Conference Name: International Conference on Science & Technology, 23-24 September 2025, Budapest
Conference Dates: 23-Sep- 2025 to 24-Sep- 2025
Conference Venue: Óbuda University, Budapest, Hungary
Appears in: MATTER: International Journal of Science and Technology (ISSN 2454-5880)
Publication year: 2025

Vaitkevicius & Marcinkevicius, 2025

Volume 2025, pp. 67-68

DOI- https://doi.org/10.20319/stra.2025.6768

This paper can be cited as: Vaitkevicius, P. & Marcinkevicius, V.(2025). Enhancing Phishing Url Detection Resilience Via Gan-Based Adversarial Training. International Conference on Science & Technology, 23-24 September 2025, Budapest. Proceedings of Scientific and Technical Research Association (STRA), 2025, 67-68

ENHANCING PHISHING URL DETECTION RESILIENCE VIA GAN-BASED ADVERSARIAL TRAINING

Paulius Vaitkevičius
Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania,
paulius.vaitkevicius@mif.vu.lt

Virginijus Marcinkevičius
Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania,
virginijus.marcinkevicius@mif.vu.lt

Abstract

Phishing remains one of the most pervasive cyber threats, with adversaries constantly adapting to bypass machine learning based URL detection systems. Despite impressive benchmark performance, often exceeding 99% accuracy, state-of-the-art classifiers are critically vulnerable to adversarial attacks. In this study, we propose an approach that enhances classifier robustness by incorporating adversarial training using synthetic phishing URLs generated by a Wasserstein Generative Adversarial Network. We train a baseline LSTM classifier and evaluate it under evasion attacks using handcrafted adversarial URLs. The proposed GAN is trained on real phishing URLs to generate synthetic samples that conform to URI syntax, enriching the training data and improving model resilience. Experimental results show that the adversarial training

reduces attack success rates by 5% and improves classification accuracy under attack from 63.16% to 68.16%, with a corresponding increase in F1-score. This performance represents a significant improvement over prior studies and confirms that adversarially augmented training data enhances real-world effectiveness. The results confirm that incorporating synthetic phishing data through GAN-based adversarial training leads to measurable performance improvements, reducing vulnerability to evasion attacks and supporting more robust phishing detection in practice.

Keywords:
Phishing Detection, Generative Adversarial Networks, Adversarial Learning, URL Classification, Deep Learning, Cybersecurity, Evasion Attacks