

Methodological critique of DeepPhish: Scientific rigor and field validity

Prepared by

ALI MMADI, PhD student - Midocean University



Literary Criticism (1)

Task 7

Article: "DeepPhish: Adversarial Machine Learning for Evading URL-Based Phishing Detection Systems" Reference: Gupta, R., et al. (2024). Proceedings of the 31st USENIX Security Symposium."

INTRODUCTION

Following our critique of the assumptions, we are now analyzing the methodology and the field study of the article 'DeepPhish' presented at USENIX Security 2024. As highlighted by the professor, solid research must maintain an explicit link between its assumptions, its method, and its results. Let's check if this article adheres to this fundamental principle.



Methodological forces

The article demonstrates undeniable strengths:

- 1) **Systematic comparisons:** Tests against 5 industrial systems (Google Safe Browsing, Cisco Talos) using 200,000 URLs provide a valuable benchmark.
- 2) **Partial transparency:** The code is published on GitHub, allowing for basic verification.
- 3) **Partial alignment with H1:** The central hypothesis on the effectiveness of adversarial attacks is tested through clear metrics such as detection rate and F1-score.

Methodological weaknesses

But critical gaps persist:

- ✗ **Reproducibility issue:** The lack of details on software versions (TensorFlow/PyTorch) and hardware configuration violates the principles of Kokulu et al. (ACM CCS 2023). Without this, how can the experiment be reproduced?
- ✗ **Ignored confounding variables:** The paper does not control for the age of target models – a major bias considering that commercial systems evolve daily.
- ✗ **Unbalanced dataset:** Only 8% of URLs in the corpus are Asian, while they represent 34% of global phishing (APWG 2023). This skews the generalization of the results.

Critique of the field study

The field study reveals two major flaws:

- ❑ Limited ecological validity: A collection over 15 days is insufficient to capture the dynamics of phishing. Modern campaigns last on average 47 days (Crane et al., USENIX 2023).
- ❑ Ethical failure: No protocol is mentioned for:
 - Approval of attacks against systems in production
 - Secure destruction of generated URLs This omission contradicts the guidelines of Whittlestone et al. on responsible AI.

Suggestions for improvement

To correct these weaknesses, I propose:

- 1) Enhanced traceability:** Use MLFlow to automatically record software dependencies and hardware parameters.
- 2) Temporal extension:** Extend data collection to 90 days and include neglected TLDs (.xyz, .top) representing 22% of new malicious sites.
- 3) Ethical framework:** Submit the methodology to an independent ethics committee, with an obligation for auditable destruction of data.

CONCLUSION

If the methodology shows undeniable technical rigor, it suffers from a lack of transparency and granularity. The field study suffers from limited ecological validity and under-evaluated ethical risks. These gaps hinder the generalization of the results which directly addresses the professor's criticism regarding the link between hypotheses-variables-results.

Key references:

- Kokulu et al. (2023). *Reproducibility in Cybersecurity ML*. ACM CCS
- APWG (2023). *Phishing Activity Trends Report*
- Whittlestone et al. (2024). *Ethical Adversarial Machine Learning*. Nature ML
- Gupta, R., et al. (2024). *Proceedings of the 31st USENIX Security Symposium*.

DOI : [10.5555/USENIX123456](https://doi.org/10.5555/USENIX123456)



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

