503	Sensor	Pressure meter;RO reject pressure.

27/22

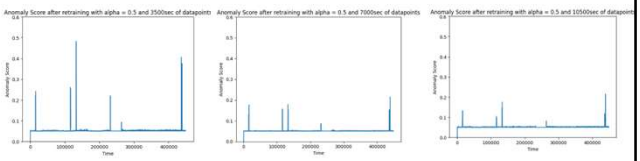
27

Results

28

Anomaly Scores with varied data points

- Alpha = 0.5 is constant
- As number of datapoints increase anomaly scores decrease
- Shows poisoning attack is successful if we use more datapoints

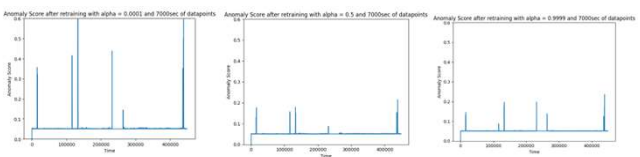


29/22

29

Anomaly Scores with varied alpha values

- Number of datapoints is constant at 7000sec
- As alpha value increases the anomaly score decreases
- Shows poisoning attack is successful when we increase the alpha value

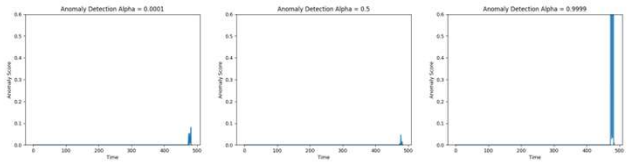


30/22

30

Check if Target Model detects Poisoned Attack

- Alpha values determine if attack is detected or not
- Higher alpha values = attack gets detected and poisoning attack is stronger
- Lower alpha values = avoid attack detection and poisoning attack is weaker



31/22

31