Trustworthy Machine Learning: Poisoning Attacks

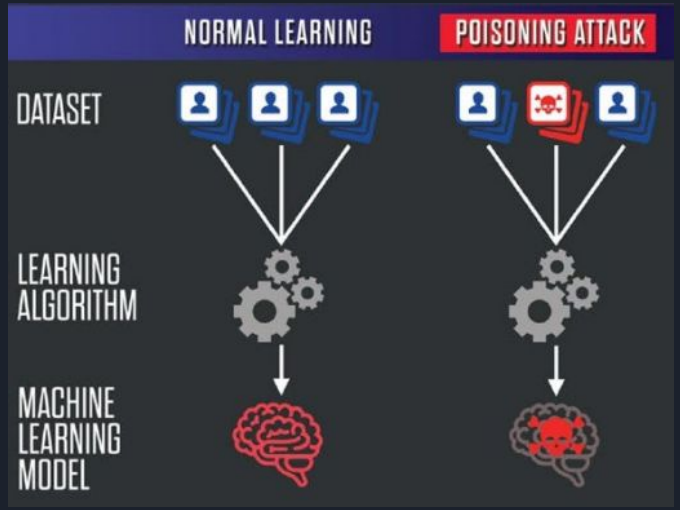
Luke Crosby
Clarence Fernandes
Pranav Kuchibhotla

What are Poisoning Attacks?

Poisoning refers to an intentional attempt to harm the results of a ML model

- It involves adding fake “data” to the dataset, which will skew the results of the model
- Labels can be misclassified
- Models themselves can also be slightly changed

In a case like medical data, this is a very serious issue as it can cause incorrect diagnoses, potentially harming the patient



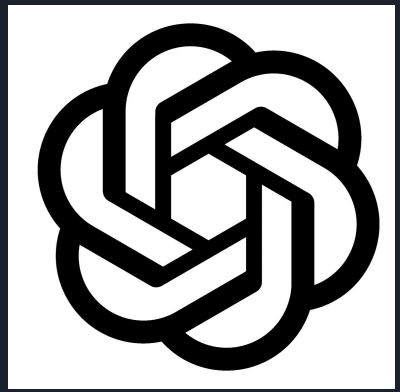
What can be done?

Artificial Intelligence is becoming increasingly mainstream every year, and as it becomes more involved in day-to-day life, it needs to remain accurate

Programs like ChatGPT are affected unintentionally by poisoning, as it takes information from the internet without knowing if it is true

There are multiple ways to combat security attacks [1]:

- **Federated Learning - User Elimination**
- Robust data validation
- Provenance tracking
- …more!



1.“Data Poisoning: The Essential Guide: Nightfall AI Security 101.” *The Essential Guide | Nightfall AI Security 101*, Nightfall AI, www.nightfall.ai/ai-security-101/data-poisoning. Accessed 10 Nov. 2024.

Related work - Other Defense Methods

Robust data validation:

- Including more data can help balance the difference between real and “poisoned” data
  - “the models’ performances will be degraded, in terms of accuracy and detection rates, if the number of the trained normal observations is not significantly larger than the poisoned data. [2]”

Provenance tracking:

- Track data sources, and data modification to decide if it is trustworthy
- Test an untrusted data set against a trusted data set, if it performs much worse, consider it poisoned
  - “ the detection effectiveness of the proposed provenance defense surpasses that of the baseline, thereby enabling the use of online and regularly re-trained machine learning models in adversarial environments where reliable provenance data can be obtained. [3]”

2. Dunn, C.; Moustafa, N.; Turnbull, B. Robustness Evaluations of Sustainable Machine Learning Models against Data Poisoning Attacks in the Internet of Things. *Sustainability* **2020**, *12*, 6434. https://doi.org/10.3390/su12166434
3. Baracaldo, Nathalie, et al. “Mitigating poisoning attacks on machine learning models.” *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3 Nov. 2017, pp. 103–110, https://doi.org/10.1145/3128572.3140450.

Federated Learning - User Elimination

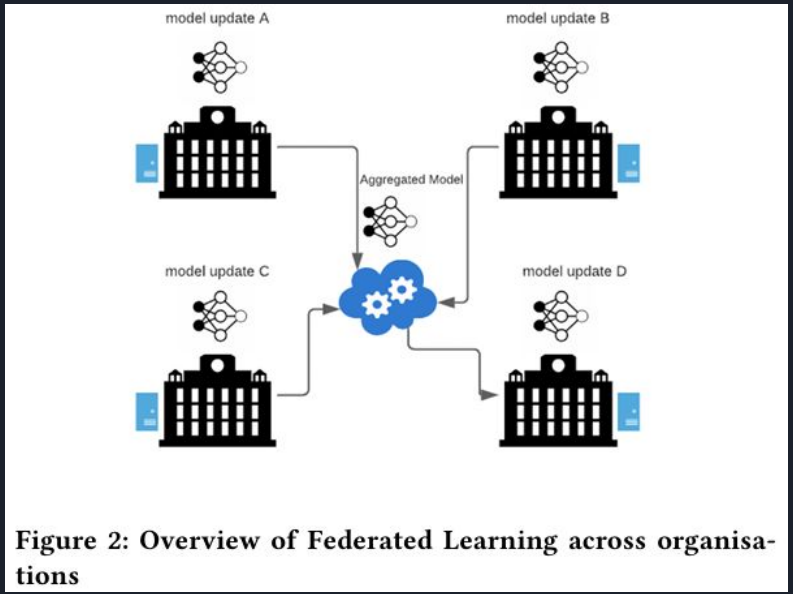


Figure 2: Overview of Federated Learning across organisations

Federated Learning: “Multiple devices collaboratively learn a machine learning model without sharing their private data under the supervision of a central server [4]”

- Especially vulnerable to poisoning

User Elimination: Loss is reported to the server when a user runs the model [5]

- A clustering technique is used to identify the users that are attackers and they are removed from the server
- Further, this method respects users privacy, which is a key aspect of FL

4. Mammen, P. M. (2021). Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.
5. Galanis, N. (2024). Defending against Data Poisoning Attacks in Federated Learning via User Elimination. *arXiv preprint arXiv:2404.12778*.

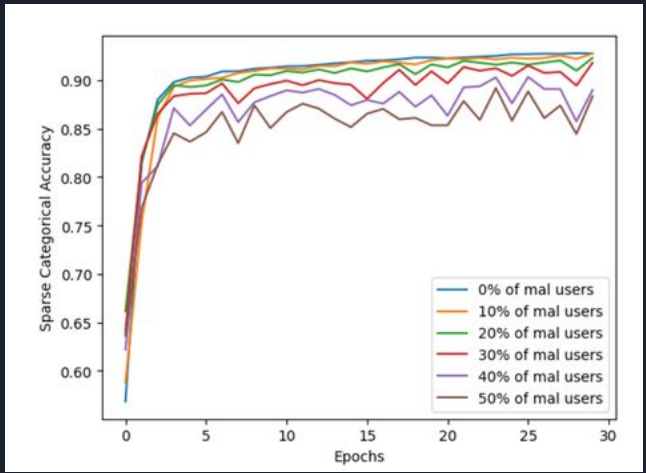
How does it work?

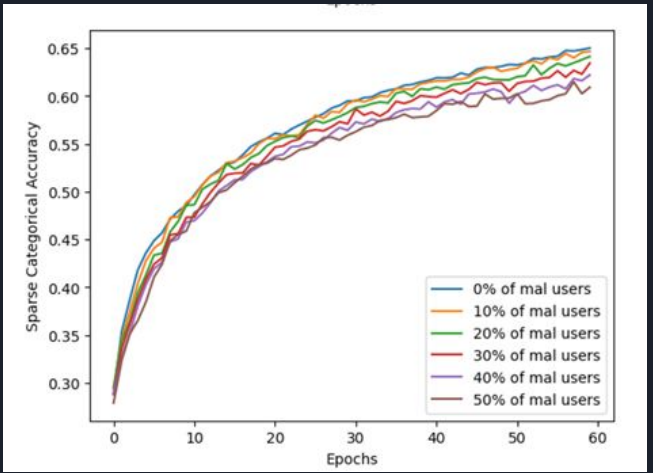
Proposed Defense Mechanism:

- Each client/user reports training loss to the server
- Differential Privacy
- Server detects anomalies
- Users identified as potentially malicious are eliminated
- This defense aims to mitigate the impact of data poisoning at each training round, enhancing model robustness from the beginning

Experimental Results

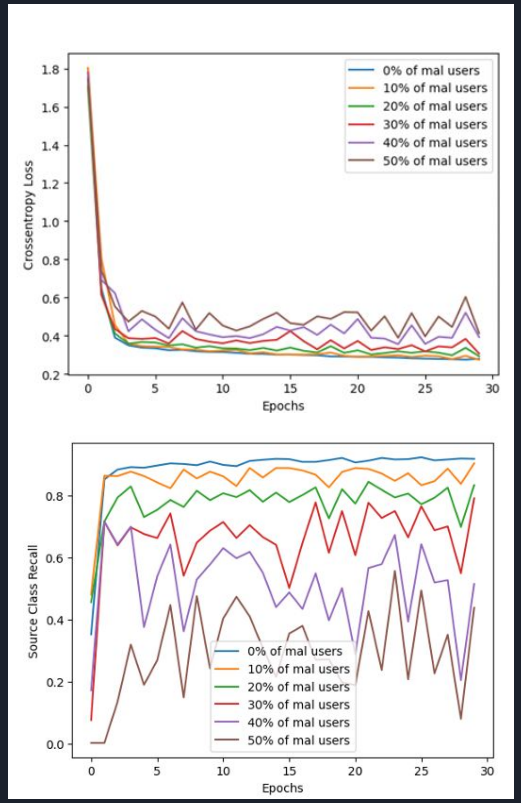
- Datasets Used: MNIST(handwritten digits) and CIFAR-10(object recognition)
- Convolutional Neural Networks (CNNs) were trained, with 30 epochs on MNIST and 60 epochs on CIFAR-10
- Key metrics: Sparse Categorical Accuracy, Cross Entropy Loss, and Source Class Recall
- MNIST CIFAR-10

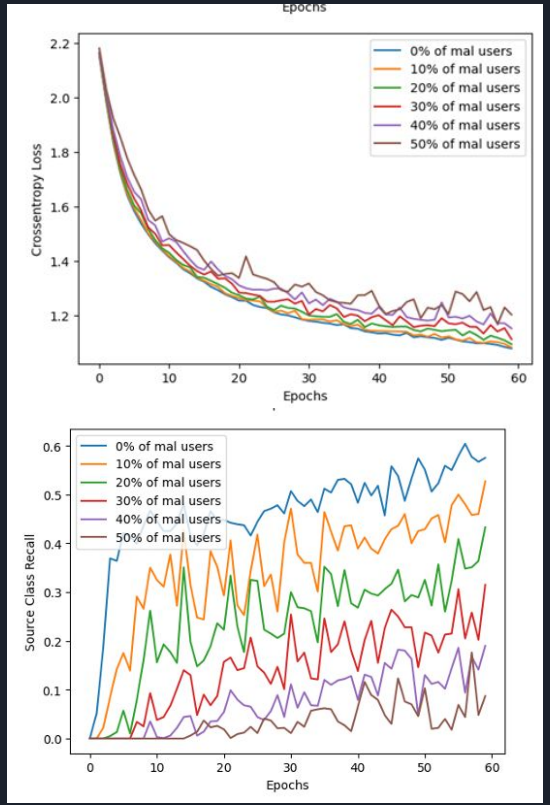




5. Galanis, N. (2024). Defending against Data Poisoning Attacks in Federated Learning via User Elimination. *arXiv preprint arXiv:2404.12778*.

Experimental Results





5. Galanis, N. (2024). Defending against Data Poisoning Attacks in Federated Learning via User Elimination. *arXiv preprint arXiv:2404.12778*.

Conclusions/Future Work

Explored Federated Learning (FL) and Data Poisoning Attacks.

- Key Findings: Malicious users (over 20%) can subtly alter FL model predictions without affecting accuracy.
- Malicious users tend to report higher loss values.

Defense Mechanism: Developed a novel approach combining user loss monitoring with Local Differential Privacy.

- Used K-Means clustering to identify and exclude attackers while preserving FL's privacy promise.
- Results: Demonstrated robust defense across MNIST and CIFAR datasets.
- Effective in both model integrity and attacker elimination.

Future Scope: Potential to generalize the method across different FL models and attack types.

- This project was limited to k-means, other models can be tested

Future Work

- Future Scope: Potential to generalize the method across different FL models and attack types.
  - This project was limited to k-means, other models can be tested

Sources

1. “Data Poisoning: The Essential Guide: Nightfall AI Security 101.” *The Essential Guide | Nightfall AI Security 101*, Nightfall AI, www.nightfall.ai/ai-security-101/data-poisoning. Accessed 10 Nov. 2024.
2. Dunn, C.; Moustafa, N.; Turnbull, B. Robustness Evaluations of Sustainable Machine Learning Models against Data Poisoning Attacks in the Internet of Things. *Sustainability* **2020**, *12*, 6434. https://doi.org/10.3390/su12166434
3. Baracaldo, Nathalie, et al. “Mitigating poisoning attacks on machine learning models.” *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3 Nov. 2017, pp. 103–110, https://doi.org/10.1145/3128572.3140450.
4. Mammen, P. M. (2021). Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.
5. Galanis, N. (2024). Defending against Data Poisoning Attacks in Federated Learning via User Elimination. *arXiv preprint arXiv:2404.12778*.