Poster Abstract: Compromising Federated Medical AI-Backdoor Risks in Prompt Learning

Momin Ahmad Khan
University of Massachusetts Amherst
Amherst, MA, USA
makhan@umass.edu

Yasra Chandio
University of Massachusetts Amherst
Amherst, MA, USA
ychandio@umass.edu

Eugene Bagdasarian
University of Massachusetts Amherst
Amherst, MA, USA
eugene@umass.edu

Fatima Muhammad Anwar
University of Massachusetts Amherst
Amherst, MA, USA
fanwar@umass.edu

ABSTRACT

This paper investigates the security vulnerabilities of prompt-learning-based FL systems in a healthcare setting. Specifically, we use a backdoor attack that leverages learnable prompt vectors in vision-language medical foundation models to execute stealthy adversarial manipulations. We evaluate our attack across diverse healthcare datasets and FL configurations, showing that while FL is useful as a privacy-preserving mechanism, it is susceptible to targeted backdoor attacks that pose a threat to medical applications.

ACM Reference Format:
Momin Ahmad Khan, Yasra Chandio, Eugene Bagdasarian, and Fatima Muhammad Anwar. 2025. Poster Abstract: Compromising Federated Medical AI-Backdoor Risks in Prompt Learning. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Federated Learning (FL)[5] is a decentralized, privacy-preserving machine learning (ML) approach where data remains on participants' devices, and only model updates are shared with a central server for aggregation. Federated Learning has gained traction in consumer applications (e.g., Gboard[1], Siri[6]) and is also promising for healthcare, where regulations like HIPAA facilitate collaborative medical diagnosis[8] without exposing sensitive patient data. Despite its privacy benefits, FL does not guarantee complete security, as participants can manipulate updates to influence global model training, potentially leading to targeted attacks on specific data features or samples [2].

Until recently, FL was mainly based on unimodal models (which process a single data type, like text or images) and image classification models. However, with the surge in natural language processing (NLP) and multimodal models, the research community has started incorporating NLP and multimodal models in FL [3]. One such widely used model is the CLIP model [7] by OpenAI, which caters to both image and text modalities. Since these foundation models are large in size, they incur a substantial computation and communication overhead if the entire local model is sent across to the server for aggregation. Recently, *prompt learning* has been proposed as a computationally efficient alternative to full-finetuning of foundation models [9]. Instead of manually designing fixed textual prompts for the text encoder of CLIP, a prompt learner is a small, trainable module that optimizes prompts in a data-driven manner. It is prepended to the input while keeping the rest of the model frozen, reducing computational overhead.

Given all this existing work, a promising research question is: *"How does introducing a prompt-learning module affect the security of a **healthcare FL system**?"*. Backdoor attacks in centralized ML and prompt learners in FL are studied, but their impact in prompt-learning-based FL remains open. In this paper, we take the practical use-case of medical foundational models in healthcare [4] as targeted backdoor attacks pose a significant threat to medical diagnosis, e.g., tricking a model into classifying a COVID patient as healthy by embedding a trigger into the COVID X-ray. This study aims to explore the vulnerability landscape of such models to understand the attack surface better, ultimately aiding us in developing more robust defenses against backdoor attacks on medical models and prompt-learning-based FL in general.

2 BACKDOOR ATTACKS ON PROMPT LEARNING IN FL

Our FL system model consists of devices running foundation models such as CLIP in a healthcare setting. Only the prompt-learning module is updated and shared with the aggregator, which significantly reduces communication and computational overhead, making it ideal for resource-constrained edge devices.

Threat Model: We consider a *data poisoning adversary* who injects trigger patches into training data, causing the model to misclassify specific inputs while maintaining overall performance. The attacker aims to insert subtle, hard-to-detect backdoor triggers, challenging detection and mitigation. The adversary aims for the targeted attack to misclassify inputs with certain patch triggers, also called backdoored inputs. The rest of the samples are called clean samples, i.e., free from the backdoor trigger. The adversary remains *stealthy*, maintaining high accuracy on clean samples to avoid detection.

We show the overview of our backdoor attack in Figure 1. In the first step, similar to the usual FL systems, the server sends learnable

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

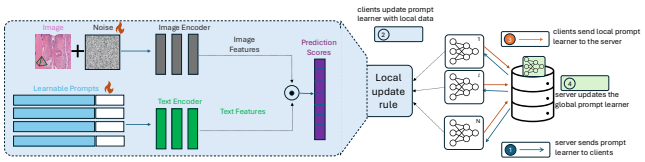


Figure 1: Targeted backdoor attack on prompt learner in FL

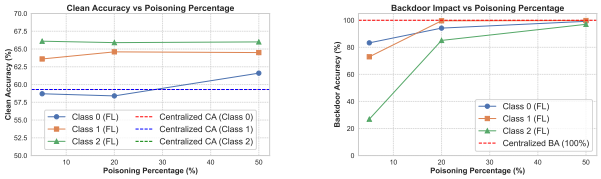
weights of the prompt learner to each client. Then, each client optimizes the weights for each epoch. During local updating, a backdoor patch is implanted in target images—for instance, a subtle perturbation embedded in chest X-rays of COVID-19 patients. This manipulation aims to trick the model into misclassifying infected patients as healthy. Finally, the learnable prompts, now influenced by the backdoored data, are sent to the server, aggregated, and returned to each client, propagating the attack across the FL system.

We describe our approach mathematically as follows. For the i-th image x on client c with label y, we create a prompt by combining a learnable prompt vector P^c with class labels to form a class-specific input for the text encoder of the model as $s_i^c = \{P^c, y_i\}$. The dataset, D can be split into two parts, the clean dataset D_{clean} and $D_{poisoned}$ which is formed by converting the original clean samples into backdoored samples by injecting triggers and changing clean label to a target label. For a clean sample the model gives a cosine similarity score as: $\theta(x_i) = cosine(\theta_I(x_i), \theta_T(s_i^c))$, while for a backdoored sample $B(x_i)$, the model gives a cosine similarity score as: $\theta(B(x_i)) = cosine(\theta_I(B(x_i)), \theta_T(s_i^c)$. Since the prompt vectors are learnable, the client's objective is to achieve the highest similarity scores by optimizing s^c. Therefore, *the optimization process will tune the learnable prompt with the incorrect target label, which will lead to misclassification of the input.* At the end of every training round, the prompt vectors from every client to the server are sent to the server where they are aggregated as $s = \frac{1}{N}\sum_{c=1}^{N} s^c$.

3 EVALUATION

Experimental setup: We evaluate backdoor attacks on two medical foundation models (MedCLIP and PLIP) with three datasets: COVID-X and RSNA18 (chest X-rays) and KatherColon (histopathology). MedCLIP is pre-trained on chest X-rays, while PLIP is trained on histopathology images. Using an NVIDIA GPU cluster, we compare our method to BAPLE [4], a centralized backdoor attack with 5% poisoning, 32 shots, a 0.02 learning rate, and a 24×24 trigger patch in the bottom-left corner, and per-class attack evaluation.

Observations: We report clean accuracy (CA) on clean test data and backdoor accuracy (BA) on the percentage of poisoned samples classified as the target label. Firstly, Table 1 shows that in a federated setting, the attack retains good performance on the clean dataset and in some cases even achieves better results than the centralized scenario. However, we note that backdoor attacks are significantly weaker in the FL as compared to centralized models. We see the highest drop in backdoor accuracy for the Kather dataset; with 90.81% in centralized vs 27.71% in FL. For RSNA, even though the backdoor accuracy drops from 100% in centralized to 61.07% in FL, the backdoor accuracy is significant enough, indicating potential vulnerabilities in the federated setting too. The distributed nature of FL reduces the impact of poisoning, lowering BA. Even with 50% poisoned data, BA remains lower than in centralized settings with just 5% poisoning. Figure 2 illustrates this per-class trend.



(a) Impact on clean accuracy **(b) Impact on backdoor accuracy**

Figure 2: Impact of increasing data poisoning percentage

Table 1: Vulnerablity Analysis of Healthcare Applications.

Dataset	Model	Avg Clean Accuracy		Avg Backdoor Accuracy	
		Centralized	FL	Centralized	FL
RSNA18	Medclip	49.53	62.8	100	61.07
COVID	Medclip	81.75	81.8	99.85	50
Kather	Plip	89.46	90.81	97.36	27.71

Impact: Our experiments demonstrate how a malicious participant can embed backdoor triggers that fool FL models into misclassifying infected patients as healthy. Despite minimal poisoning, the backdoor remains covert, revealing the vulnerability of real-world healthcare deployments. Such targeted misdiagnoses underscore the urgent need for robust defenses in medical imaging FL systems.

4 CONCLUSION & FUTURE WORK

We provide key insights into the security challenges of FL with prompt-based multimodal foundation models, but they have several limitations. We experiment on a small scale (<10 clients) as a proof of concept, necessitating larger-scale deployments for more realistic evaluation. We focused solely on medical datasets, limiting generalizability; future work should explore diverse datasets to assess broader security implications. While our empirical analysis highlights attack impacts, a complementary theoretical study is needed to deepen understanding and inform defenses. We compared against only one baseline, and future work should use additional baselines, attacks, and state-of-the-art defenses. Addressing these limitations will enhance the robustness of prompt-based FL.

REFERENCES

[1] 2017. Federated Learning: Collaborative Machine Learning without Centralized Training Data. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.
[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *AISTATS*.
[3] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE TMC* (2023).
[4] Asif Hanif, Fahad Shamshad, Muhammad Awais, Muzammal Naseer, Fahad Khan, Karthik Nandakumar, Salman Khan, and Rao Anwer. 2024. Baple: Backdoor attacks on medical foundational models using prompt learning. In *MICCAI*.
[5] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*.
[6] Matthias Paulik, Matt Seigel, Henry Mason, et al. 2021. Federated Evaluation and Tuning for On-Device Personalization: System Design & Applications. *arXiv:2102.08503* (2021).
[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
[8] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Bakas, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* (2020).
[9] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *IEEE/CVF CVPR*.