4. Valstar M. et al. Affect recognition in HCI. In Human-Centered AI. Springer, 2020.
5. Huang J. et al. Real-time AI-enhanced UX analytics. IJHCS, 2023.

УДК 004.8
Г.І. Липак, к.н.с.к., доцент, Д.А. Коломийчук
Тернопільський національний технічний університет імені Івана Пулюя

ВПЛИВ БЕКДОР АТАК НА МОЖЛИВОСТІ НАВЧАННЯ НЕРОННИХ МЕРЕЖ

H. Lypak, Ph.D., D. Kolomyichuk
Ternopil Ivan Puluj National Technical University

THE IMPACT OF BACKDOR ATTACKS ON THE LEARNING CAPABILITY OF NEURAL NETWORKS

У роботі було проаналізовано непомітні атаки зловмисників на дані часових рядів у рекурентних нейронних мережах, щоб вивчити як безпеку глибоких рекурентних нейронних мереж, так і зрозуміти властивості навчання в глибоких рекурентних нейронних мережах.

Оскільки глибокі нейронні мережі широко використовуються в прикладних областях, існує можливість зниження точності та безпеки за допомогою методів зловмисників. Метод зловмисників - це отруєння даних через бекдор, коли зловмисник збиває навчальні вибірки невеликим збуренням, щоб неправильно класифікувати вихідний клас до цільового класу. При зміні даних через бекдор зловмисник має доступ до підмножини навчальних даних з мітками, можливість впливати на навчальні вибірки та можливість змінити мітку вихідного класу на мітку цільового класу. Зловмисник не має доступу до класифікатора під час навчання або знання процесу навчання. Також було досліджено захист від зміни даних через бекдор після навчання, розглядаючи ітераційний метод для визначення пари вихідного та цільового класів у такій атаці.

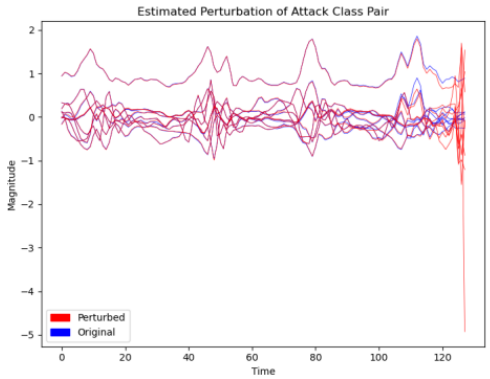

Рисунок 1 - Очікуване збурення пари класів атаки, LSTM

Методи атак через бекдор, успішно обманюють класифікатор LSTM, не знижуючи точності тестових зразків, без наявності шаблону бекдору. По-друге, метод захисту успішно визначає пару вихідних класів в такій атаці. По-третє, отруєння через бекдор у LSTM вимагає або більше навчальних зразків, або більшого збурення, ніж стандартна мережа прямого зв'язку.

Література.

2. B. Tran, J. Li and A. Madry. Spectral Signatures in Backdoor Attacks // NeurIPS. 2018. URL: https://arxiv.org/abs/1811.00636.