Adversarial Learning for Robust ML Detection of Malicious Ethereum Smart Contracts and Transactions

Tamer Abdelaziz, Ph.D.

tamer.m@nyu.edu
tamer-abdelaziz.github.io

Abstract

Machine-learning detectors for malicious Ethereum smart contracts and transactions achieve high in-distribution accuracy but collapse under evasion attacks that apply semantic-preserving mutations to bytecode or subtle perturbations to transaction graphs, enabling zero-day exploits to bypass detection in production. This project introduces AdvRobDet, a unified adversarial-learning framework that trains robust classifiers via min-max optimization over bounded perturbations on both contract embeddings and transaction sequences. Building upon our team's prior work on ML-based dumb-contract detection, exploit schooling, and analyzer evaluation [2, 3, 4, 5, 6, 7], AdvRobDet advances recent adversarial-robustness efforts in blockchain ML [8, 9, 10, 11] by jointly hardening static contract and dynamic transaction models under realistic EVM semantics. We detail the threat model, training procedure, evaluation on adversarial benchmarks, and open-source deliverables, inviting collaborations in robust optimization and real-time deployment.

Problem Formalization

Existing ML-based detectors for malicious smart contracts and transactions are limited by their vulnerability to evasion: adversaries can craft semantically equivalent bytecode mutations (e.g., opcode reordering, dead-code insertion) or minimal transaction-feature perturbations that preserve exploit intent while flipping model predictions, leading to catastrophic false-negative rates in live environments. This project solves these limitations through end-to-end adversarial training that explicitly optimizes robustness against a formal threat model of bounded semantic-preserving perturbations, producing classifiers that maintain high detection accuracy even under strong adaptive attacks.

The rapid growth of DeFi and NFT ecosystems has made automated detection of malicious smart contracts (e.g., rug-pull, honeypot, or exploit templates) and malicious transaction sequences (flash-loan attacks, oracle manipulations) a critical line of defense. Traditional static and dynamic analyzers suffer from scalability issues, while purely data-driven ML models—although fast—remain brittle against adaptive adversaries who optimize inputs to evade detection while preserving exploit semantics.

We formalize the problem under a unified threat model. A contract is represented by its bytecode embedding $\mathbf{x}_c \in \mathbb{R}^d$ (obtained via our prior EVM-native encoders) and a transaction sequence by a temporal graph snapshot $\mathbf{x}_t = (G_1, \ldots, G_k)$. The detector f_θ outputs a maliciousness score $s = f_\theta(\mathbf{x}) \in [0, 1]$. An adversary $\mathcal{A}$ crafts a perturbed input $\tilde{\mathbf{x}} = \mathbf{x} + \delta$ subject to a semantic budget $\|\delta\|_\infty \leq \epsilon$ and an equivalence oracle ensuring functional behavior is unchanged (e.g., δ only permutes equivalent opcodes or reorders non-interfering calls). The robust risk is defined as

$$R_{\text{rob}}(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\max_{\delta\in\Delta(\mathbf{x})} \ell\big(f_\theta(\mathbf{x}+\delta), y\big)\Big],$$

where $\Delta(\mathbf{x})$ is the set of allowed perturbations and ℓ is a suitable loss (e.g., binary cross-entropy). We solve the inner maximization via projected gradient descent on the perturbation (PGD) and the outer minimization via standard SGD, yielding the adversarial-training objective

$$\min_{\theta} \max_{\delta} \mathcal{L}(\theta, \delta) = \mathcal{L}_{\text{cls}}(\theta) + \lambda \mathcal{L}_{\text{adv}}(\theta, \delta).$$

AdvRobDet is a dual-stream architecture: (1) a contract branch that augments our prior multimodal EVM-FM embeddings with a certified-robust layer based on randomized smoothing over opcode substitutions, and (2) a transaction branch that processes dynamic interaction graphs via temporal GNNs with adversarial message-passing regularization. Cross-stream attention fuses contract-level and transaction-level signals for end-to-end detection of both deployed malicious contracts and in-flight attack sequences. To handle label scarcity, we incorporate semi-supervised contrastive pretraining on unlabeled mainnet data followed by adversarial fine-tuning on labeled exploit traces.

Robustness evaluation follows a rigorous adversarial benchmark pipeline. We generate three attack strengths—white-box PGD, black-box transfer from surrogate models, and semantic-constrained genetic search—and measure certified accuracy under ℓ_∞ and semantic equivalence budgets. Additional metrics include attack success rate (ASR), clean accuracy, and detection latency on a production-like mempool replay. We further quantify economic impact by simulating prevented losses on historical DeFi incidents.

This project produces open-source robust model checkpoints, an adversarial benchmark suite extending our prior analyzer evaluation work, and a deployment toolkit for validators and security oracles. Expected insights include quantitative trade-offs between robustness, accuracy, and inference cost, as well as new theoretical bounds on perturbation budgets for EVM semantics.

We invite collaborators with expertise in certified robustness, game-theoretic adversary modeling, and large-scale Ethereum data access. Industry partnerships for on-chain integration and red-team validation are especially welcome; funded PhD and postdoc positions are available. Together we will raise the bar for production-grade ML defenses against adaptive blockchain adversaries.

References

[1] T. Abdelaziz, A. Sedky Adly, B. Rossi, and M.-S. Mostafa. Identification and assessment of software design pattern violations. *Informatics Bulletin*, 1(2):6–13, 2019. https://fcihib.journals.ekb.eg/article_107517_62d89752f7d871844b0e5dd1601da4f5.pdf.

[2] T. Abdelaziz and A. Hobor. Smart learning to find dumb contracts. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1775–1792, 2023. https://www.usenix.org/conference/usenixsecurity23/presentation/abdelaziz.

[3] T. Abdelaziz and A. Hobor. Smart learning to find dumb contracts (extended version). arXiv:2304.10726, 2023. https://arxiv.org/abs/2304.10726.

[4] T. Abdelaziz and A. Hobor. USENIX'23 artifact appendix: Smart learning to find dumb contracts. USENIX Association, 2023. https://www.usenix.org/system/files/usenixsecurity23-appendix-abdelaziz.pdf.

[5] T. Abdelaziz and A. Hobor. Schooling to exploit foolish contracts. In *2023 Fifth International Conference on Blockchain Computing and Applications (BCCA)*, pages 388–395, 2023. https://ieeexplore.ieee.org/document/10338924.

[6] T. A. Abdelmegid Mohamed. Towards secure smart contracts: A deep learning approach for detecting security threats. PhD thesis, National University of Singapore, 2023. https://scholarbank.nus.edu.sg/handle/10635/247301.

[7] T. Abdelaziz, S. Alsaghir, and K. Ali. Where do smart contract security analyzers fall short? *Proceedings of the 2026 IEEE/ACM 23rd International Conference on Mining Software Repositories (MSR)*, 2026.

[8] Y. Liu et al. Generic Adversarial Smart Contract Detection with Semantics and Uncertainty-Aware LLM. arXiv:2509.18934, 2025. https://arxiv.org/abs/2509.18934.

[9] M. Kamran et al. ARCADE—Adversarially Robust Cost-Sensitive Anomaly Detection in Blockchain Using Explainable Artificial Intelligence. *Electronics*, 14(8):1648, 2025. https://www.mdpi.com/2079-9292/14/8/1648.

[10] A. Alghuried et al. Evaluating the Vulnerability of ML-Based Ethereum Phishing Detectors to Single-Feature Adversarial Perturbations. arXiv:2504.17684, 2025. https://arxiv.org/abs/2504.17684.

[11] S. Ren et al. LookAhead: Preventing DeFi Attacks via Unveiling Adversarial Contracts. arXiv:2401.07261, 2025. https://arxiv.org/abs/2401.07261.

[12] G. Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 2014. https://ethereum.github.io/yellowpaper/paper.pdf.

[13] L. Luu, D.-H. Chu, H. Olickel, P. Saxena, and A. Hobor. Making smart contracts smarter. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

[14] N. Atzei, M. Bartoletti, and T. Cimoli. A survey of attacks on ethereum smart contracts (soK). In *Proceedings of the 6th International Conference on Principles of Security and Trust (POST)*, 2017.