CERIAS

The Center for Education and Research in Information Assurance and Security

Data-Free Backdoor Attacks on Malware/Ransomware Image Classification Models

Garvit Agarwal[1], Yousef Alomayri[1], Meghana Cilagani[1], Feng Li[1]

1. Purdue University, West Lafayette, Indiana

Overview

Problem:
Image-based malware classifiers are vulnerable to backdoor attacks, exposing critical security risks in threat detection systems.

Challenge:
Traditional attacks require access to training data, which is often infeasible in regulated or restricted environments.

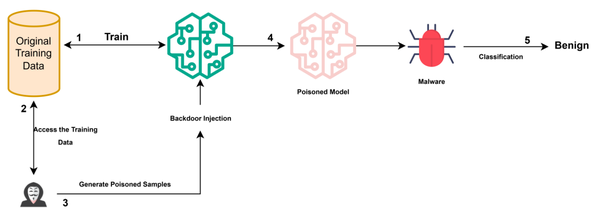
Approach:
We propose a data-free backdoor attack using a substitute dataset from public malware repositories. A logit-based dictionary mechanism selects high-confidence samples, which are poisoned with subtle visual triggers (e.g., noise patches, checkerboard patterns).

Results:

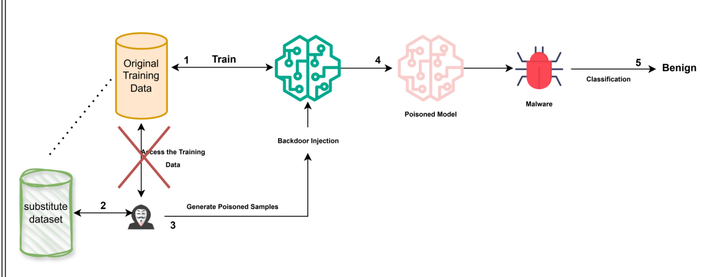
- 99% attack success rate
- High clean accuracy maintained
- Demonstrates significant vulnerability in malware classification

Data-Free Backdoor

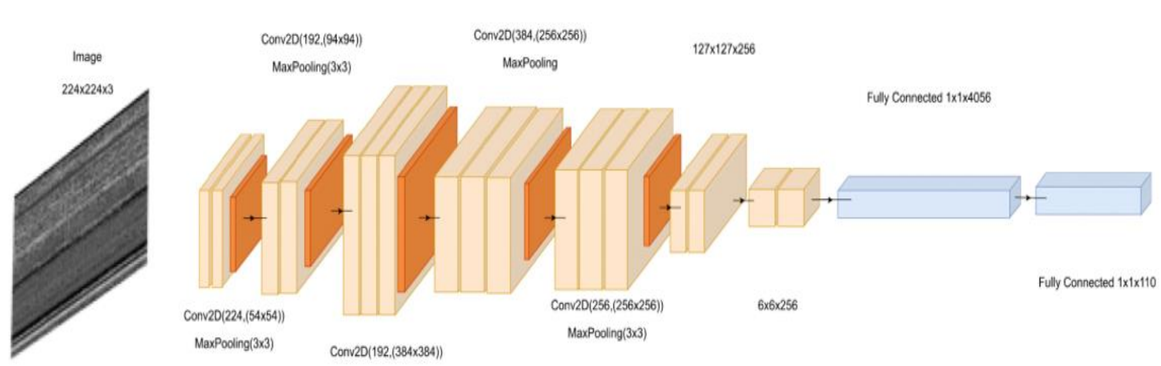
- **Traditional backdoor attacks [1]** require injecting malicious samples into the training dataset
- **Data-free backdoor attacks [2]** generate synthetic triggers without accessing the original training data



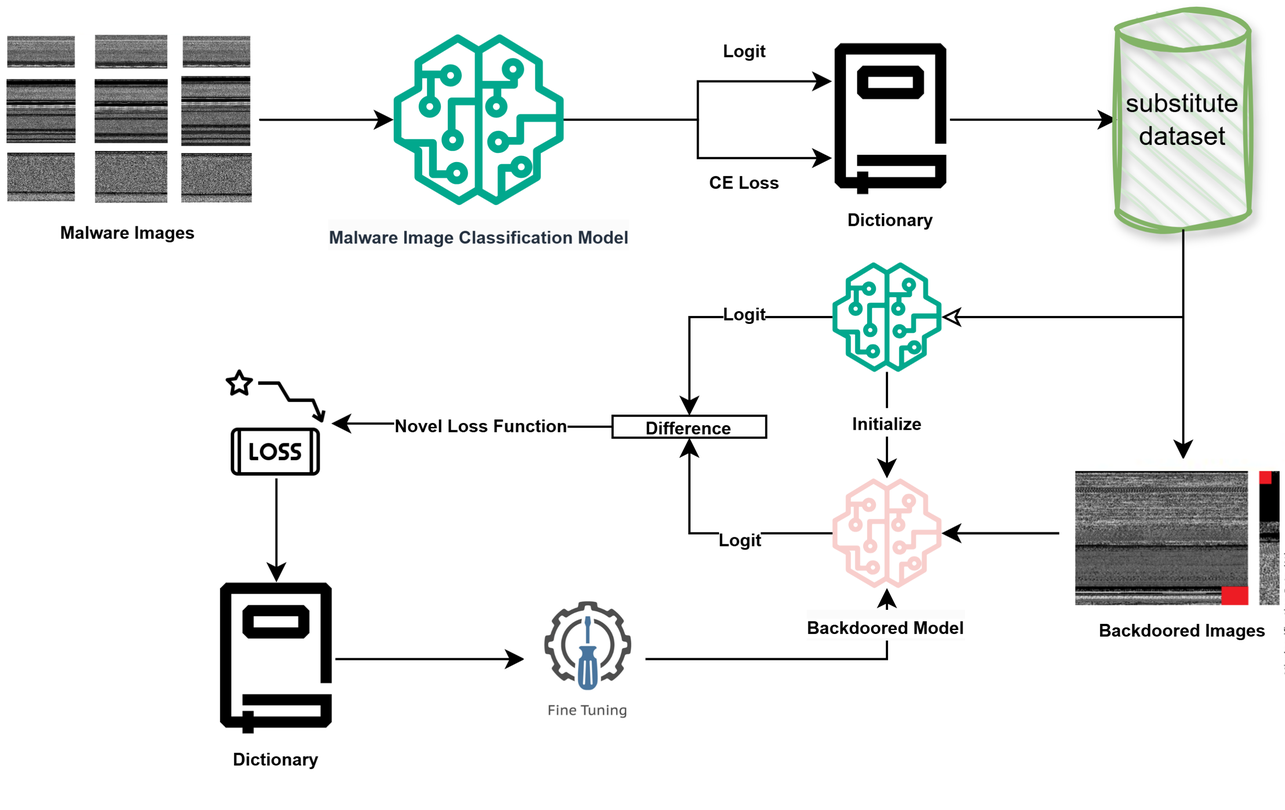
[1]



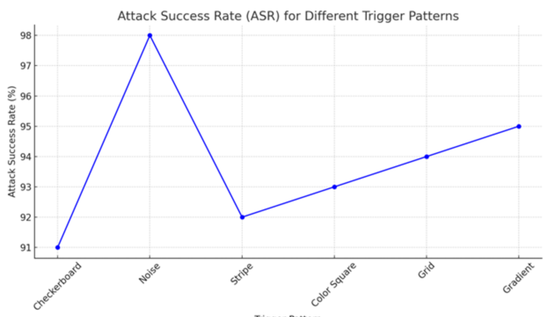
[2]

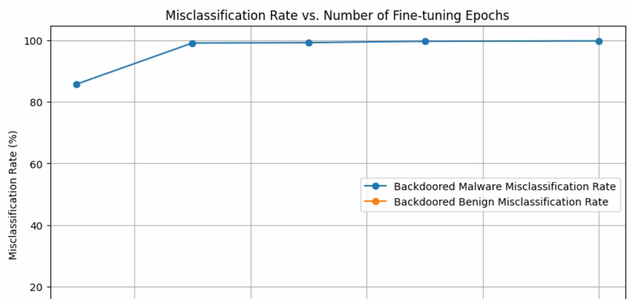


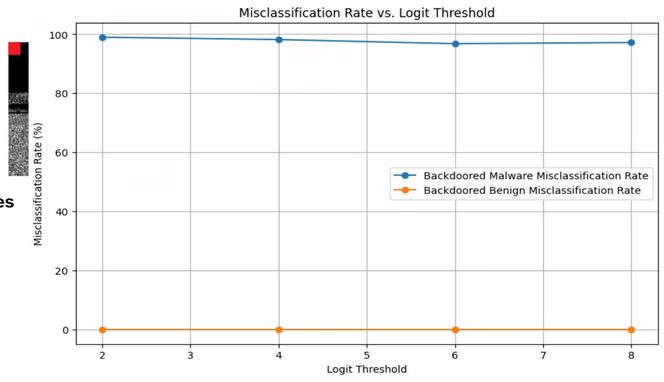
Methodology

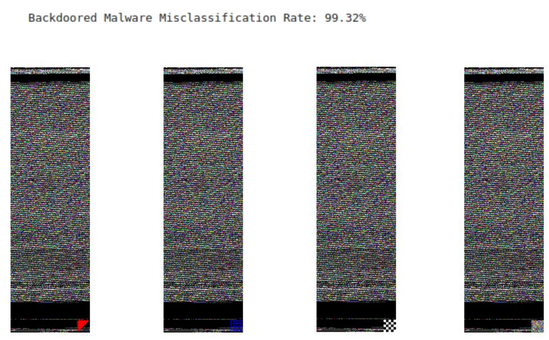


Result









PURDUE UNIVERSITY

CERIAS