1. OWASP Mobile Application Security (MAS). URL: https://mas.owasp.org/ (дата звернення: 14.05.2026).
2. Bloch J. Effective Java. 3rd ed. Boston: Addison-Wesley Professional, 2017. 412 p.
3. Griffiths D., Griffiths D. Head First Android Development: A Learner's Guide to Building Android Apps with Kotlin. 3rd ed. Sebastopol: O'Reilly Media, 2021. 930 p.

Змагальні атаки на системи виявлення вторгнень з гібридною архітектурою у мережах IoT

УДК 004.056.5 Ірина Удовик[1], Олександр Кручінін[2], Дмитро Тимофєєв[3]

Національний технічний університет «Дніпровська політехніка»,
[1]udovyk.i.m@nmu.one, [2]kruchinin.o.v@nmu.one, [3]tymofieiev.d.s@nmu.one

Сучасна еволюція цифрової інфраструктури характеризується стрімким поширенням технологій Інтернету речей (IoT) та кіберфізичних систем. Враховуючи динамічність, багатоетапність та адаптивність сучасних кіберзагроз. інтеграція методів машинного навчання (ML) та глибокого навчання (DL) у системи виявлення вторгнень (IDS) стала критичною необхідністю. Однак, в цьому випадку з'являються загрози реалізації змагальних атак (adversarial attacks) на такі IDS.

Метою даної роботи є аналіз можливих змагальних атак на IDS з гібридною архітектурою у мережах IoT.

Однією з найбільш перспективних стратегій захисту є перехід від ізольованого аналізу окремих подій до виявлення кореляційних зв'язків у часі та просторі. Традиційні IDS не враховують, що атаки в середовищах IoT поширюються через логічні взаємини між пристроями та еволюціонують через чіткі темпоральні фази. Застосування гібридних архітектур, що поєднують графові нейронні мережі (GNN) та мережі довгої короткострокової пам'яті (LSTM), дозволяє одночасно фіксувати структурні та часові динаміки атак [1].

Однак ця подвійна природа збільшує поверхню для змагальних атак. Однією із вразливістю GNN є неевклідова природа графових даних, де навіть незначна зміна ваги ребра або атрибута вузла може радикально змінити результат агрегації повідомлень через ітеративний характер навчання. У випадку LSTM вразливість криється в авторегресивній природі моделі, тобто помилка, внесена в один часовий крок, накопичується та спотворює внутрішній стан комірки пам'яті для всіх наступних кроків.

Серед таких змагальних атак можна виділити наступні:

1) Fast Gradient Sign Method (FGSM) – є однією з найбільш фундаментальних атак білої скриньки, яка використовує градієнт функції втрат щодо вхідних даних для швидкої генерації змагальних прикладів. В IoT-мережах FGSM дозволяє зловмиснику маніпулювати статистичними характеристиками пакетів (час між пакетами або розміром вікна), роблячи шкідливий потік невідрізним від нормального для GNN-класифікатора. Це особливо ефективно проти моделей, які не пройшли спеціальне змагальне навчання [2].

2) Projected Gradient Descent (PGD) – являє собою ітераційне вдосконалення FGSM, що робить її значно потужнішою атакою першого порядку. Для гібридних архітектур PGD є критичною загрозою, оскільки вона може бути налаштована на пошук "найгіршого випадку" збурення, яке обходить як структурні фільтри GNN, так і часові перевірки LSTM. Вона демонструє високий рівень успіху навіть проти захищених систем [3].

3) Temporal Adversarial Examples Attack Model (TEAM) – є спеціалізованою атакою, розробленою для експлуатації рекурентної природи RNN та LSTM у мережевих IDS. Це одна з небагатьох атак, яка безпосередньо атакує пам'ять LSTM, використовуючи інерційність моделі проти неї самої, тим самим підвищуючи рівень помилок до 96.68%. Це робить її надзвичайно небезпечною для реальних сценаріїв IoT, де трафік генерується безперервно [4].

4) Distance to Target Center (D2TC) – це атака «чорної скриньки», орієнтована на конкретні класи трафіку в IoT-мережах. Оскільки GNN-LSTM моделі часто покладаються на агреговані метрики для розрізнення типів атак D2TC дозволяє зловмиснику "розчинити" атаку в фоновому трафіку без доступу до градієнтів моделі[5].

5) Hierarchical Adversarial Attack (HAA) – також використовує стратегію «чорної скриньки», яка враховує ієрархічну структуру IoT-мереж. Використовуючи алгоритм випадкових блукань з перезапуском, атака ідентифікує ключові вузли в топології мережі. Потім вона модифікує критичні ознаки цих вузлів, щоб максимізувати вплив на представлення всього графа[6].

Дослідження цих та інших змагальних атак на IDS з гібридною архітектурою у мережах IoT є важливими для вдосконалення засобів протидії таким атакам, враховуючи реальні умови реалізації. Це є необхідною умовою для ефективного впровадження таких IDS.

1. Babenko, T.; Kolesnikova, K.; Bakhtiyarova, Y.; Yeskendirova, D.; Sansyzbay, K.; Sysoyev, A.; Kruchinin, O. (2026). Hybrid GNN-LSTM Architecture for Probabilistic IoT Botnet Detection with Calibrated Risk Assessment: Computers, 15(1), p.26. – URL: https://www.mdpi.com/2073-431X/15/1/26. (дата звернення: 25.04.2026).
2. Karma Gurung, Ashutosh Ghimire, Fathi Amsaad. (2025). Enhancing IoT Intrusion Detection Systems through Adversarial Training. – URL: https://arxiv.org/abs/2507.19739v1. (дата звернення: 25.04.2026).
3. Ade Kurniawan, Merios Gusan Putra, Dani Lukman Hakim, Mochammad Ariyanto. (2026). Temporal Adversarial Attacks on Time Series and Reinforcement Learning Systems. – URL: https://www.preprints.org/manuscript/202601.0598. (дата звернення: 25.04.2026).
4. Ziyi Liu, Dengpan Ye, Long Tang, Yunming Zhang, Jiacheng Deng. (2024). TEAM: Temporal Adversarial Examples Attack Model against Network Intrusion Detection System Applied to RNN. – URL: https://arxiv.org/abs/2409.12472. (дата звернення: 26.04.2026).
5. Islam Debicha, Tayeb Kenaza, Ishak Charfi, Salah Mosbah, Mehdi Sehaki, Jean-Michel Dricot. (2026). Targeted adversarial traffic generation: black-box approach to evade intrusion detection systems in IoT networks. – URL:

https://arxiv.org/html/2603.23438v1. (дата звернення: 26.04.2026)
6. Dimitri Galli, Andrea Venturi, Dario Stabili, Mauro Andreolini, Mirco Marchetti. (2025). Defending Network Intrusion Detection Systems Based on Graph Neural Networks Against Structural Adversarial Attacks. – URL: https://ieeexplore.ieee.org/document/11261632. (дата звернення: 27.04.2026)

On Evidence Deficits in Kleptography and the Application of Artificial Intelligence for Their Mitigation

UDK 004.4:056.57

Mykhailo Shelest[1], Yuliia Tkach[2], Oleksandr Polevod[3]

National University "Chernihiv Polytechnic",
[1]mishel3141@gmail.com, [2]tkachym79@gmail.com,
[3]oleksandr.polevod23@gmail.com

Modern information systems are increasingly viewed not only as targets of attacks, but as potentially controllable environments in which hidden intervention may occur without explicit violation of functionality. Within the kleptographic paradigm [1], the central problem is not merely the fact of compromise, but the *deficit of evidence*, i.e., the inability to reliably detect and prove the presence of hidden influence.

This work proposes a formalization of a class of evidence deficits in kleptography and introduces their classification as a distinct object of cybersecurity analysis. The proposed approach enables a transition from descriptive consideration of kleptographic threats to a systematic analysis of the limitations of the evidentiary base and the conditions under which such limitations manifest.

Unlike classical information security incidents, kleptographic interventions may leave no unambiguous technical traces, which creates fundamental constraints for their detection and attribution. In this context, we propose to treat these constraints as formalized evidence deficits that define the boundaries of applicability of traditional analysis methods.

Several key classes of such deficits can be identified.

First, the *reproducibility deficit* manifests itself in the instability of anomalies that arise only under specific conditions, such as particular system states or environmental configurations [2]. This prevents their reliable reproduction in laboratory settings. For example, in systems with dependencies on third-party libraries, behavioral deviations may occur only under narrowly defined input conditions that are not captured by standard testing procedures. A typical case involves a modified software component that exhibits correct behavior during testing but demonstrates selective or controlled behavior in real-world deployment, making it practically non-reproducible.

Second, the *causality deficit* refers to situations where an observable effect exists, but the underlying mechanism cannot be isolated or distinguished from normal system behavior.

Third, the *provenance deficit* is associated with the inability to reliably verify the origin of software artifacts, which is particularly critical in the context of supply chain attacks.