AI-ML SYSTEMS

# SECOND INTERNATIONAL CONFERENCE ON AI-ML SYSTEMS

HYBRID, BANGALORE, INDIA

12-15 OCTOBER 2022

AN INITIATIVE OF THE COMSNETS ASSOCIATION

*In association with*

acm In-Cooperation

acm SIGAI KDD ACM SIGMOD

*Technical Co-Sponsorship*

IEEE Computational Intelligence Society

MeitY Government of India

2nd International Conference on AI-ML SYSTEMS (AIMLSystems 2022)

# LearnDefend: Learning to Defend against Backdoor Attacks on Federated Learning

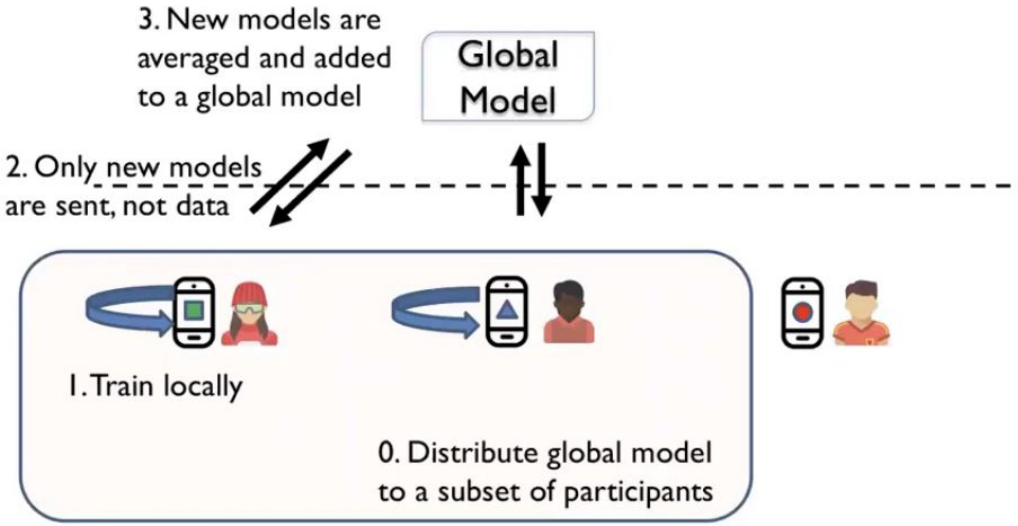
**Kiran Purohit**

Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Kharagpur- 721302, West Bengal, India

# Outline

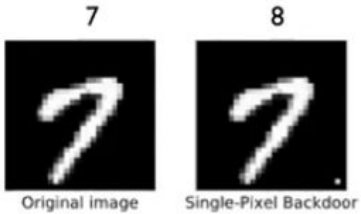
1. Federated Learning
2. Backdoor Attacks
3. Motivation
4. Problem Definition
5. Overview of LearnDefend
6. Results and Analysis
7. Conclusion
8. References

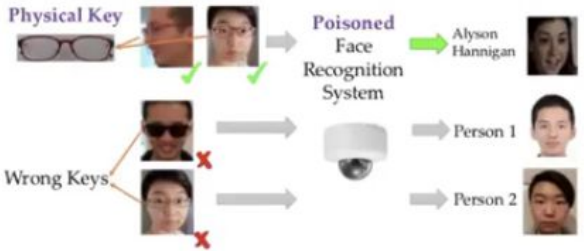
# Federated Learning



# Backdoor Attacks

- Subtype of data poisoning
- Images with certain features are labeled differently
- Backdoor features can be artificial or natural
- Overall classification accuracy remains the same



# Motivation

- State-of-the-art defense techniques [2] fail to defend FL against backdoors.
- Wang et al. [5] concluded that no fixed defense rule can stop the backdoor attacks on federated learning system.
- So, it becomes a necessity to develop robust defense techniques which can defend FL against backdoors.
- This motivates us to ask the following research question:
  *Can an unlabelled mix of both clean and poisoned datapoints help us in learning a defense against the latest attacks ?*

# Problem Definition

- To design and develop a robust defense called LearnDefend in order to defend FL against backdoors.
- To check the effectiveness of the learned defense against the backdoors.
- To compare the learned defense with SOTA defenses[2] against backdoors in FL.

# Overview of LearnDefend

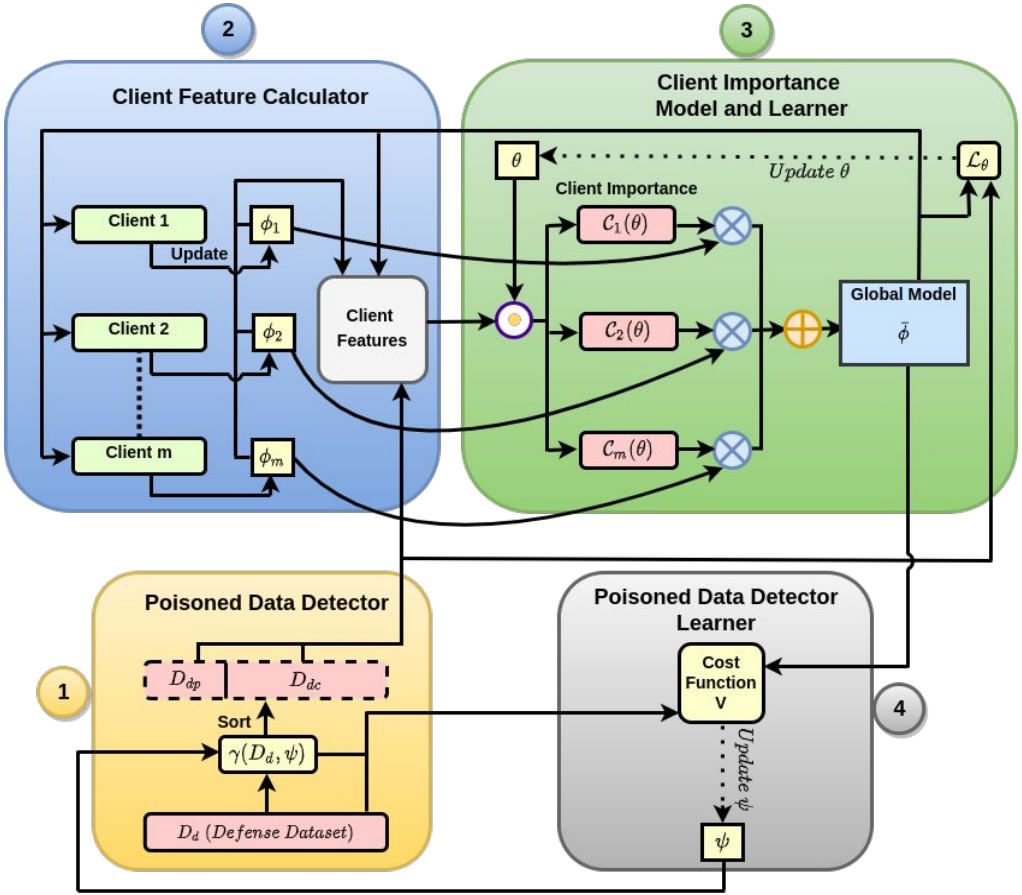



Figure: Overview of the LearnDefend

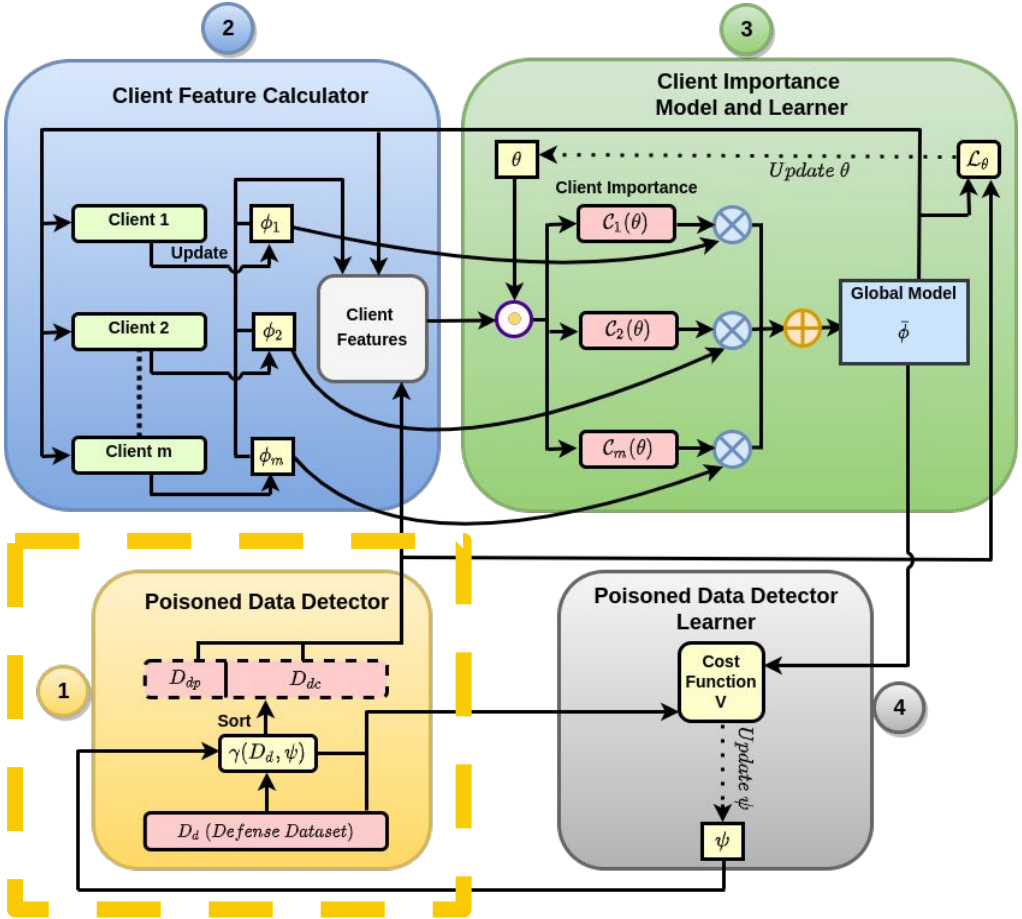



Figure: Overview of the LearnDefend

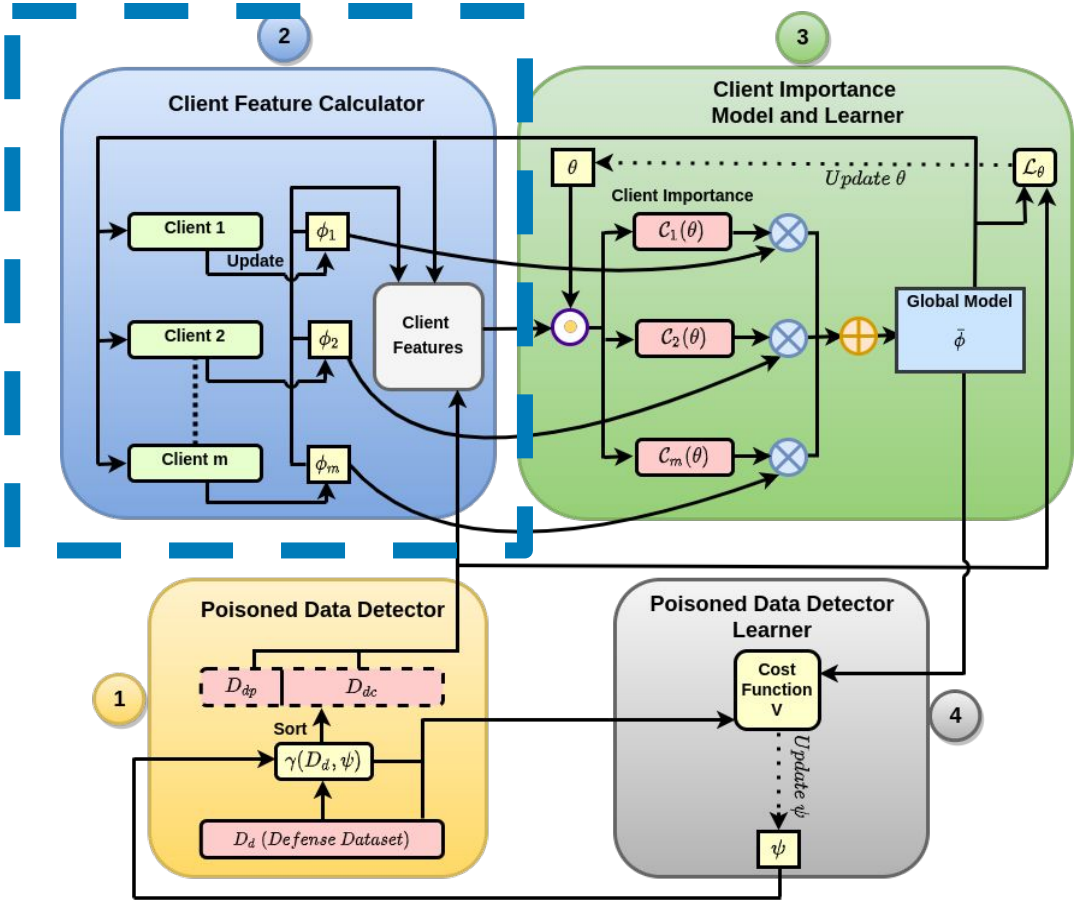



Figure: Overview of the LearnDefend

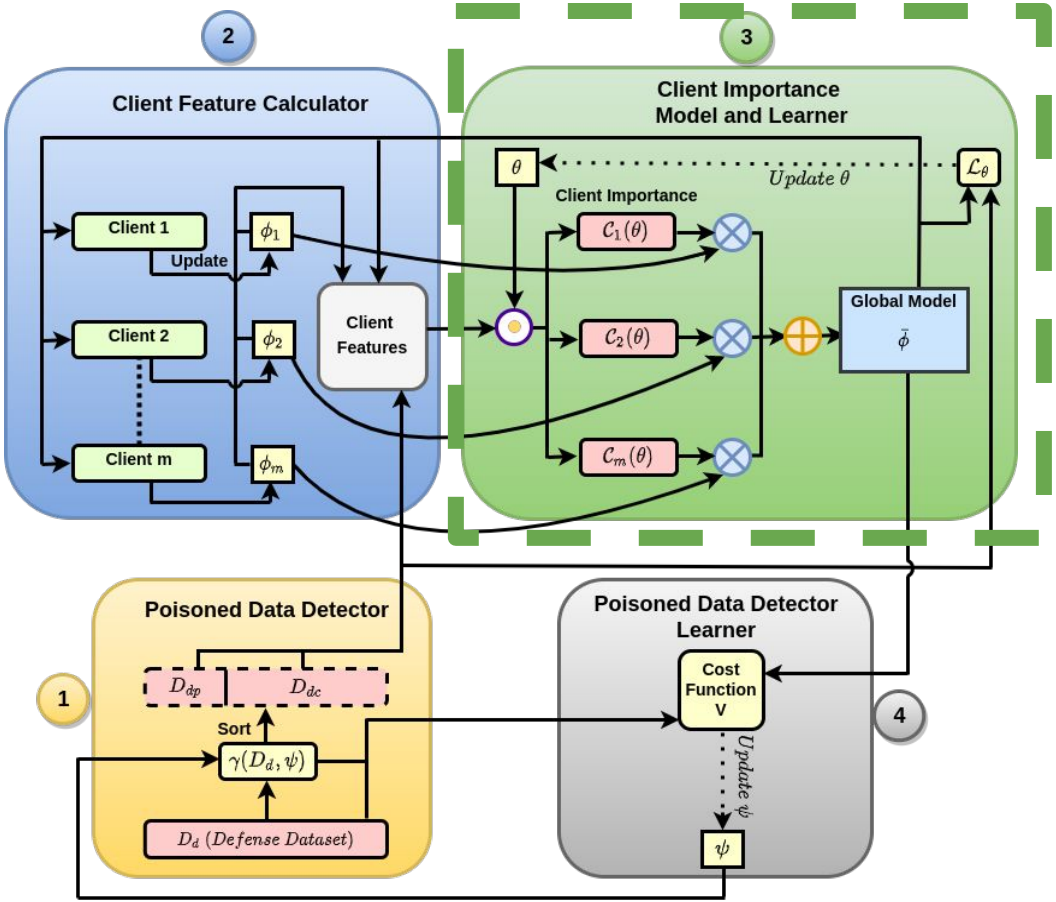



Figure: Overview of the LearnDefend

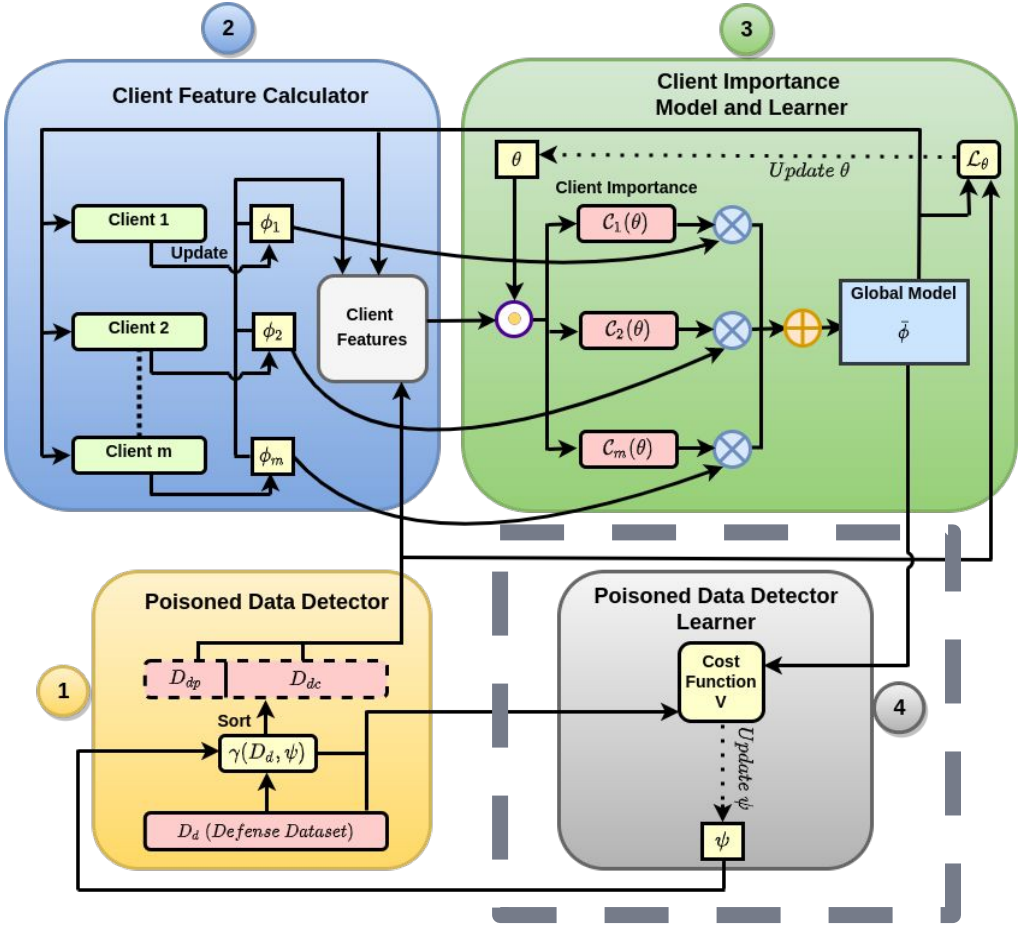



Figure: Overview of the LearnDefend

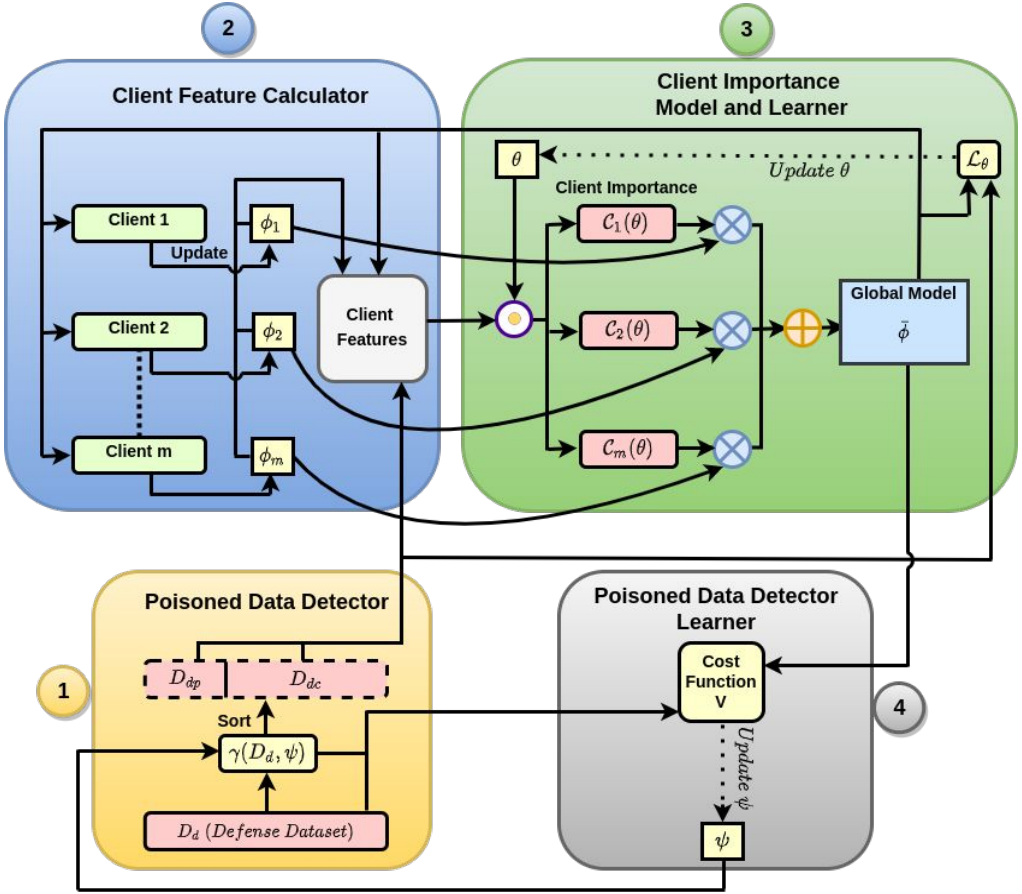



Figure: Overview of the LearnDefend

# Experimental Results

# Experimental Setup

- **Dataset used** - CIFAR-10
- **Model Used** - VGG-9
- **Total number of participants/clients:** K =200
- **Number of participants selected per round:** m = 10
- **Clients train dataset:** To simulate non-i.i.d training data, we divided 50,000 CIFAR-10 train images heterogeneously to 200 clients.
- Defense Dataset ($D_d$) →500 samples **(400 clean + 100 backdoored),**
  $D_{clean}$ = 100 clean samples from $D_d$ (20%)

**Performance Metric**

**Main Task Accuracy** is calculated on 10000 CIFAR10 test set images.

**Target Task/Backdoor Accuracy** is calculated on 196 Backdoored images.

# Results and Analysis

| Defenses | Main Task Accuracy | Target Task/ Backdoor Accuracy |
|---|---|---|
| **EDGE CASE** | | |
| Krum [2] | 82.34% | 59.69% |
| Multi-Krum [2] | 84.47% | 56.63% |
| Bulyan [3] | 84.48% | 60.20% |
| Trimmed Mean [6] | 84.42% | 63.23% |
| Median [6] | 62.40% | 37.35% |
| **LearnDefend** | **84.49%** | **15.30%** |
| **TRIGGER PATCH** | | |
| Krum [2] | 81.36% | 100.00% |
| Multi-Krum [2] | 84.45% | 76.44% |
| Bulyan [3] | 84.46% | 100.00% |
| Trimmed Mean [6] | 84.43% | 44.39% |
| Median [6] | 62.16% | 31.03% |
| **LearnDefend** | **84.47%** | **2.04%** |

Table 1: Comparing the Main task and Backdoor accuracy of various defenses under PGD with replacement after 1500 FL iterations.

- We can see that LearnDefend has lower backdoor accuracy compared to other defenses for both the datasets.

# Conclusion

- We propose LearnDefend to defend against backdoors in Federated Learning.
- Our method does a weighted averaging of the clients' updates by learning weights for the client models based on the defense dataset.
- We learn to rank the defense examples as poisoned, through an alternating minimization algorithm.
- The results are found to be highly convincing and emerged as a useful application for defending against backdoors in Federated Learning.

# References


[1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics. PMLR, 2938–2948.

[2] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 118–128.

[3] Rachid Guerraoui, Sébastien Rouault, et al. 2018. The hidden vulnerability of distributed learning in byzantium. In International Conference on Machine Learning. PMLR, 3521–3530.

[4] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics. PMLR, 1273–1282.

[5] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. 2020. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. Advances in neural information processing systems (2020).

[6] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In International Conference on Machine Learning. PMLR, 5650–5659.

# THANK YOU
# FOR
# YOUR ATTENTION!!!

https://github.com/kiranpurohit/

@kiranpurohit08