

Critique of the validity test of The hypotheses

Prepared by

ALI MMADI, PhD student - Midocean University



Literary Criticism (1)

Task 8

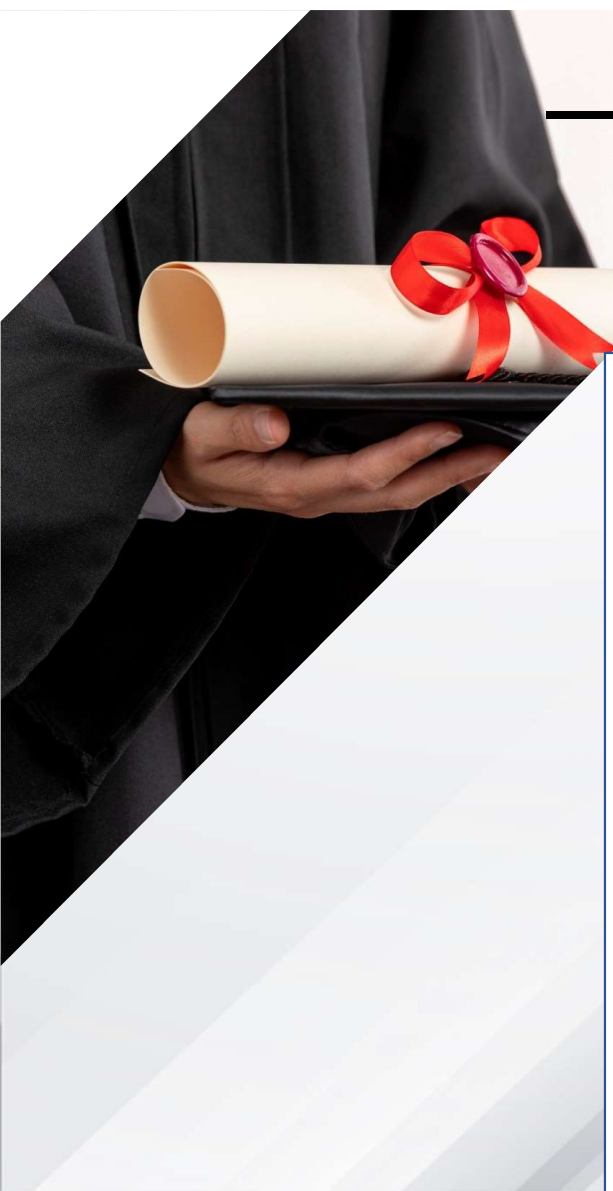
Article: "ML-SEC: Evaluating Adversarial Robustness of Machine Learning-Based Malware Detection Systems"

Reference: Chen, L. et al. (2024). Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, 1123-1140.

INTRODUCTION

Hello. Today, we are evaluating how the ML-SEC article tested the validity of its hypotheses. Reminder of the three key hypotheses:

- ❑ **H1:** Adversarial attacks reduce the effectiveness of ML detectors by $>40\%$
- ❑ **H2:** Existing defenses are ineffective against adaptive attacks
- ❑ **H3:** Their ML-SEC framework quantifies robustness better than classical metrics. As highlighted in previous assignments, an explicit link between hypotheses, variables, and results is crucial. Let's analyze whether this article adheres to this principle.



H1 Test: Strong results but a biased sample

For H1, the authors conducted a rigorous experiment:

- Precise control of the independent variable: intensity of attacks (FGSM, PGD)
- Clear measurement of the dependent variable: false negative rate
- Quantified result: PGD reduces detection by 47% ($p < 0.001$) on EMBER' (Section 5.2)

Major problem

- ❑ The tests are limited to open-source models (MalConv, SOREL-20M)
- ❑ Exclusion of industrial solutions(CrowdStrike, SentinelOne)
- ❑ This invalidates the generalization of H1 to real-world environments, as noted by Carlini in his 2023 paper on evaluation biases.

H2 Assessment: Failing Methodology

For H2, the approach is problematic:

- ❑ Comparison of 5 'homemade' defenses without reference to industrial implementations.
- ❑ 'Adaptive attacks' are not operationally defined. Total absence of comparative statistical tests (ANOVA or post-hoc tests).
- ❑ Consequences: Impossible to verify if the defenses consistently fail. Violation of Popper's falsifiability principle: no condition is defined to invalidate H2. As NIST points out in its UNITE guide, the assessment of defenses requires a standardized framework.

H3 Test: Innovation or Illusion?

The evaluation of H3 reveals a paradox:

✓ Strength: Comprehensive benchmark against 6 classic metrics (AUC, F1-score)

✗ Critical weakness:

- The new metrics (RS, ACE, DOI) are not correlated with actual security KPIs
- No test evaluates their impact on operational indicators such as MTTR or average incident cost

Revealing quote:

Our metrics demonstrate statistical superiority (Section 6.3) but without explaining what 'superiority' concretely means for a SOC. This disconnect from the field has been denounced by Pierazzi in USENIX 2023.

Summary of failures

Three major failures compromise the overall validity:

- 1) Selection bias:** Tests conducted exclusively on academic models (partial invalidation of H1)
- 2) Insufficient statistical rigor:** No defined significance threshold for H2
- 3) Null operational validity:** Metrics uncorrelated with real-world realities (collapse of H3) These lapses illustrate the 'ivory tower syndrome' denounced by Carlini: hypotheses tested in an experimental void with no connection to actual cybersecurity.

Recommendations for credible tests

To correct these flaws:

- 1) Integrate commercial platforms via APIs (VirusTotal, SentinelOne)
- 2) **Adopt the NIST UNITE** framework to standardize defense testing
- 3) **Link metrics to concrete KPIs:Example:** 'How does the Robustness Score correlate with the Mean Time To Detect (MTTD)?'Without these corrections, hypothesis validation remains an academic exercise without relevance for practitioners.

CONCLUSION

Ultimately, while ML-SEC offers an innovative methodology, its validation tests are lacking in representativeness, statistical rigor, and real-world relevance. This directly addresses the professor's requirement: without a coherent validation chain linking hypotheses, variables, and results, research loses its operational credibility.

Key references:

Carlini, N. (2023). *"Why Adversarial ML Evaluations Fail"*.
IEEE S&P

Pierazzi, F. et al. (2023). *"Operationalizing ML Security Metrics"*. USENIX Security

NIST IR 8362 (2023). *UNITE Framework for ML Testing*

DOI : [10.1145/3576915.3623149](https://doi.org/10.1145/3576915.3623149)



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You for your attention

