УДК 519.6 *http://dx.doi.org/10.30970/vam.2026.36.00000*

МЕТОДИ МАШИННОГО ТА ГЛИБОКОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ШКІДЛИВИХ URL-АДРЕС: СУЧАСНИЙ СТАН ТА ПЕРСПЕКТИВИ РОЗВИТКУ

В. Лесик

[1] *Львівський національний університет імені Івана Франка,*
вул. Університетська 1, Львів, 79000, Україна,
e-mail: volodymyr.lesyk@lnu.edu.ua

У статті проведено систематичний огляд сучасних наукових публікацій, присвячених проблематиці виявлення шкідливих URL-адрес за допомогою технологій штучного інтелекту. Проаналізовано еволюцію підходів: від традиційних методів чорних списків і евристичних правил, що характеризуються обмеженою адаптивністю, до високоефективних моделей машинного та глибокого навчання, здатних виявляти складні та приховані патерни в даних. Особливу увагу приділено порівняльному аналізу класичних алгоритмів (логістична регресія, SVM, ансамблеві методи) та сучасних нейромережевих архітектур. У роботі також детально класифіковано ознаки, що використовуються для навчання моделей, зокрема лексичні, хостові, контентні та зовнішні, а також проаналізовано їхню інформативність і вплив на якість класифікації. Окремо розглянуто питання інженерії ознак та їх комбінування з метою підвищення узагальнюючої здатності моделей. Значну увагу приділено сучасним тенденціям, таким як використання трансформерних архітектур і великих мовних моделей, які демонструють високі результати завдяки здатності враховувати контекст і семантичні зв'язки. На основі узагальнення результатів літературного аналізу було визначено ключові обмеження існуючих підходів, зокрема залежність від якості даних і обчислювальні витрати. Сформульовано перспективні напрями подальших досліджень, серед яких розробка адаптивних моделей для автоматичного формування кореляційних правил, динамічного зважування ознак у реальному часі та підвищення стійкості систем до нових типів атак.

Ключові слова: огляд літератури, шкідливі URL-адреси, машинне навчання, штучний інтелект, кібербезпека, фішинг, кореляційні правила, нейронні мережі.

1. Вступ

Активна цифровізація суспільства та перенесення значної частини сервісів і критичної інфраструктури у вебпростір зумовлюють стрімке зростання кількості кіберзагроз. Одним із найбільш поширених інструментів реалізації атак є використання шкідливих URL-адрес, які застосовуються для організації фішингових кампаній, поширення шкідливого програмного забезпечення, експлуатації вразливостей вебресурсів та координації бот-мереж. У зв'язку з експоненціальним зростанням кількості нових доменів і вебресурсів проблема оперативного виявлення шкідливих посилань набуває особливої актуальності.

Традиційні підходи до виявлення шкідливих вебресурсів здебільшого базуються на сигнатурному аналізі та використанні чорних списків. Однак ці методи мають низку обмежень, зокрема нездатність ефективно виявляти нові або модифіковані атаки, включаючи так звані атаки нульового дня. У зв'язку з цим значна частина сучасних досліджень спрямована на використання методів машинного навчання, здатних автоматично виявляти закономірності у структурі URL-адрес та поведінкових характеристиках вебресурсів [1–3].

У науковій літературі запропоновано різні підходи до вирішення цієї задачі. Значна кількість робіт зосереджена на використанні лексичних ознак URL та класичних алгоритмів машинного навчання для класифікації вебпосилань [5, 6]. Інші дослідження розглядають комбінування різних типів ознак – лексичних, мережевих та контентних – що дає змогу підвищити точність детекції [1–3, 19, 26]. Паралельно активно розвиваються підходи на основі глибокого навчання, включаючи використання трансформерних моделей та методів глибокої екстракції ознак [18, 23].

Окремий напрям досліджень пов'язаний із застосуванням асоціативної класифікації та алгоритмів видобування правил, що дозволяють виявляти інтерпретовані залежності між ознаками URL-адрес і їх належністю до певних класів кіберзагроз [16, 20]. Такі підходи забезпечують формування зрозумілих кореляційних правил, що може бути важливим для пояснюваності рішень систем кібербезпеки. Паралельно активно розвиваються методи глибокого навчання, здатні автоматично витягувати складні нелінійні ознаки з текстової структури URL та вебконтенту. Зокрема, використання трансформерних архітектур і моделей типу BERT демонструє високі результати у задачах детекції фішингових ресурсів [18]. Новітні дослідження також розглядають інтеграцію великих мовних моделей і оптимізованих глибоких архітектур для підвищення точності та адаптивності систем виявлення шкідливих URL-адрес [24, 25]. Поєднання інтерпретованих методів видобування правил із потужністю сучасних моделей глибокого навчання відкриває перспективи створення адаптивних та масштабованих систем аналізу вебзагроз.

Метою даної роботи є систематизація сучасних підходів та аналіз альтернативних методів ідентифікації шкідливих URL-адрес із застосуванням технологій штучного інтелекту. Методологія формування джерельної бази дослідження базувалася на релевантному пошуку в наукометричній базі Google Scholar за загальним запитом "Detecting malicious URL".

Процес відбору публікацій здійснювався відповідно до принципу актуальності: до аналізу залучалися праці, опубліковані протягом останнього десятиліття, з пріоритетним фокусом на найбільш сучасних розробках (2022-2026 рр.). Для забезпечення повноти огляду було також проаналізовано списки літератури обраних праць для ідентифікації досліджень, на які спирається наукова спільнота. У результаті було сформовано репрезентативну вибірку, результати аналізу та класифікації якої детально викладено в межах даної статті.

2. Аналіз традиційних підходів

Перші системи протидії фішинговим та шкідливим вебресурсам базувалися на використанні репутаційних механізмів, зокрема так званих чорних списків URL-адрес. Такі системи формують централізовані бази відомих шкідливих доменів і здійснюють перевірку шляхом прямого порівняння адреси з відповідними записами. Подібний підхід широко застосовується у практичних сервісах кібербезпеки та забезпечує швидке виявлення вже ідентифікованих загроз. Однак, як зазначено у сучасних оглядах літератури [1–3], ефективність таких систем суттєво знижується у випадку нових або модифікованих URL-адрес, оскільки між моментом появи шкідливого ресурсу та його внесенням до бази даних існує часовий лаг.

Паралельно з репутаційними системами розвивалися методи сигнатурного та евристичного аналізу. Вони базуються на використанні попередньо визначених правил, які дозволяють виявляти характерні патерни у структурі URL або у вмісті

вебсторінок. Наприклад, такі підходи можуть враховувати довжину URL, кількість спеціальних символів, використання IP-адрес замість доменних імен або наявність підозрілих ключових слів. Подібні методи дозволяють виявляти окремі структурні аномалії, однак їх ефективність обмежена через низьку адаптивність до нових технік маскування, таких як обфускація URL, використання сервісів скорочення посилань або алгоритмічна генерація доменів.

Обмеження традиційних методів стимулювали розвиток підходів, що базуються на аналізі ознак та використанні алгоритмів машинного навчання. Дослідження показують, що навіть базовий лексичний аналіз структури URL може забезпечити ефективне розпізнавання шкідливих посилань без використання зовнішніх репутаційних баз [1–3, 5, 6]. Саме використання таких ознак стало основою для створення перших моделей автоматичної класифікації вебпосилань, що дозволило перейти від реактивних систем виявлення до проактивних методів детекції, здатних узагальнювати характеристики шкідливих ресурсів на основі даних [4].

3. Класифікація та аналіз ознак шкідливих URL

Ефективність моделей автоматичного виявлення шкідливих вебресурсів значною мірою залежить від якості та репрезентативності використовуваних ознак. У сучасних дослідженнях процес формування ознак розглядається як ключовий етап побудови систем виявлення кіберзагроз, оскільки саме він забезпечує перетворення сирих даних у набір параметрів, придатних для машинного аналізу. Опираючись на сучасні наукові роботи можна виділити чотири основні групи ознак: лексичні (URL-based), мережеві або хостові (host-based), контентні (content-based) та зовнішні (external features) [1, 26].

Зазначені групи ознак відрізняються як за способом отримання, так і за обчислювальною складністю. Наприклад, URL-based характеристики можуть бути отримані миттєво без звернення до зовнішніх ресурсів, тоді як host-based або content-based параметри потребують додаткових мережевих запитів або завантаження сторінки. Тому у практичних системах детекції часто використовується комбінування декількох типів ознак, що дає змогу досягти балансу між швидкістю класифікації та її точністю.

3.1. URL-based ознаки

Ознаки, що формуються безпосередньо зі структури URL-адреси, є найбільш поширеним типом характеристик у системах виявлення шкідливих посилань. Їх основною перевагою є можливість виконання аналізу без завантаження вебсторінки, що дає змогу проводити класифікацію практично миттєво під час переходу за посиланням [5, 6].

URL-based характеристики можна умовно поділити на кілька підкатегорій. Перша з них – це лексичні та структурні ознаки. До цієї групи належать такі параметри, як загальна довжина URL-адреси, довжина доменного імені, кількість піддоменів, кількість сегментів шляху, а також використання спеціальних символів, наприклад “@”, “-”, “_” або “=”. У дослідженнях також враховується використання IP-адрес замість доменних імен, наявність номерів портів або нетипових доменів верхнього рівня.

Друга підкатегорія представлена статистичними характеристиками. Одним із

найбільш інформативних параметрів є ентропія символів у рядку URL. Висока ентропія може свідчити про використання алгоритмів генерації доменних імен (Domain Generation Algorithms, DGA), які часто застосовуються у ботнетах та шкідливих кампаніях. На відміну від легітимних доменів, які зазвичай мають зрозумілу для людини структуру, автоматично згенеровані адреси складаються з випадкових послідовностей символів.

Окрему роль відіграє аналіз текстових токенів. У шкідливих URL часто використовуються слова, пов'язані з автентифікацією або фінансовими операціями, наприклад *login*, *secure*, *update*, *verification* або *banking*. Для перетворення таких текстових характеристик у числові представлення застосовуються методи обробки природної мови, зокрема n-грамні моделі та підхід TF-IDF (Term Frequency-Inverse Document Frequency). Це дає змогу формувати векторні представлення URL-адрес, які можуть бути використані для подальшого навчання алгоритмів машинного навчання.

3.2. Host-based ознаки

Host-based ознаки описують технічні характеристики інфраструктури, на якій розміщений вебресурс. На відміну від лексичних характеристик, вони формуються на основі аналізу мережевих параметрів сервера та конфігурації домену.

До основних параметрів цієї категорії належать IP-адреса сервера, географічне розташування хоста, наявність захищеного з'єднання HTTPS, а також технічні характеристики SSL/TLS-сертифіката. Хоча сучасні фішингові сайти можуть використовувати легітимні сертифікати, їх відсутність або неправильна конфігурація часто виступає індикатором потенційної небезпеки.

Важливим фактором також є використання нестандартних портів у структурі URL, що може свідчити про спробу приховати реальну інфраструктуру ресурсу. Крім того, у ряді досліджень враховується геолокація сервера, оскільки статистичні дані демонструють підвищену концентрацію шкідливих ресурсів у певних регіонах світу. Комбінування таких параметрів із лексичними характеристиками дає змогу значно підвищити точність моделей класифікації [19, 26].

3.3. Content-based ознаки

Content-based характеристики формуються на основі аналізу HTML-структури та програмного коду вебсторінки, яка відкривається за відповідною URL-адресою. На відміну від попередніх груп ознак, цей підхід потребує повного завантаження сторінки, що збільшує час обробки запиту та може становити потенційний ризик для системи аналізу. Проте саме ці ознаки часто забезпечують найвищу точність детекції [19].

Контентні характеристики можна поділити на кілька функціональних рівнів. Перший рівень – аналіз HTML-структури сторінки. Моделі досліджують наявність підозрілих елементів, таких як приховані поля форм, iFrame-елементи, велика кількість зовнішніх посилань або форми введення конфіденційних даних. Фішингові сторінки часто використовують такі елементи для збору облікових даних користувачів.

Другий рівень включає аналіз JavaScript-коду. Дослідження показують, що активний контент може містити шкідливі скрипти, які використовують функції

на кшталт *eval()*, *document.write()* або *unescape()* для приховування небезпечних операцій. Також аналізуються скрипти, що реагують на специфічні події, наприклад спробу закрити вкладку або наведення курсора.

У деяких сучасних підходах також використовується аналіз візуальної та семантичної схожості сторінок. Алгоритми комп'ютерного зору дозволяють визначати схожість інтерфейсу сторінки із відомими вебсервісами, що є ефективним способом виявлення сайтів-двійників, які імітують легітимні ресурси для викрадення облікових даних користувачів.

3.4. Зовнішні ознаки

Зовнішні характеристики формуються на основі інформації, отриманої з незалежних джерел, таких як системи кіберрозвідки або спеціалізовані аналітичні сервіси. Використання таких даних дає змогу оцінити репутацію вебресурсу на основі історичної інформації та глобального аналізу інтернет-інфраструктури [1].

До найбільш поширених зовнішніх ознак належать реєстраційні дані домену, отримані з сервісів WHOIS. У рамках аналізу враховується вік домену, дата останнього оновлення реєстраційної інформації, рівень приватності власника та кількість доменів, пов'язаних з однією IP-адресою. Новостворені домени з високою активністю часто виступають індикатором фішингових кампаній.

Іншим джерелом інформації є репутаційні сервіси та антивірусні бази даних. Наприклад, агрегатори загроз на кшталт VirusTotal або спеціалізовані репозиторії, такі як PhishTank чи OpenPhish, дозволяють отримувати статистику виявлення конкретного ресурсу різними системами безпеки. У деяких дослідженнях ці дані використовуються як додаткові ознаки для навчання моделей, що дає змогу поєднати переваги традиційних репутаційних систем із методами машинного навчання.

Окрім цього, аналізується рівень присутності вебресурсу у глобальній мережі. Для цього використовуються показники пошукової індексації та популярності домену, зокрема рейтинги Alexa або Tranco. Легітимні ресурси зазвичай мають стабільний трафік та історію існування, тоді як шкідливі сайти часто відсутні у пошукових індексах або мають аномально низький рівень популярності.

3.5. Підсумки аналізу ознак

Підсумовуючи аналіз таксономії ознак, слід зазначити, що кожна з розглянутих груп має власну специфіку та сферу ефективного застосування. URL-based ознаки забезпечують найвищу швидкість детекції, що є критичним для систем захисту в реальному часі, тоді як контентні та хостові параметри дозволяють ідентифікувати складні та багатовекторні атаки за рахунок поглибленого аналізу інфраструктури та коду сторінки.

Проаналізований огляд літератури [1–3] демонструє чітку тенденцію до переходу від використання поодиноких параметрів до побудови багатовимірних векторів ознак. Проте збільшення кількості вхідних даних призводить до зростання обчислювальної складності та ризику перенавчання моделей. У зв'язку з цим актуальним завданням є не лише накопичення максимальної кількості ознак, а й дослідження їхньої взаємозалежності.

Саме виявлення прихованих кореляцій між лексичними аномаліями рядка URL та технічними характеристиками хоста дає змогу створювати адаптивні моделі,

здатні динамічно змінювати ваги параметрів залежно від контексту атаки. Таким чином, аналіз таксономії ознак створює необхідне теоретичне підґрунтя для розробки методів автоматичного формування кореляційних правил, що є предметом подальшого розгляду в межах даного дослідження.

4. Огляд методів штучного інтелекту в детекції шкідливих URL

Зі зростанням кількості фішингових атак, шкідливих вебресурсів та ботнет-інфраструктур традиційні методи виявлення, засновані виключно на чорних списках або сигнатурному аналізі, виявилися недостатньо ефективними. Такі підходи здатні ідентифікувати лише відомі загрози та не дозволяють оперативно реагувати на появу нових шкідливих доменів. У зв'язку з цим у сучасних дослідженнях дедалі ширше застосовуються методи штучного інтелекту та машинного навчання, які дозволяють автоматично виявляти закономірності у структурі URL та характеристиках вебресурсів.

Алгоритми штучного інтелекту використовуються для аналізу ознак, описаних у попередньому розділі, і дозволяють будувати моделі класифікації, здатні розрізняти легітимні та шкідливі посилання. З огляду літератури можна виділити три основні групи: класичні алгоритми машинного навчання, методи глибокого навчання та сучасні моделі трансформерної архітектури.

4.1. Класичні методи машинного навчання

Класичні алгоритми машинного навчання залишаються одним із найбільш поширених підходів до виявлення шкідливих URL-адрес. У більшості досліджень застосовується підхід контрольованого навчання (supervised learning), у рамках якої модель навчається на попередньо розміченому наборі даних. Кожному об'єкту, представленому набором ознак, відповідає мітка класу (наприклад, легітимний або шкідливий ресурс). У процесі навчання алгоритм формує модель, яка дає змогу визначати залежності між значеннями ознак та відповідними класами. Після завершення навчання така модель може використовуватися для класифікації нових, раніше невідомих URL-адрес.

Домінуюча роль традиційних методів машинного навчання зумовлена специфікою вхідних даних: більшість характеристик URL мають структурований вигляд і легко трансформуються у числові або категоріальні змінні. Аналіз поточної наукової бази [1–3] свідчить, що найбільш поширеними є такі алгоритми, як випадковий ліс (Random Forest), метод опорних векторів (SVM), логістична регресія, наївний байєсівський класифікатор та метод k-найближчих сусідів (KNN). Особливу увагу в сучасних дослідженнях приділено ансамблевим підходам на основі градієнтного бустингу (XGBoost, LightGBM), які забезпечують високу точність детекції при роботі з великими вибірками.

Основними перевагами таких підходів є відносно невелика обчислювальна складність, висока швидкість навчання та можливість інтерпретації отриманих результатів. Водночас ефективність класичних моделей значною мірою залежить від якості сформованих ознак, тому процес інженерії характеристик відіграє ключову роль у побудові точних систем детекції.

4.2. Методи глибокого навчання

Методи глибокого навчання дозволяють автоматично витягувати складні закономірності з даних без необхідності ручної інженерії ознак. У задачах виявлення шкідливих URL такі моделі здатні безпосередньо аналізувати текстову структуру посилань або вміст вебсторінок [23].

Найбільш поширеними архітектурами, що розглядаються в сучасних дослідженнях, є згорткові нейронні мережі (CNN) та рекурентні структури, зокрема мережі з довгою короткостроковою пам'яттю (LSTM). Згорткові нейронні мережі демонструють високу ефективність у виявленні локальних структурних аномалій та специфічних n-грам у назвах доменів. У свою чергу, архітектури LSTM краще адаптовані до моделювання часових та послідовних залежностей, що є критично важливим для ідентифікації алгоритмів динамічної генерації доменів (DGA) та виявлення прихованих закономірностей у довгих ланцюжках редиректів [24].

Незважаючи на високу предиктивну здатність, моделі глибокого навчання характеризуються значною обчислювальною складністю та потребою у великих масивах навчальних даних. Зазначені обмеження обумовлюють необхідність розробки методів оптимізації та використання пріоритетних коефіцієнтів для ознак [21]. Такий підхід дає змогу підвищити стабільність та швидкість роботи нейромережевих моделей у реальних умовах експлуатації, зберігаючи високу точність класифікації в умовах постійної зміни ландшафту кіберзагроз [3].

4.3. Трансформери та мовні моделі

Найбільш актуальним напрямом досліджень у 2024-2025 роках стало впровадження архітектур на основі механізмів уваги (Attention mechanisms) та великих мовних моделей (LLM) для аналізу кіберзагроз. На відміну від рекурентних мереж, трансформерні моделі дозволяють паралелізувати обробку послідовностей та ефективно вловлювати глобальні контекстуальні залежності в структурі URL-адрес [24, 25].

Ключовим підходом у цьому сегменті є використання моделей сімейства BERT (Bidirectional Encoder Representations from Transformers). Зокрема, застосування архітектур ELECTRA та RoBERTa дає змогу здійснювати глибоку екстракцію ознак на рівні токенів, що значно підвищує точність ідентифікації складної обфускації та прихованих редиректів [18, 24]. Такі моделі демонструють здатність до розуміння семантичного наповнення URL, що робить їх стійкими до незначних модифікацій зловмисних посилань, які зазвичай обходять класичні ML-алгоритми. Дослідження 2024-2025 років підтверджують, що комбінування глибокого навчання з механізмами уваги (Attention mechanisms) дає змогу досягти показників точності понад 99%, одночасно знижуючи кількість хибнопозитивних спрацювань [3, 25].

Окрему увагу в останніх працях [25] приділено розробці легковагових (lightweight) рішень на базі LLM. Вони поєднують потужність глибокого навчання з оптимізованими методами обробки природної мови (NLP), що дає змогу використовувати їх у системах реального часу з мінімальною затримкою. Інтеграція таких моделей у гібридні фреймворки, що включають механізми оцінки пріоритетності ознак [21], дає змогу адаптувати систему до динамічної зміни ландшафту загроз, забезпечуючи високу генералізаційну здатність моделей на нових, раніше невідомих доменах [3].

4.4. Порівняльний аналіз ефективності алгоритмів

Аналіз результатів експериментальних досліджень, представлених у розглянутій літературі, дає змогу провести порівняльну оцінку ефективності різних архітектур машинного та глибокого навчання. Ключовими метриками для оцінки якості моделей традиційно виступають точність (Accuracy), повнота (Recall) та F1-міра.

Узагальнені результати щодо поширеності використання алгоритмів та досягнутих показників точності наведено в табл. 1. Відсоткові значення, подані в дужках, відображають приблизну частку досліджень у проаналізованій вибірці літератури, в яких відповідний алгоритм або клас моделей застосовувався як один з основних методів класифікації.

Таблиця 1

Порівняння підходів штучного інтелекту для виявлення шкідливих URL

Категорія	**Найпопулярніші алгоритми**	**Accuracy**	**Ключова перевага**
Класичний ML	Random Forest (60%)	до 99.70%	Баланс швидкості та стійкості
Ансамблі	XGBoost (30%)	до 99.98%	Комплексний підхід
Deep Learning	CNN / LSTM (22%)	до 99.98%	Автоматичне виявлення ознак
Трансформери	BERT / ELECTRA / AraBERT	до 99.7%	Аналіз контексту та структури

Як свідчать наведені дані, класичні моделі машинного навчання (Random Forest, Logistic Regression, SVM та інші) характеризуються стабільними й високими результатами. Набільш поширеним є алгоритм Random Forest, який згадується у 60% проаналізованих робіт. Проте в більшості випадків класичні алгоритми ML поступаються ансамблевим методам за показниками точності. Алгоритми градієнтного бустингу (XGBoost, LightGBM та інші) демонструють високий рівень ефективності при роботі зі структурованими ознаками та часто виступають базовими моделями у прикладних дослідженнях.

Найвищі значення точності (до 99.98%), згадані у працях, досягаються при використанні глибоких нейронних мереж і трансформерних архітектур [25]. Це пояснюється їх здатністю автоматично виділяти складні нелінійні залежності та враховувати контекстні характеристики URL-адрес. Водночас підвищення точності супроводжується зростанням обчислювальних витрат, що зумовлює актуальність досліджень, спрямованих на оптимізацію моделей, зокрема через методи відбору та зважування ознак [21].

4.5. Статистичний аналіз структури наукових досліджень

Для аналізу динаміки розвитку галузі та систематизації сучасного стану досліджень у сфері виявлення шкідливих URL-адрес було проведено статистичний аналіз обраних наукових публікацій за період 2016-2025 рр. Усі роботи у вибірці (26 джерел) були класифіковані за домінуючим підходом: оглядові дослідження,

класичне машинне навчання (ML), глибоке навчання (DL), гібридні моделі та альтернативні методи (асоціативні правила, оптимізація ознак тощо).

Візуалізація структури проаналізованих праць (рис. 1) відображає динаміку розвитку підходів у межах сформованої вибірки. Попри те, що наведена діаграма ілюструє лише приблизний розподіл типів досліджень за останні роки, вона дає змогу простежити тенденцію до ускладнення використовуваних архітектур. Аналіз підтверджує, що класичні методи машинного навчання (ML) зберігають стабільну актуальність. Водночас у публікаціях за останні 5 років спостерігається активне впровадження моделей глибокого навчання (DL) та гібридних підходів, що обумовлено необхідністю автоматизації екстракції ознак та підвищення адаптивності систем до нових типів загроз.

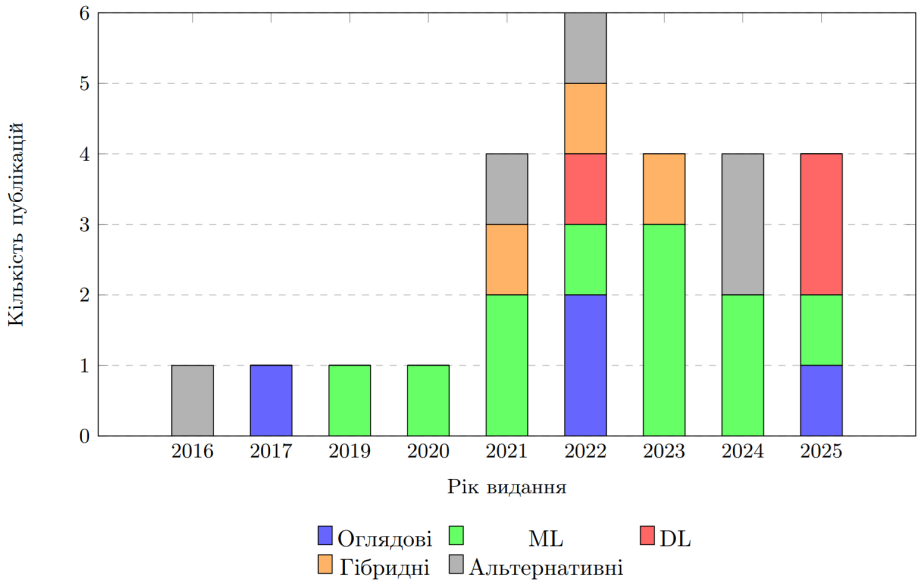

Рис. 1. Динаміка зміни наукових підходів до виявлення шкідливих URL-адрес за роками (на основі проаналізованої вибірки)

Узагальнений аналіз структури вибірки підтверджує, що найбільшу частку (близько 42%) все ще становлять експериментальні роботи на базі класичного машинного навчання [4–9, 12–15, 19]. Висока репрезентативність даного сегмента свідчить про те, що імплементація традиційних ML-моделей залишається пріоритетним підходом завдяки їхній швидкодії та відносно низьким обчислювальним витратам.

Сегмент альтернативних підходів, що охоплює методи асоціативних правил, алгоритми оптимізації ознак та інші, складає 19% від загальної кількості [11, 16, 20–22]. Оглядові праці становлять 15% [1–3, 23] і слугують базою для систематизації наявних рішень.

Гібридні підходи [10, 17, 26] та дослідження у сфері глибокого навчання [18, 24, 25] представлені у рівних пропорціях (по 12% кожна). Активізація цього напряму у 2024-2025 рр. ілюструє поступове впровадження нейромережевих архітектур для автоматизації екстракції ознак та підвищення адаптивності систем до складних, раніше невідомих типів кіберзагроз.

5. Проблеми та перспективи розвитку систем детекції

Проведений аналіз сучасних методів ідентифікації шкідливих URL-адрес дає змогу виділити низку критичних проблем, які залишаються невирішеними, незважаючи на високу точність існуючих нейромережевих архітектур.

Першою суттєвою проблемою є динамічна зміна ландшафту кіберзагроз. Неважливо наскільки модель є ефективною на певному тренувальному наборі даних, завжди можна знайти вразливість, яку вона не зможе ефективно ідентифікувати. Зловмисники постійно вдосконалюють методи обфускації, використовують сервіси коротких посилань та алгоритми динамічної генерації доменів (DGA), що вимагає постійного вдосконалення систем захисту. Як наслідок, статичні моделі машинного навчання швидко втрачають свою актуальність і потребують регулярного перенавчання на нових вибірках [3, 10].

Другим викликом є висока обчислювальна складність сучасних моделей на базі глибокого навчання та трансформерів. Хоча вони демонструють найвищу точність й адаптивну здатність, їх використання в реальному часі для аналізу всього мережевого трафіку обмежене апаратними ресурсами. Це зумовлює потребу в оптимізації процесів екстракції ознак та перехід до методів інтелектуального зважування параметрів [25].

У межах даного дослідження найбільш перспективним вектором розвитку вбачається перехід від використання фіксованих векторів ознак до розробки адаптивних систем на основі кореляційних правил. Проаналізована література [19, 21] підтверджує, що не всі ознаки URL-адреси є однаково інформативними в різні проміжки часу або для різних типів атак. Наприклад, лексичні аномалії можуть бути визначальними для фішингу, тоді як хостові ознаки – для детекції бот-мереж.

Перспективи подальшої роботи полягають у розробці моделей для автоматичного знаходження кореляційних правил між різними групами ознак. Це дозволить створити адаптивну систему, яка здатна:

- визначати реальні ваги ознак у режимі реального часу;
- автоматично відсіювати малоінформативні параметри для конкретного контексту, знижуючи обчислювальне навантаження;
- підвищувати стабільність детекції в умовах еволюції методів атаки за рахунок виявлення прихованих зв'язків між структурними елементами URL та мережевими характеристиками хоста [21].

Результатом впровадження таких правил має стати створення гнучкої моделі, яка не просто класифікує посилання, а адаптує свою внутрішню структуру до актуальних тенденцій у вебпросторі, забезпечуючи баланс між швидкістю роботи та точністю виявлення загроз.

6. Висновки

У межах проведеного дослідження здійснено комплексний аналіз сучасної наукової бази, присвяченої методам ідентифікації зловмисних вебресурсів. Огляд продемонстрував, що традиційні підходи, які базуються на репутаційних списках та сигнатурному аналізі, вичерпали свій потенціал через неспроможність протидіяти динамічним загрозам та атакам типу "нульового дня". Це зумовило домінуючу роль технологій штучного інтелекту як основного інструменту забезпечення кібербезпеки в сучасному мережевому просторі.

Аналіз таксономії ознак дозволив встановити, що найвищу предиктивну здатність демонструють гібридні моделі, які інтегрують лексичні характеристики URL-адрес із хостовими параметрами інфраструктури та контентним аналізом цільових сторінок. При цьому порівняльна оцінка алгоритмів підтвердила, що ансамблеві методи машинного навчання, зокрема Random Forest та XGBoost, забезпечують точність детекції на рівні 95.0% – 99.98%, що є оптимальним балансом між швидкістю та якістю класифікації. Глибоке навчання та трансформерні архітектури, такі як BERT та ELECTRA, попри вищу обчислювальну складність, демонструють показники понад 99.4% та є найбільш ефективними в умовах складної обфускації даних.

Ключовим результатом аналізу став висновок про необхідність переходу від статичних моделей до адаптивних систем, здатних враховувати пріоритетність ознак у реальному часі. Виявлена проблема концептуального зсуву в даних та еволюція методів генерації зловмисних доменів обґрунтовують доцільність подальших досліджень у напрямі автоматизації знаходження кореляційних правил. Розробка методів динамічного зважування параметрів дозволить створити гнучку архітектуру, яка мінімізує обчислювальні витрати шляхом ігнорування малоінформативних ознак та підвищує загальну стійкість системи до нових типів кібератак, що є перспективним вектором для розробки сучасних засобів захисту інформаційного простору.

Список використаної літератури

1. Sahoo D. Malicious URL Detection using Machine Learning: A Survey / D. Sahoo, C. Liu, S.C.H. Hoi // arXiv preprint. – 2017. – arXiv:1701.07179. – DOI: http://dx.doi.org/10.48550/arXiv.1701.07179.
2. Aljabri M. Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions / M. Aljabri, H. Altamimi, S. Albelali, M. Al-Harbi, H. Alhuraib, N. Alotaibi, A. Alahmadi, F. Alhaidari, R. Mohammad, K. Salah // IEEE Access. – 2022. – DOI: http://dx.doi.org/10.1109/ACCESS.2022.3222307.
3. Tian Y. From Past to Present: A Survey of Malicious URL Detection Techniques, Datasets and Code Repositories / Y. Tian, Y. Yu, J. Sun, Y. Wang // arXiv preprint. – 2025. – arXiv:2504.16449. – DOI: http://dx.doi.org/10.48550/arXiv.2504.16449.
4. Cherradi M. Malicious URL detection using machine learning techniques / M. Cherradi, H. El Mahajer // International Journal of Digital Signals and Smart Systems (IJDIIC). – 2025. – Vol. 4, № 2. – P. 41–52. – DOI: http://dx.doi.org/10.59461/ijdiic.v4i2.187.
5. Dhotre A. Malicious URLs Detection using Lexical Features based on Machine Learning / A. Dhotre // IJSRD. – 2023. – Vol. 11, Iss. 8.
6. Joshi A. Using Lexical Features for Malicious URL Detection – A Machine Learning Approach / A. Joshi, L. Lloyd, P. Westin, S. Seethapathy // arXiv preprint. – 2019. – arXiv:1910.06277. – DOI: http://dx.doi.org/10.48550/arXiv.1910.06277.
7. Rashid J. Phishing Detection Using Machine Learning Technique / J. Rashid, T. Mahmood, M. Nisar, T. Nazir // Proc. 2nd Int. Conf. on Smart Tools and Technologies (SMART-TECH). – 2020. – DOI: http://dx.doi.org/10.1109/SMART-TECH49988.2020.00026.
8. Alazaidah R. Website Phishing Detection Using Machine Learning Techniques / R. Alazaidah, A. Al-Shaikh, M. Almousa, H. Khafajeh, G. Samara, M. Alzyoud, N. Al-shanableh, S. Almatarneh // Journal of Statistics Applications & Probability. – 2024. – Vol. 13, № 1. – P. 119–129. – DOI: http://dx.doi.org/10.18576/jsap/130108.
9. Shayan A. Classification of Malicious URLs Using Machine Learning / A. Shayan, M.A. Nasser, A.B. Samma, S. Abad, H. Gholamy, M. Aslani // Sensors. – 2023. – Vol. 23, № 18. – Art. 7760. – DOI: http://dx.doi.org/10.3390/s23187760.

10. Alsaedi M. Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning / M. Alsaedi, F.A. Ghaleb, F. Saeed, J. Ahmad, M. Alasli // Sensors. – 2022. – Vol. 22, № 9. – Art. 3373. – DOI: http://dx.doi.org/10.3390/s22093373.
11. Yuan J. A Novel Approach for Malicious URL Detection Based on the Joint Model / J. Yuan, Y. Liu, L. Yu // Security and Communication Networks. – 2021. – Vol. 2021. – Art. 4917016. – 12 p. – DOI: http://dx.doi.org/10.1155/2021/4917016.
12. Coste C.I. Malicious Web Links Detection – A Comparative Analysis of Machine Learning Algorithms / C.I. Coste // Studia Universitatis Babes-Bolyai Informatica. – 2023. – Vol. 68, № 1. – P. 21–36. – DOI: http://dx.doi.org/10.24193/subbi.2023.1.02.
13. Wejinya G. Machine Learning for Malicious URL Detection / G. Wejinya, S. Bhatia // ICT Systems and Sustainability. Advances in Intelligent Systems and Computing. / M. Tuba, S. Akashe, A. Joshi (eds). – Singapore: Springer. – 2021. – Vol. 1270. – P. 463–472. – DOI: http://dx.doi.org/10.1007/978-981-15-8289-9_45.
14. Mummadi B.T. Detection of Phishing Websites Using Supervised Learning / B.T. Mummadi, N. Puligundla // International Journal of Intelligent Systems and Applications in Engineering (IJISAE). – 2022. – Vol. 10, № 3. – P. 171–186.
15. Bouijij H. Machine Learning Algorithms Evaluation for Phishing URLs Classification / H. Bouijij, A. Berqia // Proc. 4th Int. Symp. on Advanced Electrical and Communication Technologies (ISAECT). – 2021. – P. 1–5. – DOI: http://dx.doi.org/10.1109/ISAECT53699.2021.9668489.
16. Kumi S. Malicious URL Detection Based on Associative Classification / S. Kumi, C. Lim, S.-G. Lee // Entropy. – 2021. – Vol. 23, № 2. – Art. 182. – DOI: http://dx.doi.org/10.3390/e23020182.
17. Kinger S. Malicious URL Detection Using Machine Learning / S. Kinger, P. Nirmal, A. Shrivastav, A. Sharma, S. Saindane // Proc. 6th Int. Conf. on Contemporary Computing and Informatics (IC3I). – 2023. – P. 1062–1068. – DOI: http://dx.doi.org/10.1109/IC3I59117.2023.10397872.
18. Elsadig M. Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on BERT Features Extraction / M. Elsadig, A.O. Ibrahim, S. Basheer, M.A. Alohali, S. Alshunaifi, H. Alqahtani, N. Alharbi, W. Nagmeldin // Electronics. – 2022. – Vol. 11, № 22. – Art. 3647. – DOI: http://dx.doi.org/10.3390/electronics11223647.
19. Hamadouche S. Combining Lexical, Host, and Content-based features for Phishing Websites detection using Machine Learning Models / S. Hamadouche, O. Boudraa, M. Gasmi // EAI Endorsed Transactions on Scalable Information Systems. – 2024. – Vol. 11, № 6. – DOI: http://dx.doi.org/10.4108/eetsis.4421.
20. Jeeva S.C. Intelligent phishing url detection using association rule mining / S.C. Jeeva, E.B. Rajsingh // Hum. Cent. Comput. Inf. Sci. – 2016. – Vol. 6. – Art. 10. – DOI: http://dx.doi.org/10.1186/s13673-016-0064-3.
21. Rafsanjani A.S. Enhancing Malicious URL Detection: A Novel Framework Leveraging Priority Coefficient and Feature Evaluation / A.S. Rafsanjani, N. Binti Kamaruddin, M. Behjati, S. Aslam, A. Sarfaraz, A. Amphawan // IEEE Access. – 2024. – Vol. 12. – P. 85001–85026. – DOI: http://dx.doi.org/10.1109/ACCESS.2024.3412331.
22. Reyes-Dorta N. Detection of malicious URLs using machine learning / N. Reyes-Dorta, P. Caballero-Gil, C. Rosa-Remedios // Wireless Netw. – 2024. – Vol. 30. – P. 7543–7560. – DOI: http://dx.doi.org/10.1007/s11276-024-03700-w.
23. Do N. Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions / N. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, H. Fujita // IEEE Access. – 2022. – Vol. 10. – P. 36543–36564. – http://dx.doi.org/10.1109/ACCESS.2022.3151903.
24. Turk F. Malicious URL Detection with Advanced Machine Learning and Optimization-Supported Deep Learning Models / F. Turk, M. Kilicaslan // Appl. Sci. – 2025. – Vol. 15, № 18. – Art. 10090. – DOI: http://dx.doi.org/10.3390/app151810090.

25. Kibriya H. Lightweight malicious URL detection using deep learning and large language models / H. Kibriya, R. Amin, S.S. Alshamrani et al. // Sci Rep. – 2025. – Vol. 15. – Art. 43044. – DOI: http://dx.doi.org/10.1038/s41598-025-26653-2.
26. Aljabri M. An Assessment of Lexical, Network, and Content-Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models / M. Aljabri, F. Alhaidari, R. Mohammad, S. Mirza, D. Alhamed, H. Altamimi, S. Chrouf // Computational Intelligence and Neuroscience. – 2022. – DOI: http://dx.doi.org/10.1155/2022/3241216.

Стаття: надійшла до редколегії 02.02.2026
доопрацьована 24.02.2026
прийнята до друку 04.03.2026

MACHINE AND DEEP LEARNING METHODS FOR MALICIOUS URL DETECTION: CURRENT STATE AND FUTURE PROSPECTS

V. Lesyk

Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine,
e-mail: volodymyr.lesyk@lnu.edu.ua

The article presents a systematic review of contemporary scientific publications devoted to the problem of detecting malicious URLs using artificial intelligence technologies. The evolution of approaches is analyzed, ranging from traditional blacklist-based methods and heuristic rules, which are characterized by limited adaptability, to highly efficient machine learning and deep learning models capable of identifying complex and hidden patterns in data. Particular attention is paid to the comparative analysis of classical algorithms (logistic regression, SVM, ensemble methods) and modern neural network architectures. The paper also provides a detailed classification of features used for model training, including lexical, host-based, content-based, and external features, and analyzes their informativeness and impact on classification performance. Special consideration is given to feature engineering techniques and their combination in order to improve the generalization ability of models. Significant attention is devoted to current trends, such as the use of transformer architectures and large language models, which demonstrate high performance due to their ability to capture context and semantic relationships. Based on the generalization of the literature review results, the key limitations of existing approaches are identified, in particular the dependence on data quality and computational costs. Promising directions for future research are formulated, including the development of adaptive models for automatic generation of correlation rules, dynamic feature weighting in real time, and improving system robustness against new types of attacks.

Key words: literature review, malicious URLs, machine learning, artificial intelligence, cybersecurity, phishing, correlation rules, neural networks.