УДК 004.8
О.О. Ясінський; Г.І. Липак, к. соц. ком.
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ПРОБЛЕМА РЕПРЕЗЕНТАТИВНОСТІ НАБОРІВ ДАНИХ У ТЕОРЕТИЧНОМУ МОДЕЛЮВАННІ СУЧАСНИХ КІБЕРЗАГРОЗ

O.O. Yasinskyi; H.I. Lypak Ph.D

THE PROBLEM OF DATA SET REPRESENTATIVENESS IN THEORETICAL MODELING OF MODERN CYBERTHREATS

У 2025 році темпи еволюції кіберзагроз значно перевищують швидкість оновлення більшості дослідницьких наборів даних, що використовуються для тренування моделей систем виявлення вторгнень. За оцінками фахівців, понад 40% нових типів атак, зафіксованих протягом останніх двох років, не мають ані загальнодоступних еталонних наборів даних, ані формалізованих поведінкових профілів. Це створює критичний розрив між теоретичними результатами, отриманими в умовах лабораторного тестування, та реальними можливостями ML-моделей протистояти модернізованим, адаптивним і часто зашифрованим атакам у корпоративних мережах. Проблема репрезентативності даних стає ключовим обмеженням для розвитку IDS, оскільки сучасні атаки є динамічними, контекстними й дедалі краще оптимізованими для обходу алгоритмічних механізмів захисту.

Ефективність систем виявлення вторгень (англ. Intrusion Detection System, IDS) на базі машинного навчання (англ. *Machine learning*, ML) критично залежить від якості даних, на яких вони навчаються. Фундаментальна проблема полягає у значному розриві між високою лабораторною точністю ML-моделей та їхньою практичною неспроможністю виявляти новітні кіберзагрози. Це відбувається тому, що більшість популярних наборів даних є застарілими або неповно відображають характеристики атак.

NSL-KDD, що є покращеною версією набору даних KDD99, досі залишається найбільш поширеним для академічних порівнянь. Однак його трафік частково синтетичний, а загрози абсолютно неактуальні у 2025 році, якщо порівнювати з діями APT (англ. Advanced Persistent Threat, розвинена стала загроза) чи атаками де є шифрування. Моделі, навчені на ньому, фактично вчаться розпізнавати "артефакти" самого набору даних.

Набір даних CICIDS2017 був значним кроком уперед, оскільки він містив дані 2017 року з реалістичними профілями атак (DDoS, XSS, Botnet). Проте за понад 8 років він також застарів. Крім того, його атаки були "профільованими" – виконаними за чітким сценарієм – що не відображає реальні "тихі" та повільні (low-and-slow) атаки типу APT, які маскуються під легітимний трафік.

Проблема в тому, що дослідники продовжують публікувати роботи, де їхні нові нейронні мережі показують високий відсоток точності на NSL-KDD. Це створює хибне відчуття вирішеної проблеми. Насправді така модель просто ідеально "вивчила" застарілий, нерелевантний набір даних.

При перенесенні у реальну корпоративну мережу, модель стикається з трафіком, профілі якого ніколи не були представлені під час тренування.

В таблиці 1 подано аналіз загроз, які відсутні в сучасних популярних наборах даних.

Таблиця 1. Загрози, що відсутні в наборах даних

Тип загрози	Опис	Причина застарілості
Adversarial AI Attacks	Атаки, що цілеспрямовано "обманюють" ML-моделі, вносячи ледь помітні зміни в трафік.	Набори даних не містять таких прикладів.
Fileless Malware	Атаки "без файлів", що живуть у оперативній пам'яті системи.	Більшість наборів даних аналізують мережевий трафік, а не поведінку хоста.
Атаки на IoT	Специфічні атаки на "розумні" пристрої та промислові системи.	Використовують специфічні протоколи (MQTT, Modbus), яких немає в CICIDS чи NSL-KDD.
Зашифрований C&C трафік	Командні центри, що спілкуються через зашифровані канали	Аналіз метаданих зашифрованого трафіку – задача, не відображена в старих наборах даних.

Висновок. Використання застарілих та нерепрезентативних наборів даних або наборів даних, що швидко втрачають актуальність, є серйозною перешкодою для розвитку ефективних IDS. Теоретичне моделювання показує завищені результати, що не транслюються у реальну безпеку. Вирішення проблеми не релевантних наборів даних вимагає нових підходів, а саме використання генеративних моделей для імітації нових атак, навчання на реальних даних та фокус на метаданих замість зашифрованого вмісту.

Література

1. Sharafaldin I., Lashkari A. H., Ghorbani A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSP. 2018. Vol. 1. P. 108–116.URL: https://www.scitepress.org/papers/2018/66398/66398.pdf

2. Tavallaee M., Bagheri E., Lu W., Ghorbani A. A. A detailed analysis of the KDD CUP 99 data set. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2009. URL: https://www.researchgate.net/publication/48446353_A_detailed_analysis_of_the_KDD_CUP_99_data_set

3. Ring M., Wunderlich S., Grüdl D., Landes D., Hotho A. A survey of network-based intrusion detection data sets. Computers & Security. 2019. URL: https://arxiv.org/abs/1903.02460 .

4. Engelen T., van Ede T., Doerr C. Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study. Proceedings of the 2021 Workshop on Theoretic and Experimental Benchmarking of BGP Attacks and Defenses (WTMC 2021). 2021.URL: https://www.researchgate.net/publication/353107141_Troubleshooting_an_Intrusion_Detection_Dataset_the_CICIDS2017_Case_Study