

СОЦІАЛЬНА ІНЖЕНЕРІЯ В ЕПОХУ DEEPFAKES: АНАЛІЗ ЗАГРОЗ ТА МЕТОДИ ПРОТИДІЇ НА ОСНОВІ СИНТЕЗУ ДАНИХ

Вінницький національний технічний університет

Анотація

У роботі досліджено еволюцію методів соціальної інженерії під впливом технологій штучного інтелекту. Проаналізовано специфіку використання діпфейків для імітації особистості та обходу систем безпеки. Запропоновано підхід до виявлення аномальної активності шляхом синтезу технічних подій із контекстними даними користувача.

Ключові слова: Соціальна інженерія, Діпфейки, Штучний інтелект, Кібербезпека, Контекстно-орієнтований підхід, Синтез даних, Генеративно-змагальні мережі (GAN), Збагачення даних, Динамічна оцінка ризику, Архітектура нульової довіри, Горизонтальне переміщення, SIEM-системи

Abstract

The paper examines the evolution of social engineering methods in the era of artificial intelligence. It analyzes the specifics of using deepfakes for identity impersonation and bypassing security systems. A proactive approach to detecting anomalous activity is proposed by synthesizing technical security events with the user's contextual data.

The study highlights the transition from traditional text-based phishing to high-tech "AI-driven Social Engineering," which utilizes real-time audio and video synthesis. The research identifies key challenges in detecting such attacks, particularly the "contextual blindness" of standard monitoring tools. The proposed method of dynamic risk assessment and data enrichment aims to improve the accuracy of threat detection and reduce the rate of false positives in modern security operations centers

Keywords: Social Engineering, Deepfakes, Artificial Intelligence (AI), Cybersecurity, Context-Aware Approach, Data Synthesis, Generative Adversarial Networks (GAN), Data Enrichment, Dynamic Risk Scoring, Zero Trust Architecture, Lateral Movement SIEM Systems

Вступ

Стрімкий розвиток штучного інтелекту у 2026 році призвів до якісної трансформації загроз у сфері кібербезпеки. Традиційні методи соціальної інженерії еволюціонували у високотехнологічну галузь — «AI-driven Social Engineering», де основним інструментом стали діпфейки (Deepfakes). Використання генеративно-змагальних нейромереж (GAN) дозволяє зловмисникам створювати гіперреалістичні підробки аудіо- та відеоданих, які успішно експлуатують довіру користувачів та обходять класичні системи ідентифікації.

Актуальність даної теми зумовлена виникненням «кризи довіри» в цифрових каналах зв'язку. Зловмисники все частіше використовують клоновані голоси та відеомаски для імітації керівництва (Executive Impersonation), що створює критичні ризики для корпоративного сектору. Традиційні засоби захисту виявляються малоефективними проти таких атак через «контекстну сліпоту» систем моніторингу, які не здатні розрізнити легітимну взаємодію від синтезованого контенту в реальному часі.

Об'єктом дослідження є процеси реалізації атак соціальної інженерії з використанням діпфейків. Предметом — методи виявлення синтезованого контенту та стратегії захисту на основі синтезу технічних подій із контекстними даними. Метою роботи є аналіз сучасних Deepfake-загроз та обґрунтування переходу до контекстно-орієнтованих моделей захисту для ефективної протидії соціальній інженерії нового покоління.

Результати дослідження

1. Проблематика виявлення Deepfake-атак

Процес ідентифікації атак, що базуються на використанні діпфейків, становить складну науково-технічну проблему через високу реалістичність синтезованого контенту. В основі сучасних Deepfake-технологій лежить архітектура генеративно-змагальних мереж (GAN), де одна нейромережа

намагається створити фальсифікацію, а інша — виявити її, що в процесі навчання призводить до створення контенту, який неможливо розрізнити людськими органами чуття.

Основними чинниками, що ускладнюють своєчасне реагування на такі інциденти, є:

Неефективність сигнатурних методів захисту: При реалізації горизонтального переміщення (Lateral Movement) з використанням діпфейків дії зломисників не мають чітких технічних ознак зламу, а проявляються як легітимна активність адміністратора чи керівника.

Формати та збір даних: Події безпеки часто зберігаються у форматах (наприклад, EVTX), які потребують попередньої обробки та конвертації у структуровані дані для подальшого аналізу алгоритмами машинного навчання.

Проблема «контекстної сліпоти»: Класичні правила кореляції аналізують технічний факт успішного з'єднання або запуску процесу ізольовано від бізнес-логіки, що не дозволяє відрізнити реальний сеанс від маніпулятивного.

Лавиноподібне зростання помилкових спрацювань (False Positives): Спроба підвищити чутливість стандартних систем моніторингу призводить до феномену «втоми від тривоги» (Alert Fatigue), коли критичні аномалії ігноруються аналітиками через великий обсяг інформаційного «шуму».

Особливу небезпеку становлять атаки типу «Low and Slow», де зломисники розтягують свої маніпуляції у часі, маскуючись під щоденні робочі процеси. Використання мов програмування з широким спектром бібліотек для Data Science (таких як Python) є необхідним для реалізації складних алгоритмів статистичного аналізу та виявлення відхилень від моделі легітимної поведінки, що недоступно в стандартних «коробкових» рішеннях. Таким чином, головна проблема полягає у необхідності автоматизованого синтезу технічних логів із організаційними даними для побудови точного еталону «нормальної поведінки».

2. Застосування контекстно-орієнтованого підходу (Context-Awareness) для виявлення синтетичних загроз

Для подолання обмежень класичних методів детекції, які демонструють «сліпоту до контексту», пропонується впровадження контекстно-орієнтованого підходу (Context-Awareness). Цей підхід передбачає трансформацію процесу аналізу: перехід від перевірки ізольованих медіафайлів чи сесій до аналізу багатовимірних інформаційних об'єктів у режимі реального часу. Реалізація підходу здійснюється через механізм збагачення даних (Data Enrichment), що дозволяє автоматично доповнювати технічні логи метаданими з зовнішніх джерел (Active Directory, HR-системи, CMDB)

В рамках архітектури захисту від діпфейків пропонується виділити три ключові рівні контексту:

- Контекст ідентичності (Identity Context): Базується на інтеграції з кадровими системами для врахування ролі, посадових обов'язків та статусу співробітника. Наприклад, відеодзвінок із запитом на доступ до бази даних від маркетолога є аномальним, навіть якщо візуальний образ та голос повністю відповідають його особистості.
- Контекст активів (Asset Context): Визначає критичність інформаційних систем та вузлів мережі. Система автоматично підвищує пріоритет перевірки на «глибину синтезу» (liveness detection), якщо запит на маніпуляції з даними стосується контролерів домену або фінансових серверів.
- Ситуаційний та часовий контексти (Environmental & Temporal Context): Передбачають формування еталону «нормальної поведінки» (Baseline) на основі історичних даних. Аналіз відхилень у типовій геолокації входу, пристрої підключення або часових рамках активності дозволяє виявити ознаки використання діпфейк-масок, які зазвичай застосовуються зломисниками у нетипові для користувача періоди.

Застосування такого багатоваріантного аналізу перетворює систему моніторингу на інтелектуальне середовище, здатне виявляти приховані маневри зломисника на основі поведінкових аномалій, а не лише жорстких технічних правил. Це дозволяє ідентифікувати складні атаки, де візуальна та звукова складові здаються легітимними, але організаційний контекст вказує на високу ймовірність компрометації каналу зв'язку.

3. Оптимізація правил кореляції SIEM шляхом динамічного синтезу та оцінки ризиків

Після проведення аналізу вразливостей стандартних систем захисту, пропонується вдосконалена модель правил кореляції. Вона базується на переході від спрощеної бінарної логіки («тривога/ігнорування») до методу динамічної оцінки ризику (Risk Scoring) на основі синтезованих даних. На відміну від традиційного підходу, де інцидент створюється автоматично при досягненні певного порогу подій, оптимізована модель використовує алгоритм зваженого ризику.

В межах цієї моделі технічна вага кожної події (наприклад, відеодзвінок або вхід у систему) коригується коефіцієнтом аномальності організаційного контексту:

- Мінімальний ризик: Якщо адміністративну дію або сеанс зв'язку ініціює профільний фахівець у межах свого робочого часу з легітимного сегмента мережі, система не генерує тривогу, відфільтровуючи легітимну активність як інформаційний шум.
- Середній ризик: Виконання аналогічної дії користувачем із невідповідною роллю або з нетипового пристрою призводить до негайного зростання показника ризику. У такому разі система автоматично активує додаткові протоколи верифікації (наприклад, запит на liveness detection).
- Критичний ризик: Спроба доступу до критично важливих активів (контролери домену, фінансові системи) у неробочий час із використанням медіапотуку, що має ознаки цифрового синтезу, призводить до експоненціального зростання показника ризику та негайного створення пріоритетного інциденту для SOC.

Такий підхід дозволяє перетворити SIEM-систему з інструменту пасивного накопичення логів на проактивний засіб виявлення прихованих маневрів зловмисника. Навіть при використанні ідеального дідфейку, поведінковий патерн атаки неминуче відхилитиметься від нормального профілю співробітника, що зафіксує алгоритм зваженого ризику. Це значно підвищує точність виявлення етапу «Lateral Movement» та вирішує проблему «втоми від тривоги» (Alert Fatigue), фокусуючи увагу аналітиків виключно на подіях із високим ризик-балом.

4. Стратегія комплексного захисту та проактивного моніторингу в умовах синтетичних загроз

Ефективна протидія атакам соціальної інженерії, що базуються на використанні дідфейків, вимагає перегляду парадигми захисту від пасивного накопичення логів до проактивного управління інцидентами. Враховуючи обмеженість класичних методів, стратегія безпеки у 2026 році повинна базуватися на поєднанні технологічних інструментів та організаційних протоколів.

Ключовими елементами такої стратегії є:

- Впровадження архітектури нульової довіри (Zero Trust Architecture): Жоден сеанс зв'язку, незалежно від візуальної чи звукової автентичності, не вважається довіреним за замовчуванням. Кожна критична операція повинна підтверджуватися через два або більше незалежних каналів (Multi-Channel Verification), що нівелює успіх атаки навіть при повній компрометації одного з медіа-каналів.
- Інтеграція ШІ-детекторів у реальному часі: Використання спеціалізованих алгоритмів для аналізу артефактів цифрового стиснення та мікро-виразів обличчя, які дозволяють ідентифікувати синтетичний контент безпосередньо під час відеоконференцій.
- Автоматизація парсингу та конвертації даних: Оскільки сирі логи (наприклад, формат EVTХ) непридатні для прямої обробки алгоритмами машинного навчання, критично важливим є створення конвеєрів автоматизованого перетворення даних у структуровані формати (CSV, JSON) для подальшої кластеризації та виявлення аномалій.
- Security Awareness Training 2.0: Проведення регулярних тренінгів для персоналу, спрямованих на розвиток критичного мислення та вивчення методів ручної перевірки співрозмовника (наприклад, прохання змінити ракурс камери або задати питання, що вимагає специфічного спільного контексту, недоступного ШІ).

Реалізація запропонованого методу синтезу технічних та організаційних даних дозволяє трансформувати SIEM-систему в інтелектуальний засіб захисту. Це забезпечує не лише виявлення етапу горизонтального переміщення (Lateral Movement), а й створює фундамент для побудови стійкої до зовнішніх маніпуляцій інфраструктури, здатної функціонувати в умовах постійного зростання складності методів кіберзлочинців.

Висновки

У ході проведеного дослідження було проаналізовано проблематику виявлення сучасних кіберзагроз на етапі горизонтального переміщення (Lateral Movement) в умовах активного використання зловмисниками технологій штучного інтелекту. Встановлено, що класичні методи моніторингу на базі SIEM-систем, які спираються на статичні правила кореляції та сигнатурний аналіз, втрачають свою

ефективність при протидії атакам типу «AI-driven Social Engineering» та використанні технік «Living off the Land». Головним науковим та практичним результатом роботи є наступні положення:

Визначено, що основною причиною неефективності існуючих підходів є «сліпота до контексту», що призводить до нездатності систем розрізняти легітимні дії адміністраторів і маніпуляції зловмисників, які використовують дідфейки та вкрадені облікові записи.

Обґрунтовано необхідність застосування контекстно-орієнтованого підходу (Context-Awareness), який дозволяє трансформувати ізольовані технічні події у багатовимірні інформаційні об'єкти через механізм збагачення даних (Data Enrichment).

Розроблено метод удосконалення правил кореляції шляхом синтезу технічних логів із організаційним контекстом, що включає роль користувача (Identity Context), критичність активів (Asset Context) та історичний профіль поведінки (Temporal Context).

Доведено, що перехід від бінарної логіки виявлення до моделі динамічної оцінки ризику (Risk Scoring) дозволяє автоматично відфільтровувати легітимну активність, значно знижуючи рівень помилкових спрацювань (False Positives) та усуваючи проблему «втоми від тривоги» (Alert Fatigue) у аналітиків SOC.

Запропонований комплексний підхід, що поєднує технологічні інструменти детекції синтетичного контенту та архітектуру нульової довіри (Zero Trust), забезпечує проактивний захист корпоративної інфраструктури. Це дозволяє не лише вчасно ідентифікувати приховані маневри зловмисника, а й суттєво підвищити загальний рівень кіберстійкості організації в умовах постійної еволюції методів соціальної інженерії.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. ENISA Threat Landscape 2025. European Union Agency for Cybersecurity [Електронний ресурс]. – Режим доступу: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025> (дата звернення: 16.02.2026).
2. Global Cybersecurity Outlook 2026. World Economic Forum [Електронний ресурс]. – Режим доступу: https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2026.pdf (дата звернення: 16.02.2026).
3. NISTIR 8584. Face Analysis Technology Evaluation (FATE) MORPH 4B: Considerations for Implementing Morph Detection in Operations. National Institute of Standards and Technology, 2025 [Електронний ресурс]. – Режим доступу: <https://www.nist.gov/publications/nistir-8584> (дата звернення: 16.02.2026).
4. Deepfakes 2026: чому фейки тепер майже не відрізнити від реальності [Електронний ресурс]. – Режим доступу: <https://speka.ua/artificial-intelligence/deepfakes-2026-comu-feiki-stali-maize-nevidriznimi-vid-realnosti-9wrx73> (дата звернення: 16.02.2026).

Анатолій Віталійович Довгань – студент групи 2КІТС-24б, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: dovgananatoliy2002@gmail.com;

Науковий керівник: Тетяна Геннадіївна Кирилашук – асистент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця.

Anatoliy V. Dovhan – student of group 2KITS-24b, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: dovgananatoliy2002@gmail.com;

Supervisor: Tatyana G. Kyrylashchuk – Assistant Professor of the Department of Information Protection, Vinnytsia National Technical University, Vinnytsia.