Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Detection of Poisoning Attacks in Internet of Things-Based Machine Learning Systems

Rajkumar Nirmalan[1], Kandaswamy Gokulakrishnan[2]
[1]*Assistant Professor, Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, Sivakasi*
[2]*Associate Professor, Department of Electronics and Communication Engineering, Anna University, Guindy*
Corresponding Email : nirmalan@mepcoeng.ac.in

The rise of machine learning (ML) in the Internet of Things (IoT), both at the edge and in the cloud, is driving advancements in areas such as environmental monitoring and industrial automation. However, integrating ML in IoT also introduces significant security challenges. Adversaries can manipulate sensor data to alter training datasets, leading to what is known as poisoning attacks. These attacks can embed "backdoors" and "neural trojans" into models, causing targeted misclassifications, malicious actions, and significant degradation in performance. Using recently advanced provenance frameworks this study suggests a methodology for detecting harmful data by using contextual information about the handling and origin of data points in the training set. This method works both with and without a reliable test set. By relying on accurate provenance information, our method can effectively identify and mitigate poisoning threats in IoT environments, ensuring the integrity and reliability of machine learning systems.
Keywords: Machine Learning, Internet of things, poison case, smart grid model.

1. Introduction

Rapid advancements in machine learning (ML) and its connection to the Internet of Things (IoT) have occurred in recent years. These days edge sensors can gather the data needed to train machine learning models. These learned models can anticipate events and keep an eye on sensor data in real-time causing linked devices to take particular actions. To ensure safety an ML model can for instance automatically engage the vehicle's braking system when it detects a stop sign. However, deploying ML within IoT frameworks introduces specific security vulnerabilities, as malicious actors may exploit the training data by manipulating sensors. Poisoning attacks enable adversaries to introduce backdoors and neural trojans, provoke intentional misclassification or erroneous behavior, and substantially degrade overall system performance. One well-known instance of a poisoning attack that happened

outside of the Internet of Things was when Microsoft trained its chatbot Tay to write tweets that resembled those of a person. When some individuals started tweeting harsh words, Tay responded by sending out tweets that were also offensive [1]. After just sixteen hours, Microsoft was compelled to take down the bot. Similar assaults, such as those intended to evade environmental control, are conceivable in IoT applications.

Current methods for locating toxic data points concentrate on studying the training set. But provenance data often exists, especially in IoT systems, and this information can help identify harmful data points. The term "provenance data" describes the genealogy or meta-information associated with a data element elucidates the methodologies engaged in its creation, provenance, and evolution. This may encompass specifics regaing the apparatus that generated the data, including the firmware iteration, user identification, and temporal marker. [2] In this study, scholars present a preemptive strategy to detect harmful data. before deploying models by utilising data provenance. This technique divides the untrusted data into categories based on a high correlation between sample-to-sample poisoning likelihood within each group using provenance meta-data. After the training data has been suitably divided into segments, the efficacy of the classifier developed with and without that specific group is evaluated to determine. the data points in each segment collectively [3]. To our knowledge, this approach is the first defence tactic that filters unreliable data metrics and mitigates contamination threats through the application of data lineage.

The Method of Probability of Sufficiency (MPS) and Reject on Negative Impact (RONI) method are two previous techniques that identify toxic data by assessing how each individual data point affects the trained model's performance. These two approaches compare the efficacy of the model assessed on a robust data set in order to assess it. This is how our system also assesses performance when a reliable data set is provided [4]. However, our strategy amplifies the influence of hazardous data and allows for improved detection rates by examining all of the segment data collectively. A significant benefit given the massive volumes a notable advantage of data gathered across various Internet of Things (IoT) contexts is that the detection methodology exhibits enhanced scalability, as it reduces the frequency with which the model must undergo retraining to a minimal fraction of the overall quantity of unreliable data points. [5]. Ultimately, demonstrating how provenance information, a topic overlooked by both RONI and PS, is addressed, allows our method to identify toxic data in situations where trusted data is not available.
The following are the contributions made by this paper:

1) To develop a generalized supervised learning model suitable for Internet of Things contexts, suggest a novel approach for identifying and screening toxic data. Specifically, this approach takes advantage of data provenance to find sets of data that have a strong correlation in their likelihood of being contaminated [6].
2) Two versions of our provenance-based defense are offered to address situations where datasets that are entirely untrusted and partially trusted are accessible.
3) Assess our method's performance in identifying toxic data produced by and discover that models trained on both fully untrusted and partially trusted data sets perform much better when our defense is used as a filter before training. Further demonstrates that, on average, our technique performs faster and more efficiently

than RONI [7].

4) Then, given a partially trustworthy data set, this provides our provenance defense to detect toxic data. The next part, will provide an additional methodology to handle data that is completely untrusted, talk about potential collaboration and targeted assaults, and present ways to protect against them [8]. Afterward, this analyses our methods empirically. it wraps up by presenting relevant work.

2. TERMINOLOGY, THREAT MODEL, AND MOTIVATION

As illustrated in Figure 1, this presents a Provenance-based poison detection service in this study. Our method uses provenance information to remove harmful data from IoT observations [9]. This leverage recently suggested frameworks to guarantee the immutability and incapability of provenance data from IoT contexts to be altered. Smart grids and SCADA systems are two further IoT scenarios where adversaries can find it advantageous to contaminate an ML model [10].

This takes into account an opponent in our threat model whose objective is to lower the ML model's accuracy. For instance, managers of polluting factories can try to make the classifier worthless by drastically lowering its overall performance. As an alternative, they might work to prevent the model from learning the negative consequences of a certain chemical, which would lower the accuracy for a given input [11]. An antagonist typically can manipulate only a limited array of data sources within operational systems; attaining control over all data sources may be impractical or excessively costly. For instance, a factory manager is likely restricted to altering only the sensors present within their facility excluding those located in other establishments. As a result, erroneous data will usually emanate from specific sensors and locations. In essence, this contends that the perpetrator can solely amend data elements that exhibit unique provenance signatures [12].

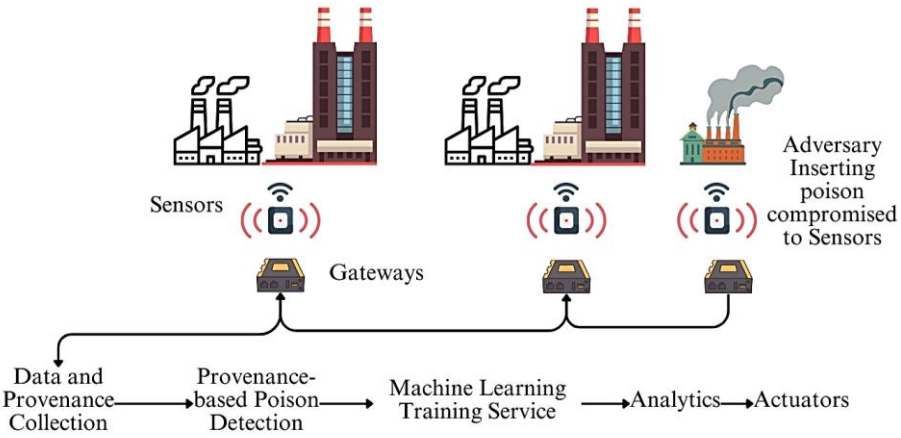

Figure 1: Poison detection service Scenario

3. DEFENSE AGAINST DATA OF LIMITED TRUSTWORTHINESS

In this discourse, this elucidates a provenance-oriented countermeasure against data poisoning in contexts where the acquired information is subject to incomplete trust. By

"incomplete trust," this indicates that certain segments of the collected data are accepted as reliable and remain uncorrupted. Incomplete trust in training data can be obtained in real-world circumstances either via trusted sources of data or by hand curation of the collected data. To guarantee the accuracy of the information gathered, the regulator might, for instance, physically keep an eye on a few sensors [13].

The technique is independent of the particular supervised machine learning algorithm employed and, theoretically, can also be utilized with unsupervised algorithms [14]. To make it easier to compare in evaluate the efficacy of the trained models, our analysis is confined to supervised learning methodologies. The ensuing inputs are employed in this approach:
1) a monitored computational learning procedure;
2) a training dataset characterized by a degree of partial trust is segmented into two components, namely a trusted subset and an untrusted subset, gathered with the intent of educating the machine learning classifier;
3) a reliable and reputable provenance dataset that encompasses metadata elucidating the provenance and lineage of every data element within the untrusted segment of the training dataset;
4) a provenance attribute that signifies the manner in which toxic elements will be grouped within the unverified segment of the data collection.
The methodology process is shown in Figure 2 complete with pseudocode using the inputs mentioned above. Each erroneous data point in the training set is first linked to its provenance record in the process. The untrustworthy dataset is split into segments, each of which shares a consistent value for the selected provenance attribute to detect and remove potentially harmful data [15]. One way to divide up the dataset would be to use the original equipment or location where the data was gathered. Following that classifiers are trained with and without each segment to examine it for possible data poisoning. If the segment is deemed contaminated and eliminated the classifier that was trained without it (filtered model) beats the one that was trained with it (unfiltered model) on the test set.

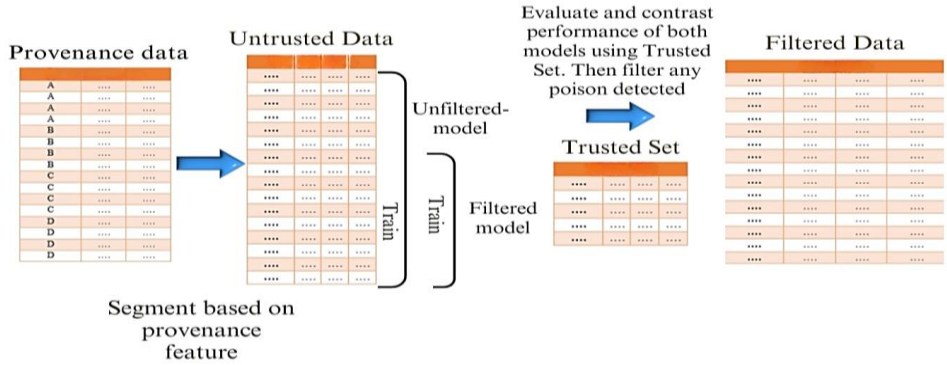

Figure 2: A Defense Mechanism for Partially Trusted Data

4. PROTECTION OF SEMI-TRUSTED DATA

In addition, this present a calibration process that attempts to comprehend the impact of eliminating a valid section from the training set. This allows us to determine the threshold at which the removal of a segment should degrade the classifier's performance before this

consider the gadget to be poisonous. The calibration process is carried out through a series of trials in which:
1) A single portion of data is selected at random from the untrusted collection, while a distinct portion of validated data is randomly extracted from the trusted collection.
2) Legitimate data is used both when training classifiers and when not.
3) The performance disparity on the remaining reliable data points is recorded.

To get a good idea of the distribution for the performance change, the user should try as many times as required. The user can select a threshold based on his or her requirements by using this estimate. With these circumstances, this suggests repeating the following procedure multiple times while evaluating a particular unreliable section. The first randomly selected segments range from 10 to twenty. The model is then trained on both the segment being evaluated and the segments that were chosen at random. Following the creation of a new model utilizing only the randomly chosen segments, the outcomes are compared. This process needs to be carried out many times in order to account for the inherent variability in the outcomes. A segment is deemed hazardous and eliminated from the dataset if the mean performance change exceeds the designated threshold.

5. PROTECTION OF COMPLETELY UNTRUSTED DATA

However, achieving a partially trusted data collection may be challenging or impossible due to the expense of human data verification and real-time constraints that hinder data verification. This presents a procedure to apply our strategy to entirely untrusted data sets.
1) Based on the chosen provenance attribute, divide the data into segments by signature.
2) Randomly allocate a segment of the data to the training set while designating the remaining segment to the evaluation set for each section.
For every signature in the chosen provenance feature, perform to develop two distinct models, the first model will utilize the entirety of the training dataset, while the second model will be constructed by excluding the relevant segment from the training dataset.

5.1 Targeted Attacks

The previous method prevented compromised devices A and B from directly influencing the evaluation of their data points. However, compromised devices can collude by inserting elements into the evaluation set that conceal compromised status. Device B can also add data points that facilitate the identification of compromised but authentic devices. False Negative Attacks occur when device A adds data points to the decision boundary, while device B adds points between the original and updated decision boundaries. The model trained without data from device A will be less accurate, as points from device B are wrongly classified, giving the false impression that despite being harmful, device A supplied valid data. False Positive Attacks occur when device B places points slightly outside the decision boundary and farther away from the true boundary, producing false positives.

5.2 Experimental Evaluation

This produced a model to evaluate our strategy that is trained using data points from numerous devices in the Internet of Things. The production of poison involves two different methods that have been previously suggested in the literature. The goal of both mechanisms

is to support vector machines (SVMs). For each type of toxin, this adhered to the protocol below.

Initially, the quantity of devices within the system was ascertained and the pertinent poisoning process was employed to produce legitimate and hazardous data points. To generate provenance data for each data point the number of contributing data points for each device was the same. Two defenses, RONI and RONI with calibration, were used to compare our method. Since that performed better, this solely presents the Calibrated RONI findings and utilizes it as a baseline1. This utilized the same size for the trustworthy set in order to compare the two approaches. As a result, the calibration, validation, and baseline portions of the trusted set utilized for the Calibrated RONI are divided.

Furthermore, this constructed a separate, autonomous testing dataset comprising 5,000 genuine data instances that were designated explicitly for benchmarking purposes. This assessed the accuracy of four distinct models utilizing this benchmarking dataset: the optimal detection model, which was exclusively trained on authentic data instances; the no-defense model, which incorporated all data instances submitted to the system; and the provenance defense model, which was developed subsequent to the exclusion of data instances identified as harmful by our defense mechanisms.

Thereafter, this categorized an untrusted segment as harmful if the performance deviation observed during the calibration trials surpassed the average value plus one standard deviation. Naturally, this threshold can be adjusted to enhance recall at the potential detriment to precision or vice versa. An alternative approach involves employing a cross-validation dataset to optimize this parameter. Ultimately, the user may also perform statistical analyses to corroborate the assertion that an untrusted segment is authentic if they can replicate the performance change distribution observed in the calibration trials. The generated distribution alongside a p-value could then be utilized to refine the threshold parameter.

5.3 Potency in the presence of poison I

This study's synthetic dataset and methodology which also used to assess our defense against the poisoning attack. There are two features and two different classes in the dataset. The numbers displayed are based on the average outcomes that were determined.

5.4 Impact of a trusted set size in contexts with partial trust

In this experiment, the total number of legitimate training instances is 1000, whereas the aggregate of detrimental training instances amounts to 200. There were merely two and 10 devices in total honest and dishonest, respectively. Figure 3 illustrates the findings, which indicate that to observe a notable enhancement over the absence of a defense mechanism, the provenance defense mandates a minimum of 100 data points within the trusted dataset. It reaches an accuracy that closely approximates perfect detection by the time 380 data points are utilized. The provenance defense consistently outperforms the baseline.

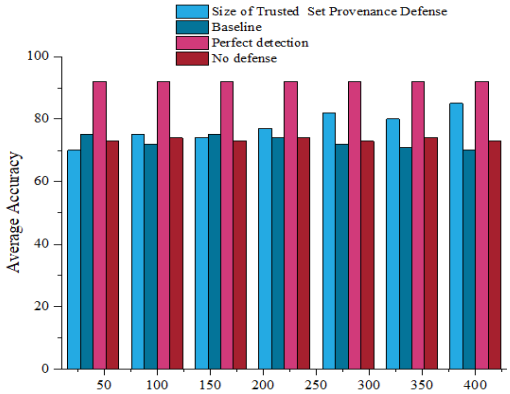

Figure 3: Influence of augmenting the dimensions of the reliable subset on the mean precision attained in the context of poison I

5.5 Impact of increasing poison concentration

The effect of developing the poison concentration in the untrusted group was examined in this experiment. With the combined contribution of 10 devices, each contributing 100 data points 1000 training points were obtained. The amount of poison added in each trial was then adjusted by changing the number of compromised devices from 1 to 7. There was a 300-data point limit on the reliable dataset. Figure 4 illustrates the results, showing that even as the proportion of poisoned data nears 70%, our approach consistently improves the final classifier's performance and generally outperforms the baseline.

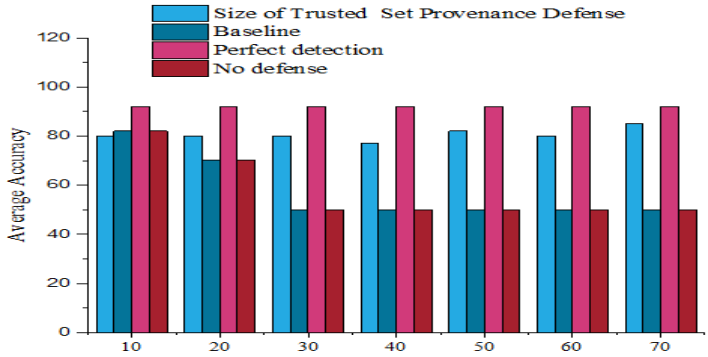

Figure 4: accuracy percentage under poison I.

Runtime: This assertion is valid despite the fact that both the provenance technique and the foundational approach possess the capacity for parallel processing. As a result, the foundational framework would necessitate O(m) times additional resources, encompassing memory and CPU cores, even in scenarios of complete parallelization. Furthermore, it is anticipated that the computation time will be O(m) times prolonged in a non-parallelized framework. This corroborate this through empirical data obtained from our previous experiment, in which this assessed the time intervals required to filter data sets of diverse sizes utilizing both the provenance and foundational strategies, while maintaining a constant number of devices. The outcomes are depicted in Figure 5. Our results suggest that our

approach is more appropriate for the extensive data sets typical of Internet of Things applications, as the provenance strategy exhibits a notable speed advantage relative to the foundational methodology.

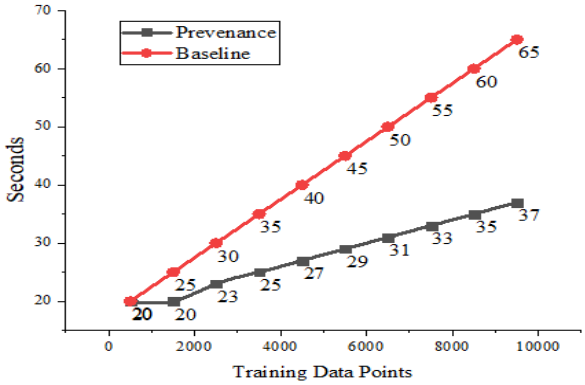

Figure 5: Mean execution duration for semi-trusted techniques as a variable of the quantity of training data instances.

5.6 Effectiveness Under Poison II

The evaluation tested the defense mechanisms against harmful data, focusing on a specific attack vector incorporated into a Support Vector Machine (SVM) using gradient ascent. The figures represent the average results from five experimental iterations. In these experiments, there were only two devices in total—one honest and one compromised. The total number of training data points was 120, with 20 being harmful. The reliable dataset consisted of 120 points, unless noted otherwise. This methodology aims to gauge how well our defense can handle such targeted poisoning attacks.

5.7 Impact of a trusted set size in contexts with partial trust

The impact of raising the trusted set size while maintaining the same values for the other parameters is seen in Figure 6. The provenance defence significantly enhances the final classifier's performance, even at 90 data points. By contrast, before the baseline can outperform no defence, at least 120 data points are required. Our provenance method's capacity to increase model accuracy has converged to the point where it performs almost as good as perfect detection after 150 data points.

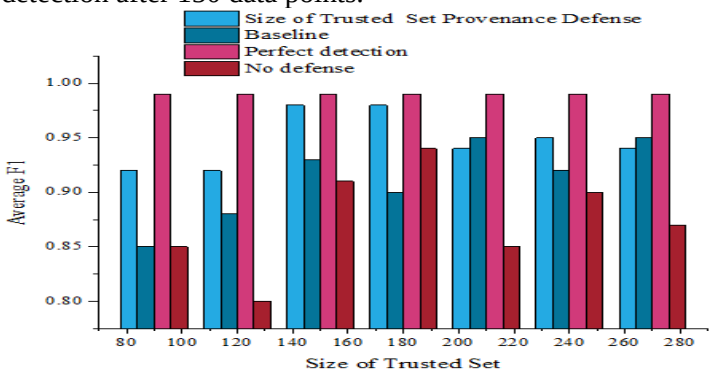

Figure 6: The impact of increasing the size of the trusted set on the average accuracy

Impact of poisoning percentage in totally untrusted environments: The evaluation set may contain toxic data in completely untrusted situations. Therefore, only a small percentage of the gathered data must be toxic for our approach to be able to identify toxic data. In this experiment, this measure how well our approach filters toxic data as the untrusted data set gets progressively more toxic.

This are unable to use Calibrated RONI as a baseline in completely untrusted situations because it depends on a trustworthy set. Rather, this contrast our approach with both no defence and flawless detection. The results of adding more poison in a completely untrusted context are seen in Figure 7. This are able to successfully improve the final classifier's performance when fewer than 25% of the data is contaminated. Nevertheless, our technique can no longer outperform no defence if 25% of the data is tainted.

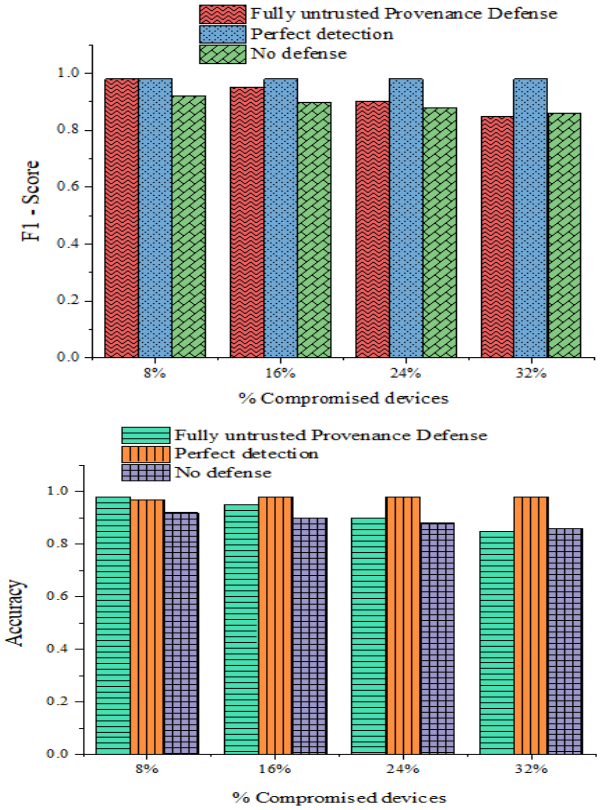

Figure 7: Fully untrusted defense under poison II

6. CONCLUSION

The integration of machine learning in IoT environments, where data is dynamically collected and learned from online, makes these systems highly vulnerable to poisoning attacks. Despite extensive documentation of such vulnerabilities, effective countermeasures are scarce. This study presents a groundbreaking method for detecting and filtering toxic data using data provenance, which tracks the origin and history of data points. This novel approach represents the first use of data provenance to counteract poisoning attacks. It

explores potential weaknesses, including collusion attacks, and proposes solutions for completely untrusted scenarios. Testing with two established poisoning techniques shows that this provenance-based defense outperforms baseline methods in both detection accuracy and runtime efficiency.

Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

Funding Information
Not Applicable

Author Contribution
Not Applicable

Data Availability Statement
Not Applicable

Abbreviation

ML - Machine Learning
IoT - Internet of Things
RONI - Reject Negative Impact
PS - Probability of Sufficiency
PUF - Physical Unclonable Functions
EPA - Environmental Protection Agency

REFERENCE

1. Borgnia, E., Cherepanova, V., Fowl, L., Ghiasi, A., Geiping, J., Goldblum, M., Goldstein, T., & Gupta, A. (2021). Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 3855–59. https://doi.org/10.1109/ICASSP39728.2021.9414803.
2. Chacon, H., Silva, S., & Rad, P. (2019). Deep learning poison data attack detection. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) pp. 971–78. https://doi.org/10.1109/ICTAI.2019.00136.
3. Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., & Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. ACM Computing Surveys (CSUR), 55(13), 1–39. https://doi.org/10.1145/3549763.
4. Dunn, C., Moustafa, N., & Turnbull, B. (2020). Robustness evaluations of sustainable machine learning models against data poisoning attacks in the Internet of Things. Sustainability: Science, Practice, and Policy, 12(16), 6434. https://doi.org/10.3390/su12166434.
5. Gao, J., Karbasi, A., & Mahmoody, M. (2021). Learning and certification under instance-targeted poisoning. In Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence pp. 2135–45. https://doi.org/10.48550/arXiv.2102.07020.
6. Jiang, W., Li, H., Liu, S., Ren, Y., & He, M. (2019). A flexible poisoning attack against machine learning. In ICC 2019 - 2019 IEEE International Conference on Communications (ICC) pp. 1–6. https://doi.org/10.1109/ICC.2019.8761822.
7. Lee, T., Edwards, B., Molloy, I., & Su, D. (2019). Defending against neural network model stealing attacks using deceptive perturbations. In 2019 IEEE Security and Privacy Workshops

(SPW) pp. 43–49. https://doi.org/10.1109/SPW.2019.00019.

8. Lin, T. (2020). Deep learning for IoT. In 2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC) pp. 1–4. https://doi.org/10.1109/IPCCC49379.2020.00010.
9. Mahloujifar, S., Ghosh, E., & Chase, M. (2022). Property inference from poisoning. In 2022 IEEE Symposium on Security and Privacy (SP) pp. 1120–37. https://doi.org/10.1109/SP46215.2022.9833618.
10. Mehra, A., Kailkhura, B., Chen, P.-Y., & Hamm, J. (2021). How robust are randomized smoothing based defenses to data poisoning? In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13244–53. https://doi.org/10.1109/CVPR46437.2021.01302.
11. Muñoz-González, L., Pfitzner, B., Russo, M., Carnerero-Cano, J., & Lupu, E. C. (2019). Poisoning attacks with generative adversarial nets. arXiv [cs.LG], 22(4), 78–93. https://doi.org/10.48550/arXiv.1906.07773.
12. Tian, Z., Cui, L., Liang, J., & Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Computing Surveys (CSUR), 55(8), 1–35. https://doi.org/10.1145/3539547.
13. Truong, L., Jones, C., Hutchinson, B., August, A., Praggastis, B., Jasper, R., Nichols, N., & Tuor, A. (2020). Systematic evaluation of backdoor data poisoning attacks on image classifiers. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 788–89. https://doi.org/10.1109/CVPRW50498.2020.00202.
14. Xiang, Z., Miller, D. J., & Kesidis, G. (2019). A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense. In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) pp. 1–6. https://doi.org/10.1109/MLSP.2019.8918807.
15. Zhao, B., & Lao, Y. (2022). CLPA: Clean-label poisoning availability attacks using generative adversarial nets. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(8), 9162–70. https://doi.org/10.1609/aaai.v36i8.20