

Список використаних джерел

1. Almtorf, A. (2025). Integrating explainable AI (XAI) into decision support systems: A framework for enhancing transparency and trust in managerial decision-making. *International Journal of Managerial Studies and Research*. doi:10.20431/2349-0349.1309002
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. doi:10.1016/j.inffus.2019.12.012
3. Goyal, B. (2025). Explainable AI (XAI) for cloud-based enterprise applications: Building trust and transparency in AI-driven decisions. *Journal of International Crisis and Risk Communication Research*. doi:10.63278/jicrcr.vi.3231
4. Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable artificial intelligence-based decision support systems: A recent review. *Electronics*, 13(14), 2842. doi:10.3390/electronics13142842
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). doi:10.1145/2939672.2939778
7. Sabharwal, R., Miah, S. J., Wamba, S. F., et al. (2025). Extending application of explainable artificial intelligence for managers in financial organizations. *Annals of Operations Research*, 354, 309–339. doi:10.1007/s10479-024-05825-9
8. Павлов, А. (2024). Порівняльний аналіз моделей машинного навчання для визначення ключових факторів успіху відкриттів роздрібних магазинів в Україні. *Grail of Science*, 43, 90–104. <https://doi.org/10.36074/grail-of-science.06.09.2024.010>

Позднякович О.Є.,

к.ф.-м.н., доцент,

Київський національний економічний університет імені Вадима Гетьмана

МОДЕЛЮВАННЯ СИСТЕМИ ПОВЕДІНКОВОГО АНАЛІЗУ ДЛЯ ПРОТИДІЇ АВТОМАТИЗОВАНИМ КІБЕРЗАГРОЗАМ

Сучасні бізнес-платформи стикаються з безпрецедентною кількістю автоматизованих атак на системи автентифікації. Методи статичного захисту, засновані на жорстких обмеженнях кількості спроб, усе частіше виявляються неефективними проти інтелектуальних ботів, які імітують людську поведінку та змінюють IP-адреси (Cloudflare, 2023). Внаслідок цього виникає необхідність переходу від простого підрахунку запитів до оцінювання «поведінкового почерку» суб'єкта. (Monrose & Rubin, 2000). Такий підхід реалізовано через створення імітаційної моделі системи захисту на базі концепції «інтегратора з витоком» (Leaky Integrator). Модель накопичує інформацію про аномалії з урахуванням частоти їх виникнення та має механізм «періоду охолодження». (Aström & Murray, 2010).

Динаміку рівня ризику $R(t)$ можна описати за допомогою диференціального рівняння:

$$\frac{dR(t)}{dt} = -k R(t) + S \sum_{i=1}^{N(t)} \delta(t - t_i) D_i,$$

де k – коефіцієнт адаптивності (охолодження) системи, S – коефіцієнт чутливості до відхилень, $N(t)$ – кількість спроб автентифікації за час t , $\delta(t - t_i)$ – входні імпульси запитів у моменти часу t_i , D_i – ступінь відхилення ознаки від профілю.

Для реалізації в програмному середовищі Python використовується дискретна форма цифрового фільтра (Smith, 1997). Для забезпечення нормалізації приросту ризику використано ЕМА-подібну форму (експоненційне ковзне середнє), яка перераховується з фіксованим кроком $dt = 0,5$ с., що дозволяє системі «охолоджувати» рівень тривоги між запитами:

$$R_{curr} = \alpha \cdot R_{prev} + (1 - \alpha) \cdot \Delta \cdot S \quad (1)$$

де $\alpha = e^{-kdt}$ визначає інерційність системи. Відхилення інтервалу вводу даних V та ентропії E нормалізуються відносно еталонних значень:

$$\Delta = w_1 \frac{|V_{norm} - V_l|}{V_{norm}} + w_2 \frac{|E_{norm} - E_l|}{E_{norm}}, \quad V_{norm} > 0, E_{norm} > 0,$$

де w_1, w_2 — вагові коефіцієнти. Після кожного кроку значення ризику обмежується діапазоном $R \in [0\%; 100\%]$, що забезпечує механізм адаптивного обмеження швидкості (Brown, 2004). Якщо за один крок надходить N запитів, за припущення однакової аномальності всіх імпульсів, формула (1) може бути апроксимована співвідношенням:

$$R_{new} = \alpha^N \cdot R_{prev} + (1 - \alpha^N) \cdot \Delta \cdot S.$$

Параметри моделі обираються на основі балансу між чутливістю до загроз (True Positives) та мінімізацією хибних блокувань легітимних користувачів (False Positives). Зокрема, коефіцієнт інерції ($\alpha = 0,9$) та чутливості ($S = 480$) підібрані так, щоб система "пробачала" поодинокі мережеві затримки (завдяки швидкому охолодженню), але досягала насичення ризику за 3–5 послідовних кроків при високих значеннях аномальності. При $\alpha = 0,9$ і $dt = 0,5$ коефіцієнт адаптивності (охолодження) системи $k \approx 0,211 \text{ с}^{-1}$. Вагові коефіцієнти $w_1 = w_2 = 0,5$ відображають початкову рівнозначність часових та ентропійних відхилень.

Система безперервно аналізує час між запитами та ентропію їх розподілу, яка обчислюється за формулою Шеннона на ковзному вікні часових інтервалів між запитами, згрупованих у гістограму з 22 бінів:

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i,$$

де p_i — ймовірність потрапляння часового інтервалу в i -тий сегмент гістограми.

Еталоном (профілем) є середній інтервал 8 с. та типове значення ентропії для людської поведінки близько 3 біт. Ентропія оцінюється для гістограми з 22 бінів, що охоплює найбільш значущий діапазон спостереження від 0 до 11 с. із кроком дискретизації системи ($dt = 0,5$ с.). Абсолютний максимум ентропії для такої системи становить $\log_2 22 \approx 4,5$ біт (що відповідає ідеально рівномірному розподілу). Людська поведінка зазвичай характеризується нерівномірним розподілом часових інтервалів, тому її ентропія нижча за максимальну. Відхиленням (Δ) вважається як занадто низька ентропія (механічна рівномірність скрипта з низькою ентропією, що прямує до 0), так і занадто висока (штучна хаотичність бота-рандомізатора з ентропією, наближеною до максимального значення $\log_2 22 \approx 4,5$ біт).

Рішення приймається за трирівневою логікою: зелена зона (ризик $< 80\%$ — доступ дозволено), сіра зона ($80\text{--}90\%$ — вимога пройти CAPTCHA), червона зона ($> 90\%$ — блокування IP).

Поріг у 80% встановлено як оптимальну межу: природний мережевий шум або повільні помилки людини генерують ризик не вище $30\text{--}50\%$, тоді як діапазон $80\text{--}90\%$ слугує безпечним буфером для перевірки сумнівних патернів через CAPTCHA без застосування жорсткого бану.

У ході експериментів було протестовано чотири сценарії атаки (табл. 1). Експерименти проводилися шляхом генерації часових послідовностей запитів із різними параметрами інтервалів та ентропії.

Експериментальні метрики реагування системи

Сценарій атаки	Середній інтервал (с)	Час детекції (с)	Запитів до реакції	Макс. ризик	Рішення системи
Брутфорс	~0,1	~0,5 – 1	≈ 5 – 10	100%	Блокування IP
Креденшл-стафінг	~0,3	~1 – 1,5	≈ 3 – 5	100%	Блокування IP
Викрадення акаунта	~4	–	–	< 30%	Доступ дозволено
Розподілена атака	~2	~10 – 12	≈ 6	< 90%	Виклик САРТСНА

Джерело: розраховано автором.

Агресивні автоматизовані атаки виявляються майже миттєво. Натомість викрадення акаунта — найскладніший сценарій, де бот імітує людину (інтервал ~4 с). Відхилення залишається малим ($\Delta = 0,25$), а між запитами відбувається 8 кроків охолодження ($\alpha^8 \approx 0,43$). Оскільки для поодиноких запитів, розділених k кроками системи, рекурсія має вигляд $R_{new} = \alpha^k \cdot R_{prev} + (1 - \alpha) \cdot \Delta \cdot S$, стаціонарний рівень ризику становить $R_{peak} = \frac{(1-\alpha) \cdot \Delta \cdot S}{1-\alpha^8} \approx 21,1\%$, тобто в рамках базової моделі цей сценарій не перевищує поріг виявлення. У випадку розподіленої атаки (інтервал ~2 с.) надмірна випадковість наближає ентропію до максимального значення, тому $\Delta \approx 0,625$. Приріст ризику частково компенсується охолодженням у 4 кроки ($k = 4, \alpha^4 \approx 0,66$). Стаціонарний рівень ризику становить $R_{peak} = \frac{(1-\alpha) \cdot \Delta \cdot S}{1-\alpha^4} \approx 87\%$, що перевищує поріг 80% і спричиняє виклик САРТСНА, блокуючи загрозу без жорсткого бану.

Впровадження моделі «інтегратора з витоком» забезпечило реалізацію адаптивного механізму оцінювання поведінки, що дозволило зменшити кількість помилкових блокувань легітимних користувачів при збереженні високого рівня безпеки. Для виявлення повільних атак (з інтервалом > 4 с.) система може бути розширена за рахунок поведінкових ознак, таких як геолокація або цифровий відбиток пристрою.

Список використаних джерел

1. Cloudflare. (2023). State of the Internet / Security: Credential Stuffing Trends. Retrieved from <https://www.cloudflare.com/learning/bots/what-is-credential-stuffing/>
2. Monroe, F., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4), 351-359.
3. Åström, K. J., & Murray, R. M. (2010). *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press.
4. Smith, S. W. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing.
5. Brown, R. G. (2004). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Dover Publications.