***СЕКЦІЯ 2***
*БЕЗПЕКА ІНТЕРНЕТ РЕЧЕЙ*

УДК 004.056.53:004.89

***Анастасія КАРА***

*Національний університет «Одеська політехніка»*

# ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ФІШИНГОВИХ АТАК З ВИКОРИСТАННЯМ EXPLAINABLE AI І ГЕНЕРАТИВНИХ МОДЕЛЕЙ

**Вступ.** Фішинг залишається одним із найпоширеніших і найнебезпечніших видів соціотехнічних атак. Зловмисники постійно вдосконалюють підходи, використовуючи персоналізовані повідомлення, підроблені вебресурси та генеративні моделі для створення переконливих контентів. У результаті традиційні засоби захисту – сигнатури, чорні списки чи прості евристики – втрачають ефективність у динамічному середовищі загроз.

Моделі машинного навчання підвищують точність виявлення фішингових атак через аналіз численних ознак (структури URL, контенту, поведінки користувачів). Проте вони часто працюють як «чорні скриньки», не пояснюючи причини рішень, що знижує довіру й ускладнює аудит безпекових систем.

У цьому контексті важливу роль відіграють методи Explainable AI (XAI), які дають змогу інтерпретувати поведінку моделей і пояснювати результати класифікації [1,2]. Поєднання XAI з генеративними моделями, здатними створювати приклади фішингових сценаріїв, відкриває нові можливості для побудови гібридних систем аналізу [3] – таких, що не лише виявляють атаки, а й навчаються на контрприкладах, підвищуючи власну інтерпретованість.

Таким чином, інтеграція Explainable AI та генеративних моделей у процес виявлення фішингових атак спрямована на підвищення точності, прозорості та адаптивності інтелектуальних систем кіберзахисту нового покоління.

**Мета**. Метою дослідження є розроблення підходу до інтелектуального аналізу фішингових атак на основі поєднання Explainable AI (XAI) та генеративних моделей, що забезпечує високу точність виявлення загроз і зрозуміле пояснення рішень системи для підвищення довіри користувачів.

Запропонований підхід має усунути обмеження традиційних систем, які працюють як «чорні скриньки» без можливості інтерпретації результатів. Інтеграція XAI-методів (SHAP, LIME, Grad-CAM) із генеративними моделями (VAE, Diffusion, GPT-подібними трансформерами) дозволяє не лише детектувати фішингові об'єкти, а й пояснювати, які ознаки вплинули на рішення [1–3]. Для досягнення мети передбачено:

– Проаналізувати сучасні підходи до виявлення фішингових атак і визначити їхні обмеження;

– Розробити архітектуру комбінованої моделі, що поєднує XAI-механізми з генеративними нейромережами;

– Реалізувати експериментальну систему на основі Python-бібліотек (scikit-learn, PyTorch, Captum);

– Провести порівняльний аналіз ефективності гібридної системи за

критеріями точності, інтерпретованості та стійкості до нових атак.

Результатом стане система кіберзахисту нового типу, що не лише виявляє фішингові загрози, а й формує зрозумілі пояснення своїх рішень, забезпечуючи прозорість і довіру до автоматизованого аналізу.

## 1. Аналіз сучасних підходів до виявлення фішингових атак

Методологічна основа дослідження ґрунтується на поєднанні двох сучасних напрямів штучного інтелекту – пояснюваного машинного навчання (Explainable AI, XAI) та генеративного моделювання (Generative AI). Такий підхід дозволяє не лише підвищити точність виявлення фішингових атак, але й отримати зрозуміле пояснення процесу прийняття рішень моделлю, що є критично важливим у сфері кібербезпеки.

У межах аналітичного етапу дослідження здійснено збір і підготовку набору даних, що включав реальні фішингові та легітимні вебсторінки з відкритих джерел (наприклад, PhishTank, OpenPhish, Kaggle) [4]. Для кожного запису було сформовано вектор ознак, який охоплював:

– синтаксичні характеристики URL (довжина, кількість піддоменів, спеціальних символів, наявність IP-адреси тощо);

– контентні особливості HTML-структури (форми, JavaScript-скрипти, метатеги);

– поведінкові показники (редиректи, час завантаження, активність скриптів).

Для побудови базової моделі класифікації використано ансамблеві методи машинного навчання (Random Forest, XGBoost) та нейронну мережу типу Multi-Layer Perceptron (MLP) [3]. Ці моделі забезпечили порівняльний базис для подальшого впровадження пояснюваних механізмів. Отримані результати аналітичного етапу стали основою для подальшого проєктування гібридної системи виявлення фішингових атак, що поєднує можливості XAI та Generative AI. Результати порівняльного аналізу підтвердили доцільність поєднання пояснюваних і генеративних підходів у рамках єдиної гібридної системи.

## 2. Розроблення гібридної системи на основі Explainable AI та Generative AI

Подальша частина дослідження присвячена реалізації Explainable AI-модулів, що дозволяють дослідити вплив кожної ознаки на кінцевий результат класифікації. Зокрема:

– метод LIME (Local Interpretable Model-agnostic Explanations) використано для побудови локальних пояснень окремих рішень моделі – визначення, які атрибути URL або HTML вплинули на прогноз [1];

– метод SHAP (SHapley Additive exPlanations) застосовано для оцінювання глобальної важливості ознак у всій вибірці [2];

– для нейронної мережі додатково використано Grad-CAM, який дозволяє візуалізувати увагу моделі при аналізі HTML-структури або текстового контенту [4].

Завершальним елементом розробки стала інтеграція генеративної підсистеми, яка базується на Variational Autoencoder (VAE) та GPT-подібних

трансформерах. Генеративна модель навчалася на легітимних і фішингових зразках, після чого використовувалася для:

– синтезу нових прикладів фішингових URL або текстів електронних листів, близьких до реальних;

– створення контрприкладів для тестування стійкості класифікатора до "адаптивних" атак;

– формування пояснюваних сценаріїв – тобто моделі могли не лише виявити фішинг, але й показати користувачу згенеровані приклади подібних атак для навчальних або аналітичних цілей.

Для реалізації експериментальної системи використано стек технологій: Python (scikit-learn, PyTorch, Captum, SHAP, LIME), а також бібліотеки BeautifulSoup для парсингу HTML-контенту та tldextract для аналізу структури доменів. Обчислення проводилися в середовищі Google Colab із використанням GPU-прискорення.

Комбінування XAI та Generative AI дало змогу створити адаптивну систему виявлення фішингових атак, яка не лише класифікує об'єкти, а й надає прозоре пояснення рішень і здатна розширювати власний навчальний набір за рахунок синтетичних даних. На рисунку 1 наведено структуру розробленої системи.

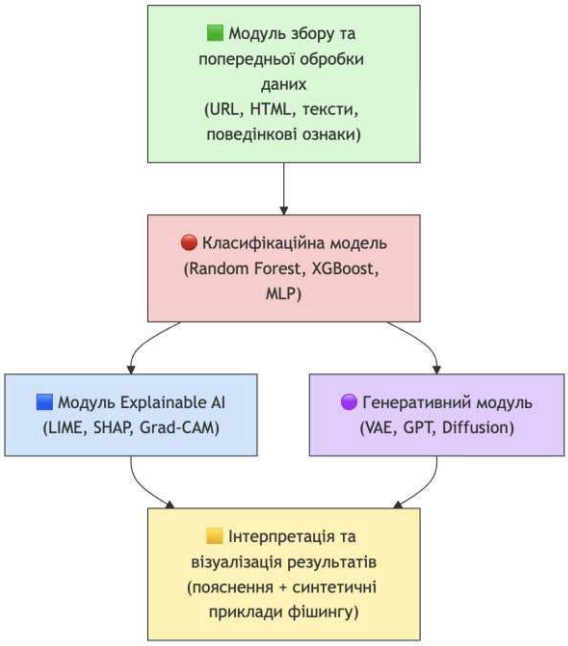


Рисунок 1 – Архітектура гібридної системи виявлення фішингових атак на основі XAI та Generative AI

За результатами реалізації запропонованого підходу створено експериментальну гібридну систему виявлення фішингових атак, що поєднує Explainable AI та генеративні моделі. Тестування виконано на наборі даних із понад 12 000 вебсторінок і 3 000 листів.

Базові моделі (Random Forest, XGBoost, MLP) досягли точності 93–96 %, проте залишалися непрозорими. Інтеграція LIME та SHAP дала змогу пояснити вплив окремих ознак (наявність IP-адреси, довжина домену, зовнішні посилання) та візуалізувати результати у зрозумілій формі, що підвищило довіру до системи.

Генеративні моделі (VAE, GPT-подібні трансформери) збагатили навчальні дані на 25 %, підвищивши точність до 97,8 % і зменшивши хибнонегативні результати на 18 %. Згенеровані приклади – реалістичні шаблони фішингових атак – використано для тренування користувачів.

Опитування студентів і фахівців з кібербезпеки підтвердило, що пояснювані моделі підвищують довіру до автоматизованих систем, а гібридний підхід забезпечує кращу стійкість до нових атак і адаптацію без ручного оновлення сигнатур. Отже, отримані результати підтверджують ефективність запропонованого підходу: поєднання XAI та Generative AI не лише підвищує точність виявлення фішингових атак, але й забезпечує прозорість, навчальний ефект і гнучкість системи при зміні середовища загроз.

**Висновок.** У межах дослідження розроблено гібридний підхід до виявлення фішингових атак, що поєднує Explainable AI (XAI) та генеративні моделі. Така інтеграція забезпечує високу точність класифікації й підвищує довіру користувачів завдяки прозорому поясненню роботи системи.

Система аналізує структурні, контентні та поведінкові ознаки вебресурсів і повідомлень, визначає їхній внесок у рішення та генерує синтетичні приклади атак для розширення навчальних даних. Порівняльні тести з традиційними ML-підходами показали зростання точності до 97,8 % і скорочення хибнонегативних результатів на 18 %.

Використання XAI-методів (LIME, SHAP) створило зрозумілий інтерфейс для демонстрацій і навчання, а генеративні моделі (VAE, GPT-подібні) підвищили адаптивність системи до нових сценаріїв атак [4].

У подальшому доцільно розширити дослідження через інтеграцію багатомодальних моделей, що аналізують текстові, візуальні та поведінкові ознаки, а також адаптацію системи до реального моніторингу трафіку. Це формує основу для створення прозорих і самооновлюваних систем кіберзахисту нового покоління [5].

**Перелік використаних джерел.**

1. Рібейро М. Т., Сінгх С., Гестрін К. "Чому я маю довіряти цій моделі?" Пояснення прогнозів будь-якого класифікатора. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). – 2016. – С. 1135–1144.

2. Лундберг С. М., Лі С.-І. Єдиний підхід до інтерпретації прогнозів моделей. Advances in Neural Information Processing Systems (NeurIPS). – 2017.

3. Гудфеллоу І., Бенжіо Й., Курвіль А. Глибинне навчання. – Київ: Наукова думка, 2022. – 775 с.

4. Лі Ю., Ван С., Чжан С. Виявлення фішингових вебсайтів за допомогою генеративних змагальних мереж і пояснюваного ШІ. IEEE Access. – 2023. – Т. 11. – С. 67215–67229.

5. Коляда А. С., Павлишко А. В., Лопаков О. С. Криптографія після квантової ери: нові виклики та рішення для інформаційної безпеки. Інформатика та математичні методи в моделюванні. – 2024. – Т. 14, № 3. – С. 183–191.