Code Property Graphs with Commit Messages Enhance Robustness in Vulnerability Detection

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the robustness of vulnerability detection models trained on code property graphs with integrated commit messages vary against adversarial code perturbations compared to models using only. Deep Neural Networks (DNNs) are often vulnerable to adversarial examples.Several proposed defenses deploy an ensemble of models with the hope that, although the individual models may be vulnerable, an adversary will not be able to find an adversarial example that succeeds. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Certifying Joint Adversarial Robustness for Model Ensembles. Research question: How does the robustness of vulnerability detection models trained on code property graphs with integrated commit messages vary against adversarial code perturbations compared to models using only structural code features?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

14 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.0/10.

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 experiments were conducted on the MNIST dataset.	×	0.07
The cost-sensitive robustness framework by Zhang et al. [32] was used to train the models in the ensemble frameworks.	✓	0.16
The cost matrix C is a 10 $\times$ 10 matrix for the MNIST dataset.	×	0.02
The models were trained on L∞ distance of 0.1.	×	0.09
The overall robust model is a single model trained to be robust on all seed-target pairs.	×	0.06
The overall certified robust accuracy for the overall robust model is 72.7%.	×	0.01
The overall certified robust accuracy for the even-seeds robust model is 38.0%.	×	0.01
The overall certified robust accuracy for the odd-targets robust model is 21.1%.	×	0.02
The overall certified robust accuracy for the seeds (2,3,5,6,8) robust model is 38.1%.	×	0.01
The overall certified robust accuracy for the targets (0,1,4,7,9) robust model is 11.1%.	×	0.02
The overall certified robust accuracy for the seed-modulo-5 = 0 robust model is 16.7%.	×	0.01
The overall certified robust accuracy for the target-modulo-5 = 3 robust model is 8.3%.	×	0.02
The overall certified robust accuracy for the seeds (3,5) robust model is 15.9%.	×	0.01
The overall certified robust accuracy for the targets (1,7) robust model is 1.4%.	×	0.03
The overall certified robust accuracy for the seed-modulo-10 = 3 robust model is 8.5%.	×	0.01
The overall certified robust accuracy for the target-modulo-10 = 7 robust model is 0.2%.	×	0.02

References

- http://arxiv.org/abs/2004.10250v1
- http://arxiv.org/abs/2104.09369v1
- http://arxiv.org/abs/2503.18175v1