DOI: 10.34185/1991-7848.itmm.2026.01.069

РИЗИКИ ВИКОРИСТАННЯ СИСТЕМ ВИЯВЛЕННЯ МЕРЕЖЕВИХ ВТОРГНЕНЬ ЯК ДЖЕРЕЛА НАВЧАЛЬНИХ МІТОК ДЛЯ НЕЙРОМЕРЕЖ

Горбатов В.С.[1] [ORCID], Журба А.О.[2] [ORCID]

[1]УДУНТ, аспірант. Україна

[2]УДУНТ, к.т.н., доцент. Україна

Анотація. *Використання сповіщень систем виявлення мережевих вторгнень (NIDS) як навчальних міток для моделей машинного навчання спричиняє виникнення систематичних похибок, що суттєво погіршують точність виявлення загроз. У дослідженні проаналізовано розбіжності між фактичними мережевими атаками та спрацюваннями сигнатурних аналізаторів, з акцентом на трьох критичних викликах: однобічній похибці маркування, самопідсиленні помилок у середовищах безперервного навчання та вразливості до навмисного отруєння даних. Зокрема, неспроможність традиційних систем ідентифікувати загрози нульового дня призводить до забруднення негативного класу, в якому пропущені атаки помилково класифікуються як безпечний трафік. Для мінімізації цих ризиків розглянуто стратегії нейтралізації, зокрема навчання на позитивних і немаркованих даних (PU-learning), слабке керування навчанням та механізми фільтрації за рівнем довіри. Впровадження надійних протоколів перевірки та методів буферизації забезпечує достовірніше виявлення вторгнень і підвищує стійкість нейромереж до мінливих кіберзагроз у динамічних середовищах.*

Ключові слова: *NIDS, машинне навчання, кібербезпека, PU-learning, дрейф концепції, самопідсилення похибки, онлайн-адаптація моделі.*

Вступ. Сигнатурні системи виявлення мережевих вторгнень вирізняються високою точністю виявлення відомих загроз за низької частоти хибних спрацювань, проте вони не здатні розпізнавати нові вектори атак [1]. Використання їхніх сповіщень як навчальних міток формує однобічну похибку: маркування атакувального трафіку є здебільшого коректним, натомість відсутність спрацювання не гарантує безпечності. Невиявлені атаки класифікуються як звичайний трафік, що викривлює навчальну вибірку. Модель оптимізується на зміщених даних і втрачає здатність виявляти пропущені системою NIDS загрози. Ця ситуація потребує застосування методів навчання на позитивних і немаркованих даних, оскільки стандартні підходи

керованого навчання є неефективними через хибне трактування всіх немаркованих зразків як абсолютно безпечних [2].

Однобічна похибка маркування. Сигнатурні NIDS пропускають значну кількість нових атак. У навчальній вибірці утворюється дисбаланс: позитивний клас (відомі атаки) представлений вузько, а негативний клас (безпечний трафік) містить приховані позитивні випадки. Навчена на таких даних модель отримує систематичне зміщення, розпізнаючи лише відомі шаблони та ігноруючи нові аномалії через їхню попередню класифікацію як безпечних [2].

Шляхи вирішення:

- Навчання на позитивних і немаркованих даних: Алгоритми цього класу оцінюють частку прихованих атак і коригують вагові коефіцієнти для компенсації зміщення. Часто використовують двокрокові схеми з попереднім пошуком аномалій серед немаркованих даних [2].
- Слабке керування навчанням: Залучення додаткових джерел інформації (евристичних правил, списків блокувань) для формування попередніх міток. Часткова анотація з високим рівнем довіри зменшує кількість хибних негативних прикладів [3].

Самопідсилення помилок під час безперервної адаптації. Під час безперервного донавчання моделі на нових даних виникає ефект накопичення похибок [4]. Якщо алгоритм не розпізнає атаку і зараховує цей трафік до безпечного, під час наступного циклу навчання хибна асоціація посилюється. Відбувається поступова деградація точності та зсув робочої концепції. Без механізму зовнішнього контролю алгоритм втрачає чутливість до аномалій, помилково адаптуючись до нових атак як до норми.

Шляхи вирішення:

- Фільтрація за рівнем довіри: Модель має оновлюватися лише на тих прикладах, щодо яких існує найвища впевненість у правильності маркування [4].
- Перехресна перевірка: Використання двох різних моделей. У разі розбіжності оцінок зразок вилучається з процесу оновлення [5].
- Попередня перевірка на еталонному наборі: Якщо після донавчання метрики погіршуються, нові дані відхиляються [5].
- Буферизація історичних даних: Збереження підтверджених атак минулих періодів у буфері запобігає забуванню попередніх класів [3].

- Активне навчання: Система самостійно відбирає найбільш суперечливі зразки для ручної експертної перевірки [6].

Вразливість до цілеспрямованого викривлення даних. Автоматичне оновлення моделі створює загрозу навмисного отруєння навчальної вибірки зловмисником. Нападник може генерувати шкідливий трафік, який не викликає спрацювання правил, змушуючи модель вивчити його як безпечний. Альтернативний сценарій передбачає створення штучних спрацювань на безпечному трафіку, що призводить до блокування легітимних користувачів. Цілеспрямоване викривлення даних здатне критично знизити точність класифікатора або створити приховані вразливості для безперешкодного доступу [5].

Шляхи вирішення:

- Контроль цілісності даних: Зразки, додавання яких до тренувальної вибірки погіршує показники моделі на контрольній множині, ідентифікуються як шкідливі та відхиляються [5].
- Адверсаріальне навчання: Включення до навчального процесу штучно модифікованих зразків та аномалій підвищує стійкість нейромережі до цілеспрямованих маніпуляцій із вхідними характеристиками трафіку [5].

Висновки

Формування навчальних вибірок на основі спрацювань сигнатурних аналізаторів супроводжується ризиками неповноти та ненадійності даних. Однобічна похибка маркування знижує рівень виявлення нових атак, самопідсилення помилок деградує модель у часі, а відсутність механізмів перевірки уможливлює цілеспрямоване викривлення вибірки. Застосування навчання на позитивних і немаркованих даних, фільтрації за рівнем довіри, буферизації, активного навчання та змагальних тренувань мінімізує зазначені ризики. Комплексне впровадження цих стратегій забезпечує стійкість нейромережевих систем виявлення вторгнень в умовах реальних динамічних середовищ.

ЛІТЕРАТУРА / REFERENCE

1. Feng Y., Sakurai K. Network Intrusion Detection: Evolution from Conventional Approaches to LLM Collaboration and Emerging Risks. URL: https://arxiv.org/abs/2510.23313.
2. Dilworth R., Gudla C. Applications of Positive Unlabeled (PU) and Negative Unlabeled (NU) Learning in Cybersecurity. URL: https://arxiv.org/abs/2412.06203.
3. Caravan: practical online learning of in-network ML models with labeling agents / Q. Zhang et al. 18th USENIX symposium on operating systems design and implementation (OSDI 24). Santa Clara, CA, 2024. P. 325–345. URL: https://www.usenix.org/conference/osdi24/presentation/zhang-qizheng.
4. Zou H. P., Caragea C. JointMatch: a unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. Proceedings of the 2023 conference on empirical methods in natural language processing / ed. by H. Bouamor, J. Pino, K. Bali. Singapore, 2023. P. 7290–7301. URL: https://doi.org/10.18653/v1/2023.emnlp-main.451.
5. Alajaji A. FortiNIDS: defending smart city iot infrastructures against transferable adversarial poisoning in machine learning-based intrusion detection systems. Sensors. 2025. Vol. 25, no. 19. P. 6056. URL: https://doi.org/10.3390/s25196056 (date of access: 04.03.2026).
6. Managing Concept Drift in Online Intrusion Detection Systems with Active Learning / C. F. et al. URL: https://www.tib.eu/de/suchen/id/base:1b52787437b97f11f6c3a39a28994f83fc750b5f.
7. Sommer R., Paxson V. Outside the closed world: on using machine learning for network intrusion detection. 2010 IEEE symposium on security and privacy. 2010. P. 305–316. URL: https://doi.org/10.1109/SP.2010.25.

RISKS OF USING NETWORK INTRUSION DETECTION SYSTEMS AS A SOURCE OF TRAINING LABELS FOR NEURAL NETWORKS

Vitalii Gorbatov, Anna Zhurba

Abstract. *The use of network intrusion detection system (NIDS) alerts as training labels for machine learning models introduces systematic biases that significantly degrade detection accuracy. This study investigates the discrepancies between actual network attacks and signature-based triggers, focusing on three critical challenges: one-sided labeling bias, error self-amplification in continuous learning environments, and vulnerability to adversarial data poisoning. Specifically, the inability of traditional NIDS to identify zero-day threats results in a polluted negative class, where missed attacks are misclassified as legitimate traffic. To address these risks, mitigation strategies are analyzed, including positive-unlabeled (PU) learning, weak supervision, and confidence-based filtering mechanisms. Implementing these robust validation protocols and buffering techniques ensures more reliable threat detection and enhances the resilience of neural networks against evolving cyber threats in dynamic network environments.*

Keywords: *NIDS, machine learning, cybersecurity, PU-learning, concept drift, error self-amplification, online model adaptation.*