Дячук О.Ю., ст. викладач
Колощук М.С., асистент
Окунькова О.О., ст. викладач
Державний університет «Житомирська політехніка»

ШТУЧНИЙ ІНТЕЛЕКТ ПІД ПРИЦІЛОМ:
ЕВОЛЮЦІЯ АТАК НА ML-СИСТЕМИ ТА ТЕХНОЛОГІЇ ЗАХИСТУ

Системи машинного навчання (ML) та штучного інтелекту (ШІ) стають невід'ємною частиною багатьох критичних сфер, таких як фінанси, медицина, транспорт і кібербезпека. Проте їх широке впровадження супроводжується зростанням ризиків, пов'язаних із кібератаками. У цьому дослідженні розглядаються нові вектори атак на ML/ШІ, оцінюється ефективність існуючих методів захисту та пропонуються покращені стратегії безпеки.

Одна з найнебезпечніших атак – це "отруєння даних" (Data Poisoning), коли зловмисники маніпулюють навчальними даними, щоб змусити модель навчитися неправильним патернам. Наприклад, якщо атака спрямована на систему розпізнавання загроз, хакер може додати у навчальний набір шкідливий код, який алгоритм помилково визначить як безпечний. Подібним чином працюють "атаки через білошум" (Adversarial Attacks) – вони використовують спеціально змінені вхідні дані, які здаються нормальними для людини, але вводять модель в оману. Достатньо незначних змін у зображенні, щоб система автономного транспорту сприйняла автомобіль за пішохода, що може призвести до критичних наслідків. Ще одним серйозним ризиком є "інверсія моделі" (Model Inversion Attack), коли хакери можуть отримати доступ до вихідних навчальних даних, аналізуючи відповіді моделі. Це особливо небезпечно у фінансових та медичних застосунках, де конфіденційність даних має першорядне значення.

Реальні випадки атак підтверджують загрозу таких векторів. Наприклад, дослідники Google виявили, що модифіковані аудіофайли можуть змусити голосових асистентів виконувати команди без відома користувача. У 2023 році хакери використали технологію діпфейків для обходу біометричної автентифікації в банківських системах, створивши підроблені відеодзвінки та отримавши доступ до рахунків клієнтів.

Для мінімізації ризиків атак необхідно застосовувати комплексні заходи захисту. Одним із найефективніших є адаптивне навчання (Adversarial Training), яке передбачає тренування моделей на спеціально змінених даних, що містять потенційні атаки. Це підвищує стійкість алгоритмів, хоча вимагає значних обчислювальних ресурсів. Важливим є також захист навчальних даних, наприклад, за допомогою криптографічних методів, таких як гомоморфне шифрування чи диференційна конфіденційність, які дозволяють мінімізувати ризик витоку інформації. Моніторинг поведінки моделей, впровадження систем виявлення аномалій та обмеження доступу до моделі є додатковими ефективними методами протидії атакам.

Аналіз існуючих підходів до захисту показує, що адаптивне навчання та моніторинг аномалій є найбільш перспективними методами, але вони мають певні недоліки. Зокрема, адаптивне навчання потребує значних обчислювальних потужностей, а системи виявлення аномалій не гарантують стовідсоткового захисту від нових векторів атак. Тому доцільно застосовувати гібридний підхід, який поєднує адаптивне навчання та традиційні методи кіберзахисту. Перспективними напрямами є автоматизоване виявлення атак за допомогою самонавчальних алгоритмів, розподілене навчання (Federated Learning) для зменшення ризику компрометації даних, а також створення фільтрів для очищення вхідних даних, що дозволяє знизити ефективність adversarial атак.

Загалом атаки на системи ML/ШІ стають усе більш поширеними, і їхні наслідки можуть бути критичними для безпеки даних та функціонування автоматизованих систем. Впровадження передових методів захисту є необхідною умовою для забезпечення стійкості сучасних ШІ-алгоритмів. Найефективнішими на сьогодні є адаптивне навчання та моніторинг аномалій, проте для подальшого підвищення рівня безпеки необхідно розвивати комплексні підходи, що поєднують класичні методи кібербезпеки з інноваційними технологіями ML. Подальші дослідження в цій сфері відіграють ключову роль у розробці нових захисних механізмів, здатних ефективно протидіяти сучасним загрозам.

Список використаних джерел:

1. Biggio B., Roli F. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition. 2022. DOI: https://doi.org/10.1016/j.patcog.2021.107384.
2. Papernot N., McDaniel P. Towards the Science of Security and Privacy in Machine Learning. IEEE Transactions on Information Forensics and Security. 2018. DOI: https://doi.org/10.1109/TIFS.2018.2871657.
3. Tramèr F., Zhang F. Stealing Machine Learning Models via Prediction APIs. USENIX Security Symposium. 2016. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer (дата звернення: 11.03.2025).