УДК 004.8:004.056

Нехороших Д.М.

асистент кафедри Безпеки інформаційних технологій

Харківський національний університет радіоелектроніки

Іванський І.О.

здобувач вищої освіти бакалаврського рівня

Харківський національний університет радіоелектроніки

АНАЛІЗ ПРОЦЕСІВ ЦІЛЕСПРЯМОВАНОЇ ДЕГРАДАЦІЇ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ПРОТЯГОМ ЖИТТЄВОГО ЦИКЛУ

Анотація: У роботі розглянуто механізми цілеспрямованого погіршення роботи моделей машинного навчання на етапах підготовки даних, навчання та використання в системах кібербезпеки. Проаналізовано отруєння навчальних даних, навмисне спотворення вхідних даних і атаки шляхом впровадження зловмисних інструкцій, а також їхній вплив на точність класифікації та виявлення кіберзагроз. Визначено основні технічні й організаційні заходи підвищення стійкості моделей.

Ключові слова: машинне навчання, кібербезпека, погіршення роботи моделей, отруєння навчальних даних, навмисне спотворення вхідних даних, впровадження зловмисних інструкцій, виявлення кіберзагроз, стійкість моделей.

Вступ. Активне використання машинного навчання в кібербезпеці пов’язане з потребою автоматизовано аналізувати великі обсяги даних,

виявляти аномальну активність та реагувати на кіберзагрози. Ці моделі застосовуються для аналізу мережевого трафіку та виявлення відхилень від нормальної поведінки: машинне навчання пропонує ефективне рішення цієї проблеми, виявляючи аномалії, що відхиляються від типових моделей поведінки мережі [1]. Відповідно самі моделі також стають об’єктами атак: зловмисник може впливати на навчальні дані, вхідні параметри або процес формування рішень, знижуючи здатність системи розпізнавати загрози.

Мета доповіді. Дослідити механізми цілеспрямованої деградації моделей машинного навчання протягом життєвого циклу, визначити їхні наслідки для кібербезпеки й узагальнити підходи до підвищення стійкості моделей.

Дослідження цілеспрямованої деградації моделей машинного навчання також перебуває в площині аналізу вразливостей та управління ризиками штучного інтелекту. Нижченаведені механізми атак, зокрема отруєння навчальних даних, цілеспрямоване спотворення вхідної інформації та впровадження зловмисних інструкцій, можуть розглядатися як сценарії реалізації ризиків під час використання ШІ-систем. Їх оцінювання доцільно здійснювати з урахуванням рекомендацій міжнародних стандартів ISO/IEC 27005:2022, який визначає підходи до управління ризиками інформаційної безпеки, та ISO/IEC 23894:2023, присвяченого управлінню ризиками штучного інтелекту. Такі рішення надають можливість розглядати стійкість моделей як технічну характеристику та складову загальної системи управління інформаційними ризиками [2, 3].

Цілеспрямована деградація моделі – це погіршення її роботи внаслідок навмисного втручання зловмисника в дані, процес навчання або вхідну інформацію. Важливо розмежовувати атаки, спрямовані на зміну поведінки моделі через втручання у навчальні дані, та атаки, що експлуатують особливості її функціонування під час обробки вхідної

інформації. До першої категорії належить отруєння навчальних даних, атаки ухилення та атака на основі впровадження інструкцій спрямовані переважно на маніпулювання результатами роботи вже розгорнутих моделей.

Одним з головних механізмів деградації є отруєння навчальних даних: ідея полягає у внесенні зловмисником модифікованих або спеціально сформованих записів до навчальної вибірки, унаслідок чого модель засвоює хибні залежності. У системах кібербезпеки це може призвести до класифікації шкідливої активності як легітимної або нормальної поведінки як загрози. Отруєння навчальних даних може знизити точність класифікації до 27% у задачах розпізнавання зображень і до 22% у моделях виявлення шахрайства [4]. Загалом, ці показники отримано не для SIEM-систем, однак вони демонструють потенційний масштаб впливу на якість класифікації.

Небезпека отруєння полягає також у тому, що модель може зберігати зовнішню працездатність, але пропускати певні категорії атак під час експлуатації. Ця процедура перенавчання створює для зловмисника можливості імітувати сигнали пошкоджених датчиків на етапі навчання та спотворити процес навчання детектора таким чином, що атаки залишатимуться невиявленими під час розгортання системи [5].

Інший механізм – атаки ухилення, спрямовані на вхідні дані розгорнутої моделі. Зловмисник змінює параметри об’єкта так, щоб система надала неправильний результат класифікації. Це тип атаки, під час якої зловмисник маніпулює вхідними даними алгоритму машинного навчання з метою виявлення малих відхилень, що призводять до значних змін у вихідних даних, наприклад, помилок у прийнятті рішень [6]. У кібербезпеці подібні маніпуляції можуть допомагати обходити виявлення вторгнень або маскувати шкідливу активність у мережевому трафіку. При цьому приклади з навмисно внесеними спотвореннями мають враховувати особливості

даних, характерних для кібербезпеки: суперечливі приклади повинні відповідати обмеженням, що випливають із семантики додатка та способу представлення ознак, як-от мережевий трафік або бінарні файли програм [5].

Для генеративних моделей окремим ризиком є маніпуляції вхідними даними та інструкціями. Зокрема, методи маніпулювання вхідними даними моделей передбачають створення вхідних даних, аби вони зберігали вигляд звичайної, нешкідливої інформації, але водночас впливали на результати роботи моделі [7]. Одним із проявів є Prompt Injection – використання спеціально сформованих інструкцій для зміни поведінки мовної моделі. Більшість захисних стратегій не залежать від конкретної моделі і тому не вимагають її навчання, а також доступу до внутрішніх механізмів цільової моделі. Такі підходи становлять 63.63% розглянутих робіт; серед поширених засобів захисту визначено додаткове налаштування безпеки моделі після її навчання, а також фільтрацію вхідних і вихідних даних. [8]. У системах кібербезпеки така атака може реалізовуватися через розміщення зловмисних інструкцій у журналах подій або інших даних, які передаються генеративній моделі для аналізу. У результаті модель може проігнорувати задані правила обробки інформації, сформувати хибний висновок щодо інциденту або надати необґрунтовані рекомендації щодо реагування, що знижує достовірність автоматизованого аналізу.

Наслідком цілеспрямованого впливу на модель може бути зниження ефективності виявлення кіберзагроз. Хибнопозитивні результати змушують фахівців перевіряти легітимну активність як підозрілу, а хибнонегативні – призводять до пропуску реальних атак. Помилкові спрацьовування є однією з найважливіших проблем у системах виявлення аномалій [1]. Особливо це стосується моделей, що можуть позначати нормальну поведінку як шкідливу та перевантажувати команду безпеки сповіщеннями: моделі машинного навчання, особливо в рамках неконтрольованого навчання,

можуть помилково розпізнавати нешкідливу поведінку як зловмисну, що призводить до перевантаження команди з безпеки сповіщеннями [1]. Водночас атаки типу «adversarial» можуть суттєво знизити ефективність виявлення, збільшити частку помилкових відмов та помилкових спрацьовувань, маніпулювати межами прийняття рішень та знизити впевненість щодо надійності автоматизованих систем безпеки [9].

Під час експлуатації генеративних моделей також можуть виникати галюцинації, тобто формування правдоподібної, але недостовірної інформації. Явище не є безпосереднім наслідком цілеспрямованих атак, однак може посилювати їхні негативні наслідки, спричиняючи помилкову інтерпретацію подій і формування необґрунтованих рекомендацій у системах кібербезпеки.

Головна проблема полягає в тому, що висока точність прогнозування не завжди гарантує стійкість до суперечливих атак [9]. Тому оцінювання моделей має враховувати і стандартні показники класифікації, і здатність зберігати коректність за умови маніпуляцій навчальними та вхідними даними.

Захист моделі має охоплювати перевірку даних перед навчанням та безпосередньо контроль її роботи після розгортання. Статистичне виявлення аномалій допомагає виявляти підозрілі зміни в даних, а навчання на прикладах із навмисно внесеними спотвореннями – підвищувати стійкість моделі шляхом включення спеціально сформованих прикладів до навчання. Виявлення статистичних аномалій та суперечливе навчання успішно пом’якшили наслідки отруєння та відновили точність у середньому на 15–20%. Додатковий рівень стійкості забезпечує ансамблеве навчання, яке може зменшувати кількість помилок, спричинених адверсарними ін’єкціями даних [4].

Після розгортання моделі необхідно безперервно контролювати її поведінку. Рекомендовано визначати й відстежувати показники належного

функціонування моделі, зокрема зміни її поведінки, що дає змогу швидше виявляти аномалії [6]. Також має охоплюватись захищене керування даними, валідацію моделі, стійке формування ознак, пояснюваність, моніторинг, людський нагляд і регулярний аудит безпеки [9]. Потенційно небезпечні рекомендації та дії здатні вплинути на доступність або захищеність інфраструктури, мають перевірятися фахівцями; автономне виконання небезпечних операцій доцільно обмежувати.

Цілеспрямоване погіршення функціонування моделей машинного навчання може відбуватися на етапах підготовки даних, навчання та експлуатації. Основними механізмами такого впливу є отруєння навчальної вибірки, цілеспрямоване спотворення вхідних даних і впровадження зловмисних інструкцій у запити до генеративних моделей. Їхні наслідки охоплюють зниження точності, пропуск кіберзагроз, збільшення кількості хибних спрацювань і формування недостовірних рекомендацій. Забезпечення стійкості моделей потребує комплексного підходу, що поєднує контроль даних, навчання на прикладах із навмисно внесеними спотвореннями, використання сукупності моделей, перевірку їхньої коректності, постійний моніторинг функціонування та нагляд фахівців. Водночас оцінювання відповідних загроз і впровадження захисних заходів доцільно здійснювати в межах загальної системи управління ризиками інформаційної безпеки з урахуванням специфіки ШІ-систем.

Список використаних джерел

1. Kulia V., Petrenko O. Detection of anomalous actions in the network based on machine learning. Computer and information systems and technologies : Seventh International Scientific and Technical Conference. Kharkiv : NURE, 2024. P. 40–41. URL: https://openarchive.nure.ua/entities/publication/f3d84a73-839c-4aa7-8b97-95f9a5308c10;
2. ISO/IEC 27005:2022. Information security, cybersecurity and privacy protection – Guidance on managing information security risks. 2022;
3. ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management. 2023;

4. Kure H. I., Sarkar P., Ndanusa A. B., Nwajana A. O. Detecting and Preventing Data Poisoning Attacks on AI Models. 2025 PhotonIcs & Electromagnetics Research Symposium (PIERS). Abu Dhabi, 2025. 9 p. URL: https://arxiv.org/abs/2503.09302;
5. Oprea A., Vassilev A. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Gaithersburg : National Institute of Standards and Technology, 2023. NIST AI 100-2e2023 ipd. 75 p. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023.ipd;
6. European Union Agency for Cybersecurity. Securing Machine Learning Algorithms. 2021. 70 p. DOI: https://doi.org/10.2824/874249. URL: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Securing%20Machine%20Learning%20Algorithms.pdf;
7. National Cyber Security Centre. Understanding adversarial attacks against Machine Learning and AI. 2026. 14 p. URL: https://www.ncsc.gov.uk/sites/default/files/2026-08/Understanding-adversarial-attacks-against-Machine-Learning-and-AI.pdf;
8. Correia P. H. B., Achjian R. W., Caetano de Oliveira D. E. G., Maria Y. A., Hayashi V. T., Lopes M., Miers C. C., Simplicio M. A. A Systematic Literature Review on LLM Defenses Against Prompt Injection and Jailbreaking: Expanding NIST Taxonomy. 2026. arXiv:2601.22240. 27 p. URL: https://arxiv.org/abs/2601.22240;
9. Tochukwu N. B., Ezeaku-Ezeme C. Adversarial Attacks on Machine Learning-Based Cybersecurity Classifiers: A Systematic Analysis and Defense Framework. International Journal of Research and Scientific Innovation. 2026. Vol. 13, №7. P. 5867–5898. DOI: https://doi.org/10.51244/IJRSI.2026.1307000430.