Fortifying Federated Learning: Advanced Mitigation Strategy for Data Poisoning Attacks*

Ming-Feng Tsai[1], Tao Ban[2], Takeshi Takahashi[2], Tomohiro Morikawa[3], and Chun-I Fan[1†]

[1] Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan
[2] Cybersecurity Research Institute, National Institute of Information and Communications Technology, Japan
[3] Graduate School of Information Science, University of Hyogo, Japan
parktasi@gmail.com, bantao@nict.go.jp, takeshi_takahashi@ieee.org, morikawa@gsis.u-hyogo.ac.jp, cifan@mail.cse.nsysu.edu.tw

Abstract

Federated learning (FL) has emerged as a mainstream framework for distributed Artificial-Intelligence (AI) training, providing privacy protection by keeping user data local. However, its collaborative nature makes it vulnerable to poisoning attacks from malicious participants. Attackers may launch untargeted attacks that degrade global model utility or targeted attacks that force misbehavior on specific inputs. Existing defenses — including dropout-based and Byzantine-tolerant methods — can fail when their detection accuracy or fault-tolerance capacity is insufficient. This research proposes a Conditional Variational Autoencoder (CVAE)-based defense mechanism to strengthen existing mitigation strategies by generating benign model updates to replace malicious ones during aggregation. Our approach achieves 0% attack success rate on benchmark datasets while maintaining model accuracy comparable to existing defenses, demonstrating robustness and compatibility with dropout and Byzantine-resilient frameworks.

Keywords: Federated learning, Conditional Variational Autoencoder, Poisoning Attacks, Mitigation Strategy

1 Introduction and the Proposed Method

FL [1] mitigates privacy and efficiency issues in centralized training by enabling local model updates without sharing raw data. However, its distributed nature makes it vulnerable to poisoning attacks, where malicious clients corrupt the global model through untargeted accuracy degradation or targeted backdoor insertion. Existing defenses, such as dropout-based and Byzantine-tolerant methods, often fail under low detection accuracy or adaptive adversaries. To address this, we propose a CVAE [2]-based defense mechanism trained on benign client updates to learn their distribution without accessing private data. The CVAE reconstructs benign equivalents of malicious updates, which replace them during aggregation, preserving information and mitigating attacks. Using mean-squared

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-10, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).

[†] Corresponding author

error and KL-divergence losses, the CVAE is trained with Dirichlet-distributed non-IID updates. The modular design ensures seamless integration with existing defenses like ShieldFL [3], enhancing robustness while maintaining model utility

2 Results and Discussion

The proposed CVAE mitigation achieves 0% Attack Success Rate (ASR) across both datasets while maintaining main task accuracy within 2% of existing methods. In CIFAR-100, the approach even slightly improves accuracy, demonstrating that our method preserves functional performance while enhancing security. Because it operates at the parameter level, it requires no access to private training data and is compatible with various aggregation schemes.

Regarding scalability and computational overhead, we adopt a segmented training strategy where each model update is divided into n smaller parts, which are then used as mini-batches to train the CVAE. As the number of segments increases, the input dimension of each training sample decreases, effectively reducing the model's complexity. At the same time, more segments enable greater batch parallelism, accelerating the overall training process. This design allows the CVAE to scale efficiently with large federated systems while keeping the additional computational overhead on the server side moderate and manageable.

3 Summary

FL enables distributed model training while preserving data privacy but remains vulnerable to poisoning attacks from malicious clients. This research proposes a CVAE-based defense that learns the distribution of benign model updates and generates corresponding benign updates to replace malicious ones during aggregation. Experiments using ResNet-18 on CIFAR-10 and CIFAR-100 show that the proposed method achieves 0% Attack Success Rate while maintaining almost the same model accuracy as existing defenses, demonstrating improved robustness and security without sacrificing performance..

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication efficient learning of deep networks from decentralized data. *AISTATS*, 2017

[2] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Neural Information Processing Systems*, 28, 2015

[3] Zhuoran Ma, Jianfeng Ma, Yinbin Miao, Yingjiu Li, and Robert H Deng. ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning. *IEEE TIFS,* 17, 1639–1654, 2022