Вплив реалістичних умов реалізації змагальних атак проти систем виявлення вторгнень на методи захисту

УДК 004.056.5

Олександр Кручінін[1], Дмитро Тимофєєв[2], Сергій Мацюк[3]

Національний технічний університет «Дніпровська політехніка»,

[1]kruchinin.o.v@nmu.one, [2]tymofieiev.d.s@nmu.one, [3]matsiuk.s.m@nmu.one

В умовах зростання складності та обсягів кіберзагроз, системи виявлення вторгнень (англ. Intrusion Detection System, IDS) відіграють критично важливу роль у забезпеченні безпеки мереж. Перспективним напрямком розвитку IDS є інтеграція методів машинного навчання (англ. Machine Learning, ML), що дозволяє виявляти як відомі, так і нові типи атак. Однак, IDS на базі ML є вразливими до змагальних атак. Тому актуальною є задача розвитку методів та засобів протидії таким атакам [1]. Для оптимізації цих методів та засобів протидії є доцільним врахування реалістичних умови реалізації змагальних атак [2].

Метою даної роботи є аналіз реалістичних умов реалізації змагальних атак проти IDS, які використовують ML, із врахуванням моделей загроз та порушника.

Окремі методи протидії змагальним атакам проти IDS, як правило, адаптовані до конкретного виду атак. Одним з методом реалізації більш універсального рішення є використання ансамблів моделей, тобто інтеграція кількох моделей машинного навчання та поєднання різних стратегій ідентифікації. Але таке рішення має свої недоліки: складність реалізації порівняно з одиночними моделями; підвищені вимоги до обчислювальних ресурсів та енергоспоживання; збільшення затримки у виявленні та реакції на атаки; збільшення трудомісткості налаштування та оновлення.

Слід зазначити, що для конкретних інформаційно-комунікаційних систем (ІКС) існує обмежена кількість реальних сценаріїв реалізації змагальних атак. Тобто множина таких атак та, відповідно, актуальних методів та засобів протидії є обмеженою.

На сьогодні для класифікації змагальних атак використовується ряд ознак [3]:

1) Рівень знань зловмисника про IDS («біла», «сіра» або «чорна» скриня).

2) Мета атаки (конфіденційність, цілісність, доступність).

3) Стратегія атаки (ухилення, отруєння, оракула).

4) Фаза атаки (із впливом на тренувальні дані, без впливу на тренувальні дані).

5) Простір атаки (ознак, задач).

6) Спрямованість (спрямована, не спрямована)

Частина цих ознак залежить від конкретних умов функціонування ІКС, тобто від моделі загроз та порушника.

Для демонстрації такого впливу можна розглянути декілька сценаріїв реалізації змагальних атак проти IDS для різних ІКС з різними профілями зловмисників:

1) об’єкт атаки: веб-сервер. Зловмисник, не маючи інформації про внутрішню роботу IDS на базі ML, що захищає веб-сервер, намагається обійти його захист. Для цього він використовує автоматизовані інструменти для сканування веб-сервера з метою виявлення вразливостей, таких як SQL injection або Cross-Site Scripting (XSS). Зловмисник поступово змінює параметри своїх запитів, наприклад, кодує шкідливі SQL-команди або змінює структуру XSS-скриптів, щоб обійти правила IDS. Оскільки зловмисник не знає, які саме ознаки використовує IDS для виявлення атак, він намагається внести невеликі зміни у велику кількість параметрів, щоб знайти комбінацію, яка дозволить обійти захист. Його кінцевою метою є успішна експлуатація вразливості веб-сервера, не будучи виявленим IDS. Ця атака за рівнем знань зловмисника про IDS – «чорна» скриня, без впливу на тренувальні дані, націлена на зміну або пошкодження даних або моделі машинного навчання, не спрямована;

2) об’єкт атаки: промислове обладнання. Інсайдер, маючи повний доступ до інформації про IDS на базі ML, яка контролює роботу промислового контролера (PLC) технологічного обладнання, планує диверсію. Він знає архітектуру IDS, має доступ до тренувальних даних та алгоритму навчання. Використовуючи ці знання, він розробляє змагальні приклади, які маніпулюють вхідними даними контролера, наприклад, значеннями тиску, температури та швидкості обертання.

Мета маніпуляцій - змусити контролер працювати в небезпечному режимі, не викликаючи при цьому підозри у IDS. Кінцевою метою є спричинення аварії промислового обладнання, залишаючись непоміченим системою захисту. Ця атака за рівнем знань зловмисника про IDS – «біла» скриня, без впливу на тренувальні дані, націлена на зміну роботи контролера та обхід захисту для саботажу.

Таким чином, на основі аналізу моделі загроз та порушника можна визначити реалістичні умови та результати реалізації змагальних атак проти IDS, а значить сфокусувати увагу на обмеженому переліку таких атак. Це дозволить оптимізувати ефективність застосування методів та засобів протидії таким атакам за умови мінімізації вимог до обчислювальних потужностей та затримки у виявленні та реакції на атаки.

1. Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. applied sciences, 9(20), 4396. – URL: https://www.mdpi.com/2076-3417/9/20/4396. (дата звернення: 25.04.2025).

2. Ennaji, Sabrine & Benkhelifa, Elhadj & Mancini, Luigi. (2025). Toward Realistic Adversarial Attacks in IDS: A Novel Feasibility Metric for Transferability. – URL: https://arxiv.org/abs/2504.08480. (дата звернення: 25.04.2025).

3. Alotaibi, A.; Rassam, M.A. Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet 2023, 15, 62. – URL: https://doi.org/10.3390/fi15020062. (дата звернення: 25.04.2025).