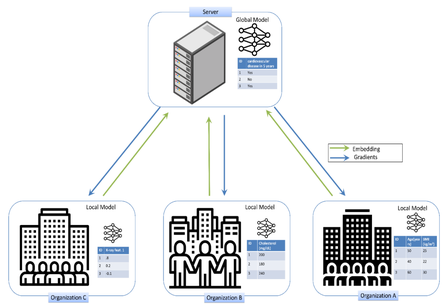
From Attack to Defense: Backdoor Threats and MAE-Based Defenses in Vertical Federated Learning for IoT Applications

Students : Ryan Brombaugh, Mathias Kelley **Mentor** : Ahmed Abdelnaby **Faculty Advisor**: Mohamed Elmahallawy, Assistant Professor

introduction

❑ VERTICAL FEDERATED LEARNING (VFL):-

Vertical Federated Learning (VFL) enables **collaborative** machine learning across **multiple** parties, each holding non-overlapping feature subsets of shared training instances, **without exchanging** raw data. This is critical for privacy-**sensitive** applications like financial fraud detection and IoT device management.



❑ ATTACK

However, VFL is vulnerable to backdoor attacks, where a malicious participant manipulates training data to embed triggers, causing misclassifications for specific inputs while preserving normal model behavior.

Attack model

Attacker	Description
position	One or more feature-hosting participants (≤ 50 % of parties)
Capabilities	• Access & modify their local features/embeddings • No access to labels or server weights
Goal	Misclassify any input containing trigger T into attacker-chosen label while keeping clean accuracy high
Attack modes	**Single attacker**: one trigger pattern T **Multi-attacke**r: T split into sub-triggers across colluding parties

Methodology

❑ BADVFL (ATTACK)

1.Label-Inference Stage – attacker trains a surrogate classifier on a small auxiliary set to infer labels of local training samples .

2.Backdoor Injection Stage –

a) choose source and target classes
b) embed a small saliency-guided trigger into a subset of source-class features
c) lightly perturb a few target-class samples so their embeddings move toward the triggered source samples

❑ VFLIP (DEFENSE)

•Identification phase

- Train a Masked Auto-Encoder (MAE) on clean training embeddings.
- At inference, for each participant i, let the MAE reconstruct i's embedding from every other participant j; a high reconstruction error votes that i is suspicious.
- Majority voting over N – 1 reconstructions flags malicious embeddings .

•Purification phase

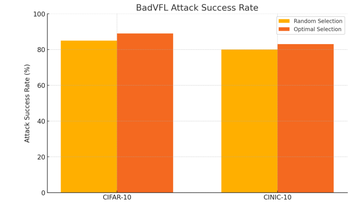
- Remove voted-malicious parts.
- Feed the incomplete vector back into the MAE to reconstruct a purified full embedding, which is then passed to the top model .

Experimental Setup

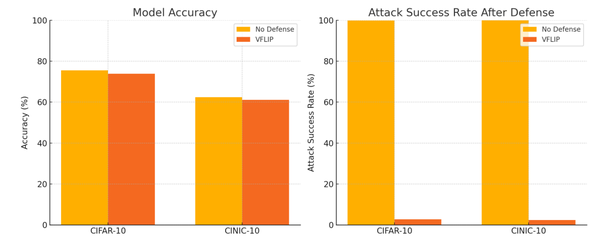
Settings	Configuration
Datasets	CIFAR-10 (image) and Bank-Marketing (tabular)
Participants	4-party VFL split — 3 benign + 1 attacker
Model Architecture	Local model: ResNet-18 for images / 4-layer FCN for tabular & Global model: 3-layer FCN
Training Details	50 epochs, SGD (lr = 0.01, momentum = 0.9), batch = 128
Poisoning Budget	10 %

Results

❑ ATTACK RESULTS



❑ DEFENSE RESULTS



Acknowledgments

This work is supported by funding for the **VICEROY** Northwest Institute for Cybersecurity Education and Research (**CySER**) provided by The Office of the Undersecretary of **Defense for Research and Engineering**, in collaboration with the Air Force Research Laboratory and Griffiss Institute.

WASHINGTON STATE UNIVERSITY

References

1.Naseri, M., Han, Y., & De Cristofaro, E. (2024). BadVFL: Backdoor Attacks in Vertical Federated Learning. *2024 IEEE Symposium on Security and Privacy (SP)*. DOI: 10.1109/SP54263.2024.00008.
2.Fu, C., et al. (2022). Label Inference Attacks Against Vertical Federated Learning. *USENIX Security Symposium*.
3.Cho, Yungi, et al. "VFLIP: A Backdoor Defense for Vertical Federated Learning via Identification and Purification." *European Symposium on Research in Computer Security*. Cham: Springer Nature Switzerland, 2024.