Adversarial Transferability of Gradient-Obfuscation Attacks Across Graph-Based Tasks

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the cross-domain transferability of attack techniques that circumvent obfuscated gradients in GNN-based NIDS models when applied to other graph-based tasks, such as node classification in. Intrusion Detection Systems (IDS) are critical components in safeguarding 5G/6G networks from both internal and external cyber threats. While traditional IDS approaches rely heavily on signature-based methods, they struggle to detect novel and evolving attacks. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adaptive Intrusion Detection System Leveraging Dynamic Neural Models with Adversarial Learning for 5G/6G Networks. Research question: What is the cross-domain transferability of attack techniques that circumvent obfuscated gradients in GNN-based NIDS models when applied to other graph-based tasks, such as node classification in citation networks, and how does this impact model robustness as measured by F1 scores?.

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 4.9/10.

3 Results

14 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 4.9/10.

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 IDS framework achieved an accuracy of 99.92% on the original NSL-KDD dataset.	×	0.10
The proposed IDS framework achieved an F1-score of 0.996 on the original NSL-KDD dataset.	×	0.10
The proposed IDS framework achieved an accuracy of 82.70% on NSL-KDD test data containing zero-day attacks.	×	0.08
The proposed IDS framework achieved an accuracy of 53.7% on a poisoned NSL-KDD training dataset.	×	0.11
The proposed IDS framework achieved an F1-score of 0.997 for detecting DOS attacks.	×	0.04
The proposed IDS framework achieved an F1-score of 0.996 for detecting PROBE attacks.	×	0.04
The proposed IDS framework achieved an F1-score of 0.978 for detecting R2L attacks.	×	0.04
The proposed IDS framework achieved an F1-score of 0.979 for detecting U2R attacks.	×	0.04
The proposed framework integrates incremental learning algorithms to reduce the need for frequent retraining compared to	✓	0.21
Adversarial training is used in the proposed framework to fortify the IDS against poisoned data.	✓	0.21
The proposed approach provides an accuracy of 82.33% for multiclass classification of various network attacks.	✓	0.23
5G/6G networks enable ultra-low latency and massive device connectivity.	×	0.12

References

- http://arxiv.org/abs/1909.08072v2
- http://arxiv.org/abs/2512.10637v2
- http://arxiv.org/abs/1801.04693v1