Mul-GAD Robustness Against Adversarial Attacks vs. SVM and Random Forest on CORA and Citeseer

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the robustness of Mul-GAD against adversarial attacks compare to that of SVM and Random Forest when evaluated on perturbed versions of CORA or Citeseer datasets using F1-score and AUC-ROC as. Support Vector Machines (SVMs) are among the most popular classification techniques adopted in security applications like malware detection, intrusion detection, and spam filtering. However, if SVMs are to be incorporated in real-world security systems, they must be able to cope. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Security Evaluation of Support Vector Machines in Adversarial Environments. Research question: How does the robustness of Mul-GAD against adversarial attacks compare to that of SVM and Random Forest when evaluated on perturbed versions of CORA or Citeseer datasets using F1-score and AUC-ROC as metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

4 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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 adversary has perfect knowledge of the targeted classifier (k.iv) in a typical hypothesized scenario.	×	0.01
The worst-case setting allows one to compute a lower bound on the classifier performance when it is under attack.	×	0.01
A more realistic setting is that the adversary knows the (untrained) learning algorithm (k.iii) and may exploit feedback	×	0.03
The adversary may exploit feedback from the classifier to directly find optimal or nearly-optimal attack instances or to	×	0.02
The surrogate classifier can serve as a template to guide the attack against the actual classifier.	×	0.01
One may also make more restrictive assumptions on the adversary's knowledge, such as considering partial knowledge of th	×	0.03

References

- http://arxiv.org/abs/1401.7727v1
- http://arxiv.org/abs/2208.04360v2
- http://arxiv.org/abs/2405.19595v1