УДК 004.056:004.891

Дричак Б.Я., студент
Вовк Р.Б., к.т.н., доцент
Івано-Франківський національний технічний університет нафти і газу

МЕТОДОЛОГІЯ ЗАБЕЗПЕЧЕННЯ СТІЙКОСТІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ У ІНФОРМАЦІЙНО-КОМП’ЮТЕРНИХ СИСТЕМАХ

Широка імплементація глибоких нейронних мереж (DNN) у сучасну архітектуру інформаційно-комп’ютерних систем зумовлює критичну необхідність дослідження їхньої резистентності до адверсаріальних впливів. У період 2024–2026 рр. характер загроз еволюціонував від примітивних інвазій у вхідні дані до комплексних деструктивних маніпуляцій логікою функціонування та інтерпретованістю моделей. Зазначена тенденція трансформує проблему робастності нейромережевих структур у фундаментальний аспект парадигми кібербезпеки.

Класифікація векторів атак традиційно здійснюється за критерієм рівня інформованості суб'єкта впливу про параметри цільової моделі:

- Сценарій «білої скриньки» передбачає повний доступ до внутрішньої архітектури та градієнтів моделі. Найбільш ефективними в даному контексті є градієнтні методи, зокрема Fast Gradient Sign Method та Projected Gradient Descent.

- Сценарій «сірої скриньки» базується на частковій інформації про систему та використанні властивості переносності через моделі.

- Сценарій «чорної скриньки» обмежується аналізом вхідних та вихідних сигналів. Детерміновано зростання ефективності таких атак за рахунок інтеграції великих мовних моделей (LLM) як інструментів генерації запитів [1].

Рівень робастності моделей демонструє суттєву кореляцію з предметною областю. В задачах комп’ютерного зору інтенсифікація глибини мережі сприяє покращенню генералізації, проте в системах виявлення мережевих вторгнень (NIDS) надмірна глибина архітектури може ініціювати деградацію стійкості внаслідок дискретної природи вхідних параметрів [2]. Це обґрунтовує доцільність застосування доменно-орієнтованого підходу при проєктуванні нейромережевих компонентів. Окрему категорію становлять атаки на відмову в обслуговуванні, зокрема «губчасті атаки» (Sponge Attacks). Їхня специфіка полягає в цілеспрямованому максимізуванні обчислювальної складності обробки даних, що призводить до виснаження апаратних ресурсів, деградації енергоефективності мобільних пристроїв та підвищення латентності в хмарних сервісах.

Ескалація ризиків спостерігається в автономних транспортних системах та медичній діагностиці. У першому випадку адверсаріальний вплив здатний спричинити колапс логіки прийняття рішень, у другому – спотворити результати інтерпретації діагнозу, нівелюючи довіру до автоматизованих висновків. Розвиток LLM актуалізував загрози, пов’язані з автоматизованою генерацією дезінформації та маніпулятивним впливом на результати пошукової видачі. Емпірично доведено низьку стійкість таких моделей до числових пертурбацій, що вказує на лімітованість їхніх когнітивних спроможностей щодо логічного виведення. Для об'єктивізації оцінювання запропоновано індекс різниці стійкості. Дана метрика базується на аналізі топології внутрішнього простору ознак і дозволяє диференціювати якість розділення класів незалежно від специфіки алгоритму атаки.

Сучасний інструментарій захисту включає методи очищення представлень (representation scrubbing), дифузійні моделі для відновлення вихідних даних та алгоритми сингулярного розкладу (SVD) для фільтрації шумів. Перспективним вектором досліджень є розробка сертифікованих методів навчання (certified robustness), що гарантують стабільність виходу в межах заданого радіуса збурення [3].

Проте встановлено наявність оберненої залежності між точністю (accuracy) та робастністю (robustness). Формування оптимального балансу між цими показниками є першочерговим завданням при проєктуванні систем для конкретних галузей.

Отже, забезпечення стійкості DNN є комплексним науково-технічним завданням, що вимагає синергії новітніх метрик оцінювання, адаптивних архітектурних рішень та верифікованих механізмів захисту. Подальша парадигма розвитку галузі полягає у створенні універсальних фреймворків аналізу надійності нейромережевих систем, орієнтованих на експлуатацію в критично важливих сегментах економіки та безпеки.

Список використаних джерел

1. White, Gray & Black-Box Attacks Overview [Електронний ресурс]. URL: https://surl.lu/juyczy (дата звернення: 20.03.2026).

2. Exploring the Effect of DNN Depth on Adversarial Attacks in Network Intrusion Detection Systems [Електронний ресурс]. URL: https://arxiv.org/html/2510.19761v1 (дата звернення: 20.03.2026).

3. RDI: An adversarial robustness evaluation metric for deep neural networks based on model statistical features [Електронний ресурс]. URL: https://arxiv.org/html/2504.18556v2 (дата звернення: 20.03.2026).