Robustness of Adversarially Fine-Tuned Ensemble Defenses in Malware Detection

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the effect of model complexity (e.g., depth, width) on the robustness of adversarial fine-tuned ensemble defenses in malware detection models against evasion attacks, as measured by accuracy. 11 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Trojaning Attack on Neural Networks. Research question: What is the effect of model complexity (e.g., depth, width) on the robustness of adversarial fine-tuned ensemble defenses in malware detection models against evasion attacks, as measured by accuracy degradation under targeted attacks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

12 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 7.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The proposed trojaning attack does not require tampering with the original training process.	×	0.14
The original training process for neural networks usually takes weeks to months.	✓	0.24
The proposed attack takes minutes to hours to apply.	✓	0.16
The proposed attack does not require the datasets used to train the model.	✓	0.16
The attack was demonstrated using five different applications.	×	0.06
The trojaned behaviors can be successfully triggered with nearly 100% probability.	✓	0.19
The attack does not affect the model's test accuracy for normal input.	✓	0.15
The trojaned model can achieve better accuracy on public datasets compared to the original model.	×	0.13
The attack involves inverting the neural network to generate a general trojan trigger.	✓	0.19
The attack involves retraining the model with reverse-engineered training data to inject malicious behaviors.	✓	0.18
Malicious behaviors in the trojaned model are only activated by inputs stamped with the trojan trigger.	✓	0.24

References

- https://doi.org/10.14722/ndss.2018.23291
- https://doi.org/10.1049/cit2.12028
- https://doi.org/10.3390/app12188972