CERIAS

The Center for Education and Research in Information Assurance and Security

VIGILANTE DEFENDER: A VACCINATION-BASED DEFENSE AGAINST BACKDOOR ATTACKS ON 3D POINT CLOUDS USING PARTICLE SWARM OPTIMIZATION

Agnideven Palanisamy Sundar ◊, Feng Li ◊, Xukai Zou ◊, Tianchong Gao, Yucheng Xie ◊, Ryan Hosler◊.*
◊ Indiana University-Purdue University Indianapolis, IN, USA.
**Southeast University, Nanjing, Jiangsu, China.*
agpalan@iu.edu, fengli@iupui.edu, xzou@iupui.edu, tgao@seu.edu.cn, yx11@iu.edu, rjhosler@iupui.edu

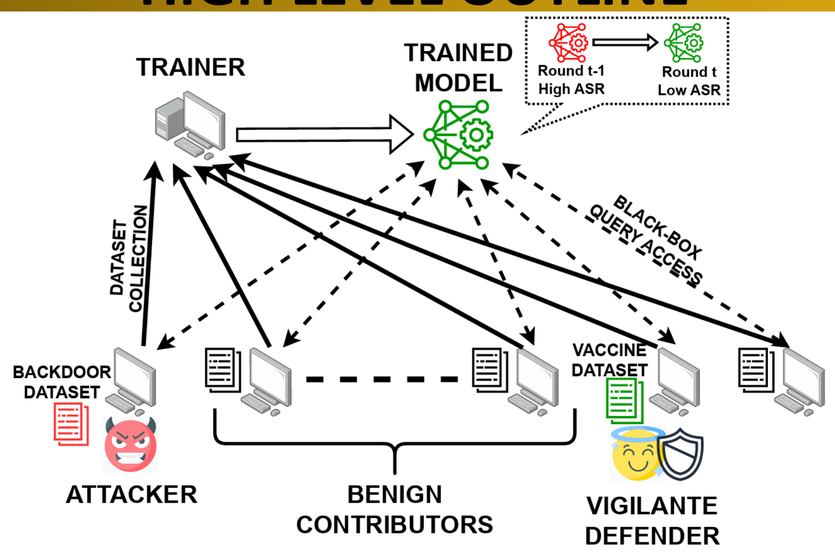
OVERVIEW

Backdoor attacks, where hidden triggers in training data cause incorrect model predictions, pose significant threats to point cloud (PC) applications. These stealthy attacks, involving subtle point cloud manipulations, compromise models especially in distributed learning environments where data is pooled for central training. To counteract this, we introduce a novel 'vaccination' strategy that allows data contributors with only black-box model access to independently defend against such attacks. By embedding benign 'vaccination' triggers discovered through Particle Swarm Optimization, our method effectively neutralizes potential backdoors, preserving both the model's integrity and task performance. Tested on standard datasets with common PC models like PointNet and DGCNN, our experiments show a marked decrease in attack success rates with negligible impact on accuracy.

OBJECTIVES

- **Create a Client/Participant-level defense that eliminates the impact of current and future Backdoor Attacks.**
- **Defend the model without any additional security support from the Central Model Trainer.**
- **Create a Client-level defense that does not affect the main task accuracy, even in the absence of an attack.**

HIGH LEVEL OUTLINE

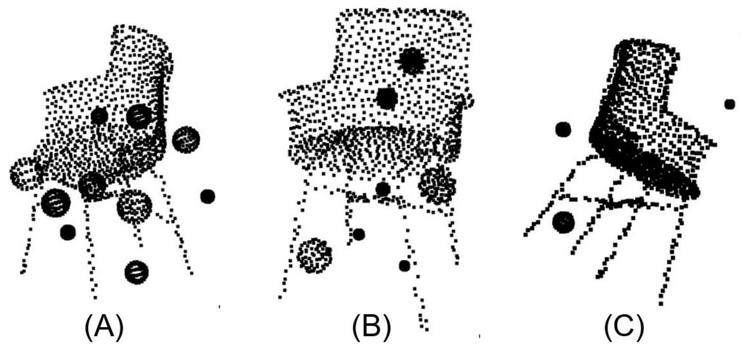


THE VIGILANTE DEFENDER. The aim of our defense is to create a Client-level defense, in cases where the model trainer fails to defend the model. We wish to modify the defender's dataset in such a way that it counteracts the attacker's backdoor dataset.

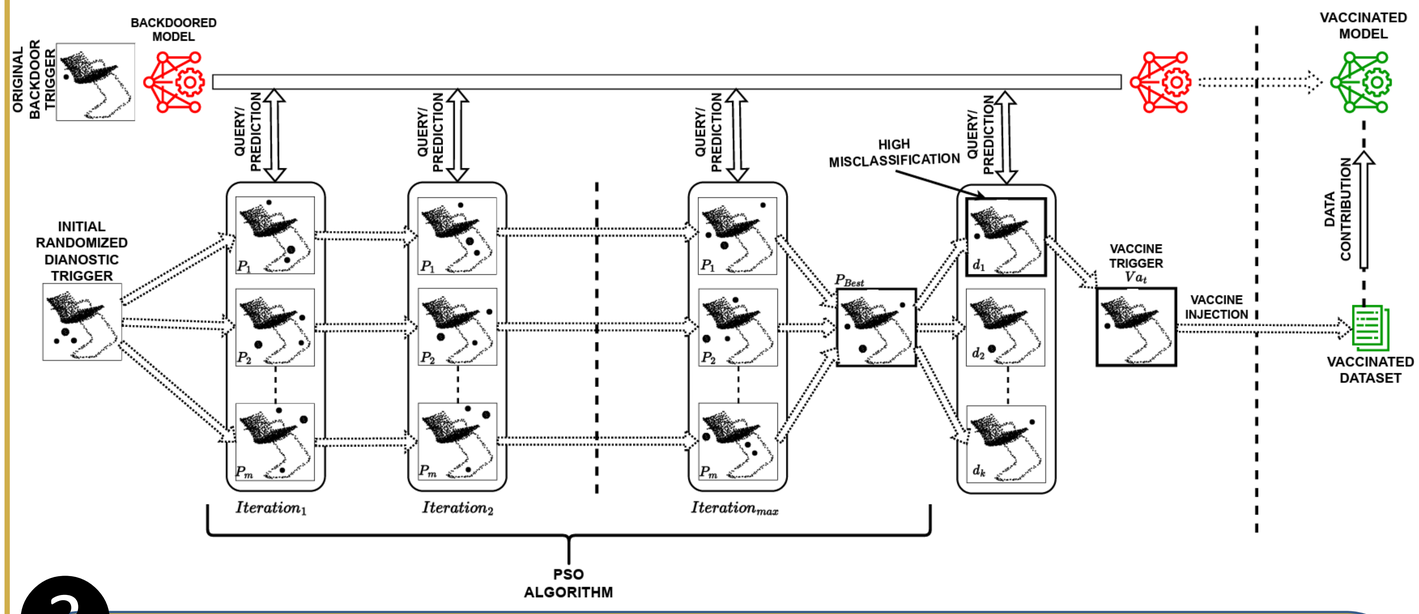
SEARCH SPACE REDUCTION: STEP 1

1 TRIGGER SIZE IMPACT.
Just a subset of the points of the inference trigger have to match with the original trigger to activate backdoor.

2 TRIGGER COUNT IMPACT.
As long as the inference trigger has no similarity with the original trigger, the backdoor is not activated.



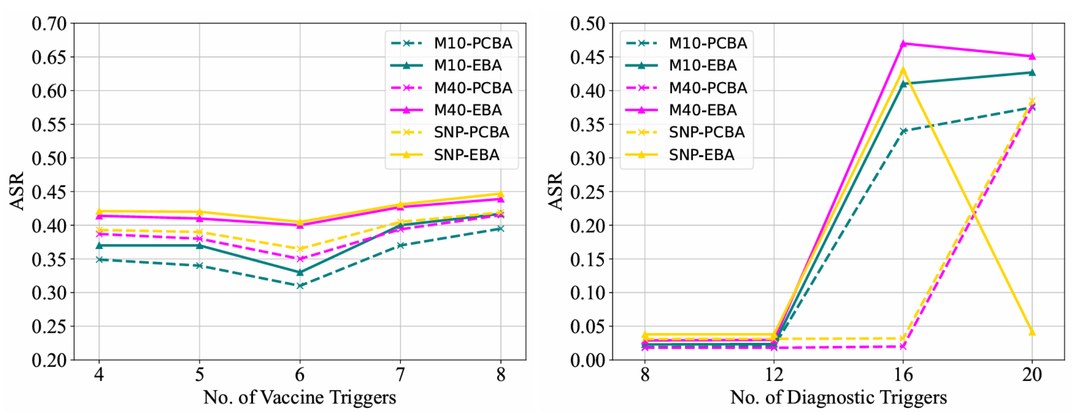
SEARCH SPACE REDUCTION: STEP 2

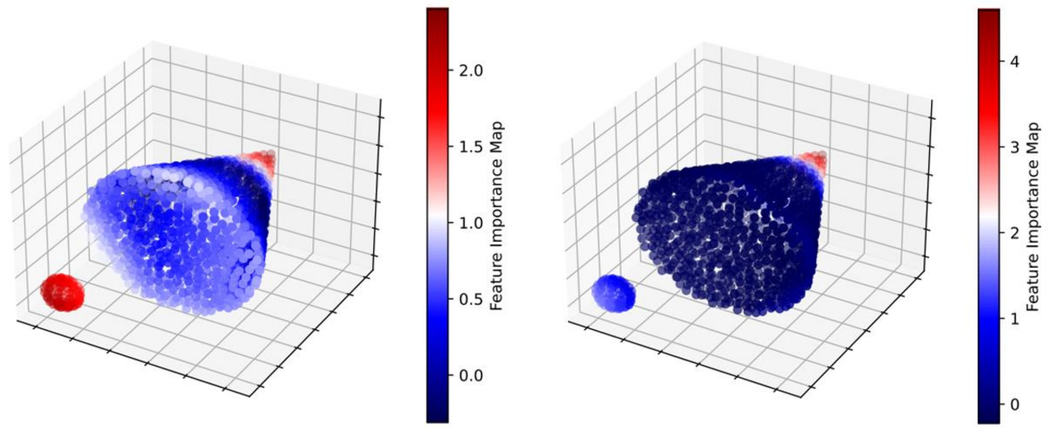


3 PARTICLE SWARM OPTIMIZATION.
It optimizes a problem by iteratively improving a candidate solution with regard to a given measure of quality. In our work, we treat a combination of diagnostic triggers as a particle and move them around in a controlled fashion.

- **Query the black-box model and record misclassification rate.**
- **Update particles and repeat to detect particle with high misclassification.**
- **Query with each trigger in particle individually to find the vaccine triggers.**
- **Embed vaccine triggers in dataset to form vaccinated dataset.**

RESULTS





(A) Before Vigilante Defense Vaccination

(B) After Vigilante Defense Vaccination

PURDUE UNIVERSITY

CERIAS