ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «СИСТЕМА ЗАХИСТУ ДАНИХ ВІД ФІШИНГОВИХ АТАК НА БАЗІ ШТУЧНОГО ІНТЕЛЕКТУ»

на здобуття освітнього ступеня бакалавра
зі спеціальності 123 Комп'ютерної інженерії
(код, найменування спеціальності)
освітньо-професійної програми Комп'ютерна інженерія
(назва)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ *Ілля БАЗІЛЕВИЧ*
(підпис) *Ім'я, ПРІЗВИЩЕ здобувача*

Виконав: здобувач вищої освіти гр. КІД-41
Ілля БАЗІЛЕВИЧ
Ім'я, ПРІЗВИЩЕ

Керівник: Ірина КУЗЬМІЧ
науковий ступінь, вчене звання Ім'я, ПРІЗВИЩЕ

Рецензент: _____
науковий ступінь, вчене звання Ім'я, ПРІЗВИЩЕ

Київ 2026

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
Навчально-науковий інститут Інформаційних технологій

Кафедра 123 Комп'ютерної інженерії
Ступінь вищої освіти Бакалавр
Спеціальність 123 Комп'ютерної інженерії
Освітньо-професійна програма Комп'ютерна інженерія

ЗАТВЕРДЖУЮ
Завідувач кафедрою Наталія ЛАЩЕВСЬКА
Ім'я, ПРІЗВИЩЕ

«_____» _____ 2026р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Базілевич Ілля Юрійович
(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: Система захисту даних від фішингових атак на базі штучного інтелекту
керівник кваліфікаційної роботи Ірина КУЗЬМІЧ,
(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)
затверджені наказом Державного університету інформаційно-комунікаційних технологій від «20» 02 2026р. № 51
2. Строк подання кваліфікаційної роботи «08» 05 2026р.
3. Вихідні дані до кваліфікаційної роботи: Науково-технічна література, методи системного аналізу, штучний інтелект, математичний апарат, емпіричні методи
Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. Комплексний аналіз, огляд та практика впровадження технологій штучного інтелекту для забезпечення безпеки даних від фішингових загроз
2. Комплексне дослідження та порівняльна оцінка ефективності алгоритмів машинного і глибинного навчання у задачах ідентифікації фішингових загроз
3. Практична реалізація та проектування прикладної системи на базі штучного інтелекту для ефективної протидії фішинговим загрозам та забезпечення надійного захисту конфіденційних даних
4. Перелік ілюстративного матеріалу: *презентація*
5. Дата видачі завдання «09» 02 2026р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Збір даних	25.02-24.03.2026	Виконано
2	Обробка матеріалу	23.03-05.04.2026	Виконано
3	Розробка теоретичної частини	06.04-10.04.2026	Виконано
4	Розробка практичної частини	11.04-20.04.2026	Виконано
5	Висновки за результатами аналізу	21.04-06.05.2026	Виконано
6	Оформлення пояснювальної записки	07.05-08.05.2026	Виконано
7	Розробка презентації	09.05-11.05.2026	Виконано
8	Передзахист	27.05-29.05.2026	Виконано
9	Здача в деканат	02.06-05.06.2026	Виконано

Здобувач вищої освіти _____ Ілля БАЗІЛЕВИЧ
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Керівник кваліфікаційної роботи _____ Ірина КУЗЬМІЧ
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня бакалавра: 81 стор., 9 рис., 3 табл., 24 джерела.

Мета дослідження – розробка та програмна реалізація системи виявлення фішингових атак на базі методів штучного інтелекту та обробки природної мови.

Об'єкт дослідження – процес ідентифікації та блокування фішингових загроз в інформаційних та комунікаційних мережах.

Предмет дослідження – алгоритми машинного навчання та технології семантичного аналізу тексту для класифікації електронних повідомлень.

Короткий зміст роботи: Стрімкий розвиток цифрових комунікацій супроводжується пропорційним зростанням кіберзагроз, серед яких фішинг залишається найпоширенішим вектором атак. Використання традиційних статичних фільтрів та сигнатурних баз втрачає свою ефективність через появу атак «нульового дня» та використання зловмисниками генеративного штучного інтелекту для створення персоналізованих повідомлень. Це зумовлює гостру необхідність у впровадженні інтелектуальних систем захисту, здатних динамічно аналізувати семантику тексту та виявляти приховані патерни маніпуляцій.

На основі проведеного дослідження було проаналізовано сучасні архітектури машинного та глибинного навчання. Для практичної реалізації було сформовано репрезентативний збалансований датасет обсягом понад 22 000 повідомлень. Шляхом векторизації алгоритмом TF-IDF та порівняльного тестування чотирьох моделей було доведено, що модель логістичної регресії забезпечує оптимальний баланс між високою точністю, швидкістю роботи та можливістю порогового налаштування. На базі обраної моделі розроблено практичний CLI-інструмент, який демонструє високу ефективність у режимі реального часу та готовий до інтеграції в сучасні SaaS-системи або поштові шлюзи.

КЛЮЧОВІ СЛОВА: ФІШИНГ, ШТУЧНИЙ ІНТЕЛЕКТ, МАШИННЕ НАВЧАННЯ, ОБРОБКА ПРИРОДНОЇ МОВИ, ЛОГІСТИЧНА РЕГРЕСІЯ, TF-IDF, КІБЕРБЕЗПЕКА.

ABSTRACT

Textual part of the qualification work for the bachelor's degree: 81 pages, 9 figures, 3 tables, 24 sources.

The purpose of the work – Development and software implementation of a phishing attack detection system based on artificial intelligence and natural language processing methods.

Object of research – The process of identifying and blocking phishing threats in information and communication networks.

Subject of research – Machine learning algorithms and semantic text analysis technologies for classifying electronic messages.

Summary of the work: The rapid development of digital communications is accompanied by a proportional increase in cyber threats, among which phishing remains the most common attack vector. The use of traditional static filters and signature databases is losing its effectiveness due to the emergence of zero-day attacks and the use of generative AI by attackers to create personalized messages. This dictates an urgent need for the implementation of intelligent security systems capable of dynamically analyzing text semantics and detecting hidden manipulation patterns.

Based on the research, modern machine and deep learning architectures were analyzed. For practical implementation, a representative balanced dataset of over 22,000 messages was formed. Through vectorization with the TF-IDF algorithm and comparative testing of four models, it was proven that the Logistic Regression model provides the optimal balance between high accuracy (98.5%), processing speed, and the ability for threshold adjustment. Based on the selected model, a practical CLI tool was developed, which demonstrates high efficiency in real-time and is ready for integration into modern SaaS systems or email gateways.

KEYWORDS: PHISHING, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, LOGISTIC REGRESSION, TF-IDF, CYBERSECURITY.

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут Інформаційних технологій

ПОДАННЯ
ГОЛОВІ ДЕРЖАВНОЇ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня бакалавра

Направляється здобувач Базілевич І. Ю. до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю 123 Комп'ютерна інженерія
(код, найменування спеціальності)
освітньо-професійної програми Комп'ютерна інженерія
(назва)
на тему: «Система захисту даних від фішингових атак на базі штучного інтелекту».
Кваліфікаційна робота і рецензія додаються.

Директор ННІ _____ Катерина НЕСТЕРЕНКО
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувач Базілевич І. Ю показав відмінну теоретичну підготовку, уміння володіти та розуміння новими комп'ютерними технологіями, користуватися навчальною і науково-технічною літературою. Працюючи над завданнями, які були поставлені керівником, проявив ініціативність, сумлінність та хист до інженерної роботи.
Все це дозволяє оцінити кваліфікаційну роботу студента Базілевича Іллі Юрійовича на оцінку «відмінно», та присвоїти йому кваліфікацію «бакалавр з комп'ютерної інженерії».

Керівник кваліфікаційної роботи _____ Ірина КУЗЬМІЧ
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

«_» _____ 2026 року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувач *Базілевич І. Ю* допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою Комп'ютерної інженерії _____ Наталія ЛАЩЕВСЬКА
(назва) *(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну бакалаврську роботу

здобувача вищої освіти Базілевич Ілля Юрійович
(прізвище, ім'я, по батькові)
на тему: «Система захисту даних від фішингових атак на базі штучного інтелекту»

Актуальність:

Бакалаврська кваліфікаційна робота присвячена одній із найгостріших проблем сучасної IT-галузі — захисту корпоративних та приватних даних від фішингових атак. В умовах, коли традиційні методи фільтрації втрачають свою ефективність через появу таргетованих атак та використання хакерами генеративного штучного інтелекту, тема дослідження, спрямована на застосування передових методів машинного навчання для семантичного аналізу текстів, є надзвичайно актуальною

Позитивні сторони:

1. У роботі наведено глибоке та структуроване теоретичне обґрунтування використання алгоритмів машинного навчання та методів обробки природної мови для виявлення маніпуляцій. Автор вдало аргументував переваги та недоліки різних архітектур.
2. Автор продемонстрував високий рівень компетентності у підготовці даних. Збір композитного датасету із шести незалежних джерел дозволив вирішити проблему дисбалансу класів та забезпечити репрезентативність навчальної вибірки.
3. Практична цінність роботи підтверджується масштабним порівняльним експериментом та створенням реального програмного продукту. Автор не просто обрав найточнішу модель, а зробив зважений інженерний вибір на користь логістичної регресії через її здатність до ймовірнісного прогнозування та високу швидкість.

Недоліки:

1. У роботі недостатньо уваги приділено глибокому аналізу новітніх трансформерних моделей безпосередньо на етапі практичної реалізації. Хоча вони згадуються в теоретичній частині, було б цікаво порівняти легковагову логістичну регресію з важкою мовною моделлю на одному датасеті.
2. Бажано було б розширити функціонал створеного CLI-інструменту, додавши до нього можливість автоматичного парсингу та аналізу вкладених посилань та технічних заголовків електронної пошти, а не обмежуватися лише текстом повідомлення. Це зробило б систему ще більш універсальною в умовах боротьби з комплексними загрозами.

Висновок: *кваліфікаційна бакалаврська робота заслуговує оцінку " відмінно", а здобувач Базілевич І. Ю заслуговує присвоєння кваліфікації: бакалавра з комп'ютерної інженерії.*

Рецензент:
науковий ступінь, вчене звання _____ *підпис* _____ *Ім'я, ПРІЗВИЩЕ*

ЗМІСТ

ВСТУП.....10
1 КОМПЛЕКСНИЙ АНАЛІЗ, ОГЛЯД ТА ПРАКТИКА ВПРОВАДЖЕННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ЗАБЕЗПЕЧЕННЯ БЕЗПЕКИ ДАНИХ ВІД ФІШИНГОВИХ ЗАГРОЗ.....11
1.1 Теоретичні основи фішингу та критична актуальність протидії цій загрозі 11
1.2 Систематизація та детальна типологія сучасних фішингових загроз.....16
1.3 Базові принципи та фундаментальні засади інтеграції технологій штучного інтелекту у сферу кібернетичної безпеки 19
1.4 Аналіз практичних кейсів та реальних прикладів впровадження систем штучного інтелекту для ефективної протидії фішинговим кампаніям.....26
2 КОМПЛЕКСНЕ ДОСЛІДЖЕННЯ ТА ПОРІВНЯЛЬНА ОЦІНКА ЕФЕКТИВНОСТІ АЛГОРИТМІВ МАШИННОГО І ГЛИБИННОГО НАВЧАННЯ У ЗАДАЧАХ ІДЕНТИФІКАЦІЇ ФІШИНГОВИХ ЗАГРОЗ.....35
2.1 Ключові метрики та методологічні критерії для комплексної оцінки ефективності алгоритмів штучного інтелекту.....35
2.2 Аналіз та застосування спеціалізованих класифікаторів машинного навчання для ефективної ідентифікації фішингових ресурсів 40
2.3 Застосування передових архітектур глибинного навчання для складного аналізу та розпізнавання фішингових загроз 45
2.4 Технології обробки природної мови як інструмент семантичного аналізу та виявлення маніпуляцій у тексті 49
3 ПРАКТИЧНА РЕАЛІЗАЦІЯ ТА ПРОЕКТУВАННЯ ПРИКЛАДНОЇ СИСТЕМИ НА БАЗІ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ЕФЕКТИВНОЇ ПРОТИДІЇ ФІШИНГОВИМ ЗАГРОЗАМ ТА ЗАБЕЗПЕЧЕННЯ НАДІЙНОГО ЗАХИСТУ КОНФІДЕНЦІЙНИХ ДАНИХ.....54

3.1 Критичне обґрунтування та аргументація вибору оптимального алгоритму для практичної інтеграції 54
3.2 Етап підготовки інформаційної бази: обґрунтування вибору датасету та методики попереднього очищення дани 58
3.3 Проектування архітектури моделі та процес її машинного навчання 62
3.4 Комплексна оцінка продуктивності та етап глибокого тестування розробленої інтелектуальної моделі 66
ВИСНОВКИ 76
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 78
Додаток А 81

ВСТУП

У сучасному глибоко цифровізованому суспільстві інформаційні технології пронизують усі сфери життєдіяльності, а електронна комунікація стала основою як для приватного спілкування, так і для функціонування глобального бізнесу. Однак ця залежність від цифрових каналів зв'язку створює ідеальне середовище для розвитку кіберзлочинності. Серед усього спектру кіберзагроз фішинг заслужено вважається однією з найпоширеніших та найнебезпечніших форм застосування соціальної інженерії. Його головна небезпека полягає в тому, що атака спрямована не стільки на злом апаратного чи програмного забезпечення, скільки на експлуатацію психологічних вразливостей самої людини.

Сьогодні фішинг еволюціонував від примітивних масових розсилок до високотехнологічних, таргетованих кампаній. Зловмисники активно використовують великі мовні моделі (LLM) для генерації граматично бездоганних, персоналізованих повідомлень, які легко оминають традиційні спам-фільтри. У таких умовах класичні системи кіберзахисту, побудовані на основі статичних правил та сигнатурних баз даних, демонструють критичну неефективність, оскільки вони не здатні реагувати на раніше невідомі атаки.

Вирішення цієї проблеми лежить у площині впровадження методів штучного інтелекту та машинного навчання. Інтелектуальні алгоритми здатні автоматично аналізувати величезні масиви неструктурованих текстових даних, розпізнавати складні семантичні зв'язки та виявляти приховані лінгвістичні маркери маніпуляцій. Метою даної роботи є комплексне дослідження, оцінка та практична розробка ефективної системи виявлення фішингових атак на базі алгоритмів машинного навчання, що здатна адаптуватися до нових загроз та функціонувати в режимі реального часу, забезпечуючи надійний захист конфіденційних даних.

1 КОМПЛЕКСНИЙ АНАЛІЗ, ОГЛЯД ТА ПРАКТИКА ВПРОВАДЖЕННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ЗАБЕЗПЕЧЕННЯ БЕЗПЕКИ ДАНИХ ВІД ФІШИНГОВИХ ЗАГРОЗ

1.1 Теоретичні основи фішингу та критична актуальність протидії цій загрозі

У сучасному глибоко цифровізованому суспільстві, де інформаційні технології пронизують усі сфери життєдіяльності, фішинг заслужено вважається однією з найпоширеніших, наймасштабніших та найнебезпечніших форм застосування методів соціальної інженерії. Цей вектор кібератак становить критичну загрозу не лише для конфіденційності приватних користувачів, але й для безпеки транснаціональних корпорацій, державних установ та об'єктів критичної інфраструктури у всьому світі. Якщо розглядати це поняття більш детально, то фішинг являє собою комплексну зловмисну практику, головною метою якої є маніпулятивне спонукання жертви до добровільної передачі високочутливої та конфіденційної інформації. До такої інформації традиційно належать автентифікаційні дані (логіни та паролі), реквізити банківських карток, внутрішня корпоративна документація, персональні дані тощо. Зловмисники досягають цієї мети шляхом майстерної візуальної та семантичної імітації довірених осіб, легітимних організацій або відомих брендів. На сьогоднішній день архітектура таких атак є мультиканальною: вони реалізуються через компрометовані електронні листи, підроблені веб-ресурси, SMS-розсилки (смішинг), голосові телефонні дзвінки (вішинг), а також через таргетовані повідомлення у популярних месенджерах та соціальних мережах.

Історичний контекст свідчить, що сам термін «фішинг» був введений в обіг ще у середині 1990-х років, у період раннього масового поширення

інтернету. Саме тоді кіберзлочинці почали активно використовувати підроблені електронні повідомлення для незаконного заволодіння паролями від облікових записів користувачів популярного на той час провайдера America Online (AOL) [1]. У своїх перших спробах хакери досить примітивно, але ефективно імітували офіційні службові сповіщення, копіюючи елементи корпоративного стилю служби підтримки, щоб шляхом обману та залякування блокуванням акаунту виманити необхідні облікові дані. Відтоді фішинг пройшов довгий шлях еволюційного розвитку. Сьогодні ця загроза трансформувалася у високотехнологічну індустрію з десятками спеціалізованих підтипів, інноваційними методами доставки шкідливого контенту та адаптацією до новітніх технологічних трендів, таких як повсюдне використання мобільних пристроїв, міграція бізнесу у хмарні сервіси та застосування потужних генеративних моделей штучного інтелекту.

Масштаби проблеми яскраво ілюструє актуальна статистика. Згідно з ґрунтовним звітом авторитетної міжнародної Антифішингової робочої групи (APWG) [2], лише протягом 2023 року у глобальній мережі було офіційно зафіксовано понад 4,7 мільйона унікальних фішингових атак, що стало абсолютним рекордом за весь час проведення подібних спостережень. Структура цілей цих атак також є досить показовою: понад 30% від їх загальної кількості були цілеспрямовано сфокусовані на фінансовому секторі, до якого входять традиційні банки, міжнародні платіжні системи та страхові корпорації, оскільки компрометація таких установ приносить зловмисникам найшвидший прямий прибуток. Другим за рівнем привабливості для хакерів виявився сектор хмарного програмного забезпечення як послуги (SaaS). Зловмисники масовано атакували корпоративні та приватні облікові записи в таких екосистемах, як Google Workspace, Microsoft 365, Dropbox, Zoom та інших аналогічних платформах, щоб отримати закріплення у внутрішніх мережах організацій.

Фундаментальна причина стабільно високої успішності фішингових кампаній полягає в тому, що вони експлуатують не стільки технічні вразливості апаратного чи програмного забезпечення, скільки глибинні

психологічні особливості та когнітивні упередження людини. Серед ключових факторів, на які спираються шахраї, виділяють безумовну довіру до авторитетів (керівництва, державних органів, банків), зниження рівня концентрації уваги під час повсякденної роботи з поштою, а також штучно створену необхідність діяти максимально швидко в умовах імітованої загрози чи сильного емоційного стресу. У переважній більшості випадків фішингові листи містять потужні вербальні стимули, що провокують користувачів на імпульсивні, необдумані дії. Використання таких тригерних фраз, як «ТЕРМІНОВО», «ПОТРІБНА ВАША ДІЯ», «КРИТИЧНЕ ПОПЕРЕДЖЕННЯ БЕЗПЕКИ», штучно підвищує психологічний тиск на людину, змушуючи її відключити критичне мислення. Механізми цього згубного впливу докладно проаналізовані та описані в сучасній науковій літературі, що знаходиться на стику поведінкової психології та кібербезпеки. Зокрема, вичерпна класифікація принципів соціальної інженерії, які найчастіше застосовуються у процесі реалізації фішингових атак, наведена в дослідженні [3].

Залежно від масштабу, методів підготовки та специфіки вибору жертв, фішингові атаки можна розділити на кілька основних категорій. Найбільш поширеним є масовий (або «килимовий») фішинг, під час якого абсолютно однакове, стандартизоване шахрайське повідомлення розсилається навмання сотням тисяч або навіть мільйонам користувачів у надії, що хоча б невеликий відсоток з них потрапить на гачок. Більш небезпечною формою є цільовий, або спеціалізований фішинг (spear phishing), коли атака ретельно готується та глибоко персоналізується під конкретну людину чи невелику групу співробітників. Для цього хакери активно збирають інформацію з відкритих джерел (OSINT), аналізують профілі жертви в професійній мережі LinkedIn, вивчають її активність у соціальних мережах та використовують дані з попередніх масштабних витоків інформації. Окремо варто виділити компрометацію корпоративної електронної пошти (Business Email Compromise, BEC) — надзвичайно складну і фінансово руйнівну форму атаки, яка полягає у високоякісній імітації внутрішнього корпоративного листування.

У таких випадках зловмисники часто діють від імені генерального директора або фінансового керівника, наказуючи підлеглим терміново переказати значні суми коштів на підставні рахунки.

На сучасному етапі розвитку технологій вкрай серйозне занепокоєння у фахівців з кібербезпеки викликає стрімке зростання масштабів так званого генеративного фішингу. Використовуючи передові великі мовні моделі (LLM), такі як GPT, BARD, Claude та їхні аналоги з відкритим вихідним кодом, зловмисники отримали можливість автоматизовано генерувати стилістично бездоганні, граматично правильні, глибоко логічні та максимально персоналізовані тексти для своїх повідомлень. Це дозволяє їм легко обходити традиційні спам-фільтри, які раніше реагували на орфографічні помилки або неприродні мовні конструкції. Такі атаки, повністю або частково згенеровані штучним інтелектом, виглядають набагато переконливіше для кінцевого користувача, а алгоритмічно виявити їх стає вкрай важко, оскільки кожне окреме повідомлення в межах однієї кампанії може бути абсолютно унікальним за формою, зберігаючи при цьому шкідливу суть [4].

Реальна і потенційна небезпека фішингу давно вийшла за межі локальних інцидентів та втрат окремих фізичних осіб. Сьогодні ця загроза стає причиною глобальних катастроф інфраструктурного масштабу. Яскравим прикладом цього є події 2021 року, коли єдиний успішний фішинговий електронний лист став причиною повномасштабного злому комп'ютерної системи компанії Colonial Pipeline, що, в свою чергу, призвело до безпрецедентної тимчасової зупинки значної частини енергетичної системи та дефіциту пального на східному узбережжі США [5]. Не менш руйнівною стала ситуація 2020 року, коли елементарний фішинговий лист виступив початковою точкою входу для сумнозвісної атаки на компанію SolarWinds. Цей інцидент спровокував ланцюгову реакцію та призвів до критичного витоку секретних даних з десятків найважливіших урядових агентств та міністерств Сполучених Штатів. Також показовим є випадок 2023 року, коли внаслідок ретельно спланованої фішингової кампанії, побудованої на бездоганній імітації сервісу

хмарного зберігання Microsoft OneDrive, було успішно скомпрометовано особисті та робочі дані понад 10 тисяч облікових записів, що належали викладачам та працівникам освітнього сектору [6].

Окремо слід відзначити специфіку фішингових загроз у контексті повномасштабної війни в Україні та безперервного поширення гібридних методів ведення бойових дій. У цьому середовищі спостерігається експоненціальне зростання кількості фішингових атак, які цинічно імітують комунікацію від ключових державних установ, благодійних фондів, національних поштових сервісів та популярних платформ надання державних послуг, таких як портал та застосунок «Дія». У цих реаліях фішингові кампанії набули рис не просто звичайного фінансового шахрайства, а перетворилися на потужний інструмент інформаційно-психологічного впливу (ІПСО), поширення дезінформації, посіву паніки та цілеспрямованого кібершпигунства за військовими та цивільним населенням. З огляду на все вищезазначене, стає абсолютно очевидним, що сьогодні фішинг являє собою не просто поширену форму шахрайства, а надзвичайно складну, гнучку, багатовекторну кібератаку, яка безперервно еволюціонує паралельно з розвитком легітимних цифрових інструментів. Відповідно, ефективна протидія цій комплексній загрозі вже не може спиратися виключно на традиційні, сигнатурні заходи безпеки, такі як класичні антивіруси чи базові поштові фільтри. Сучасні виклики вимагають обов'язкового впровадження високоадаптивних, інтелектуальних рішень на основі алгоритмів машинного навчання та штучного інтелекту, які здатні в режимі реального часу аналізувати колосальні масиви даних, виявляти глибоко приховані семантичні закономірності, фіксувати поведінкові аномалії та миттєво реагувати на найслабші сигнали підготовки фішингової атаки.

1.2 Систематизація та детальна типологія сучасних фішингових загроз

Розглядаючи фішинг як комплексне кіберзлочинне явище, слід розуміти, що він охоплює надзвичайно широкий та різноманітний спектр маніпулятивних методів, технологічних інструментів і неочевидних каналів взаємодії, головною метою яких є введення жертви в оману задля подальшої крадіжки її конфіденційних даних. Ефективність та кінцевий успіх будь-якої шахрайської кампанії критично залежать від кількох ключових факторів, серед яких провідне місце займають вибір оптимального каналу доставки шкідливого контенту, глибина персоналізації повідомлення під конкретного користувача, а також загальний рівень психологічної достовірності розробленого зловмисниками сценарію. Детальна класифікація різновидів фішингових атак є життєво необхідною, оскільки вона дозволяє фахівцям з кібербезпеки не лише чітко систематизувати наявні вектори загроз, але й значно точніше та ефективніше розробляти відповідні превентивні заходи захисту. Відповідно до сучасних досліджень, найпоширенішими та найнебезпечнішими типами таких атак на сьогоднішній день є кілька основних категорій [7].

Одним із фундаментальних і найбільш масових різновидів є класичний фішинг електронної пошти. Цей тип атаки історично був першим і досі залишається найпоширенішим інструментом у арсеналі кіберзлочинців. Його механіка полягає у масовому або цілеспрямованому надсиланні ретельно підроблених електронних листів. У процесі реалізації таких схем зловмисники найчастіше видають себе за представників офіційних та авторитетних установ: державних податкових служб, систем охорони здоров’я, популярних соціальних мереж, відомих банківських установ або навіть за колег чи керівництво компанії, в якій працює жертва. Типове шахрайське повідомлення такого формату зазвичай характеризується кількома яскравими ознаками.

Насамперед, воно завжди має надзвичайно привертаючий увагу, а іноді й відверто провокативний рядок теми, наприклад, повідомлення про раптове блокування облікового запису або неочікуване повернення значних коштів. Крім того, текст листа обов'язково містить наполегливий запит на виконання термінової дії з боку користувача, вимагаючи негайно підтвердити персональні дані або перейти за вказаним посиланням для авторизації. Третім обов'язковим елементом є наявність у тілі листа самого шкідливого посилання, що веде на скомпрометований чи підроблений веб-ресурс, або ж прикріпленого файлу в популярних форматах (документи, PDF-файли чи архіви), що містять прихований шкідливий програмний код. Для підвищення довіри атакуючі активно застосовують техніки спуфінгу, підробляючи видиму адресу відправника, реєструють фішингові домени, які візуально майже не відрізняються від легітимних, або ж майстерно маскують справжні шкідливі URL-адреси за допомогою кнопок із закликом до дії чи популярних сервісів скорочення посилань. Масштаб цієї проблеми важко переоцінити, оскільки понад вісімдесят відсотків усіх успішних несанкціонованих проникнень у захищені корпоративні мережі традиційно розпочиналися саме з вдало проведеної атаки через електронну пошту [2].

Тісно пов'язаним із попереднім методом є фішинг веб-сайтів, також відомий як веб-спуфінг. Цей вектор атаки передбачає створення високоякісних підроблених веб-сторінок, які до дрібниць імітують дизайн, структуру та фірмовий стиль реальних легітимних ресурсів, таких як портали онлайн-банкінгу, міжнародні платіжні системи чи служби доставки. Сценарій атаки зазвичай виглядає так: користувач, піддавшись на маніпуляцію, натискає на посилання у фішинговому листі або шкідливій рекламі в інтернеті, після чого потрапляє на веб-інтерфейс, який візуально є абсолютно ідентичним до зовнішнього вигляду знайомого йому оригінального сайту. Усі конфіденційні дані, включаючи логіни, паролі чи CVV-коди банківських карток, які користувач вводить у форми на такому ресурсі, миттєво передаються на сервери зловмисників. До класичних ознак веб-спуфінгу належить

використання доменних імен, які надзвичайно схожі на реальні адреси відомих брендів. Окрім цього, сучасні шахраї легко отримують безкоштовні SSL-сертифікати для своїх доменів, що забезпечує відображення протоколу HTTPS у адресному рядку і створює у жертви хибне відчуття повної безпеки. Також зловмисники дедалі частіше розміщують свої фішингові сторінки на безкоштовних легітимних хмарних хостингах та конструкторах сайтів. Варто зазначити, що такі шахрайські ресурси мають вкрай обмежений життєвий цикл, який зазвичай не перевищує двадцяти чотирьох годин. Для боротьби з цим явищем фахівці застосовують комплексні рішення: ботнети для моніторингу системи доменних імен, краудсорсингові бази даних та передові класифікатори URL-адрес на основі алгоритмів штучного інтелекту, що інтегруються у глобальні системи захисту.

В епоху тотального поширення смартфонів особливої актуальності набув смішинг — специфічний різновид фішингу, в якому як основний канал доставки шахрайських повідомлень використовуються текстові SMS-повідомлення [8]. Рівень небезпеки смішингу оцінюється як надзвичайно високий, оскільки мобільні користувачі психологічно рідше підозрюють SMS і схильні реагувати на них швидко, клікаючи на посилання без роздумів. Крім того, малі екрани смартфонів об'єктивно ускладнюють перевірку легітимності повідомлення, не завжди дозволяючи побачити повну URL-адресу або детально вивчити контактні дані відправника. Типові приклади смішингових повідомлень експлуатують тематику терміновості: це можуть бути тривожні сповіщення про блокування рахунку з вимогою перейти за коротким посиланням, фальшиві повідомлення про очікування посилки з проханням оновити деталі доставки, або ж фейкові привітання з виграшем від банку, для отримання якого необхідно підтвердити реквізити. Під час глобальної пандемії COVID-19 відбувся різкий сплеск смішинг-атак, у яких шахраї підробляли розсилки від міністерств охорони здоров'я, банків та соціальних служб. Згідно з аналітичними даними звіту щодо ландшафту загроз за 2022 рік [9], смішинг набув безпрецедентної популярності в Європі, що тісно пов'язано зі стрімким

зростанням використання QR-кодів та масовим переходом населення на послуги мобільного банкінгу.

Окрім текстових методів, кіберзлочинці активно використовують голосовий канал зв'язку, реалізуючи так званий вішинг, або голосовий фішинг. Цей вид атаки відбувається в режимі реального часу, коли зловмисники телефонують потенційній жертві, видаючи себе за працівників банку, поліції або служби підтримки. Головна мета такого дзвінка полягає у тому, щоб викликати у людини стан паніки чи терміновості та змусити її поділитися конфіденційною інформацією або самостійно перевести кошти. Технічний арсенал вішингу включає технологію підміни ідентифікатора абонента, завдяки якій на екрані жертви висвічується офіційний номер установи. Для посилення ефекту достовірності часто використовуються звукові ефекти, що імітують шум банківського кол-центру. Проте найбільшу тривогу останнім часом викликає зростання випадків deepfake-вішингу, де голос атакуючого генерується та підробляється за допомогою нейромереж. Наприклад, штучний інтелект здатен зімітувати голос генерального директора компанії, який дзвонить бухгалтеру з проханням негайно переказати кошти. Подібні інциденти вже стали реальністю; зокрема, в літературі задокументовано випадок [11], коли завдяки майстерній імітації голосу керівника шахраям вдалося виманити двісті двадцять тисяч євро.

1.3 Базові принципи та фундаментальні засади інтеграції технологій штучного інтелекту у сферу кібернетичної безпеки

Стрімкий розвиток та впровадження технологій штучного інтелекту (ШІ) радикально, на фундаментальному рівні, трансформують традиційні парадигми та методи захисту складних інформаційних систем. Особливо відчутним цей вплив є у критичній сфері виявлення, ідентифікації та

оперативного реагування на багатовекторні фішингові атаки. Сучасні інтелектуальні системи на базі ШІ володіють безпрецедентною здатністю агрегувати та аналізувати колосальні обсяги як чітко структурованих, так і повністю неструктурованих даних (таких як неформатований текст, графічні елементи, поведінкові логи) у режимі реального часу. На відміну від класичних інструментів кіберзахисту минулого покоління, які жорстко обмежені використанням статичних баз сигнатур та заздалегідь прописаних фільтрів, системи штучного інтелекту здатні автономно виявляти неочевидні аномалії, гнучко адаптуватися до появи абсолютно нових, раніше невідомих патернів кібератак та безперервно самовдосконалюватися в процесі роботи. У контексті глобальної кібербезпеки це відкриває принципово нові горизонти: від забезпечення по-справжньому проактивного виявлення високотехнологічних фішингових повідомлень ще до моменту їх потрапляння до кінцевого користувача, і аж до повністю автоматизованого, миттєвого реагування на складні інциденти без безпосереднього втручання людини-оператора.

Наріжним каменем і фундаментом переважної більшості сучасних комерційних та академічних систем виявлення фішингу виступає машинне навчання (МН) — ключовий розділ штучного інтелекту. Основна концептуальна ідея застосування МН у цій сфері полягає у розробці та тренуванні математичних моделей класифікації. Така модель, аналізуючи тисячі прикладів, алгоритмічно навчається з високою точністю розрізняти зловмисні фішингові повідомлення та безпечні легітимні листи на основі ретельно підібраного набору специфічних ознак (так званих фіч). Серед найпопулярніших та математично обґрунтованих алгоритмів машинного навчання, які сьогодні застосовуються для вирішення цього завдання, можна виділити метод випадкового лісу (Random Forest), який будує множину дерев прийняття рішень для підвищення стабільності результату; метод опорних векторів (Support Vector Machine), що знаходить оптимальну гіперплощину для розділення класів у багатовимірному просторі; класичні дерева рішень (Decision Trees); потужні ансамблеві методи градієнтного бустингу (Gradient

Boosted Trees або XGBoost), алгоритм k-найближчих сусідів (k-NN) та метод наївного Байєса (Naive Bayes), заснований на теорії ймовірностей.

Для успішного функціонування цих алгоритмів необхідна ретельна екстракція ознак з аналізованих об'єктів. Типові індикатори, які вказують на високу ймовірність фішингу, включають аномально велику кількість субдоменів у структурі URL-адреси, надмірну загальну довжину посилання, пряме використання IP-адреси замість зареєстрованого доменного імені. Окрім цього, алгоритми детально аналізують приховану структуру HTML-коду самого повідомлення, фіксують наявність маніпулятивних тригерних слів у тексті (таких як наполегливі заклики «терміново», «негайно перевірте», «обов'язково натисніть тут»), а також виявляють візуальні ознаки маніпуляції у полі теми листа, наприклад, надмірне використання великих літер (КАПСЛОК), знаків оклику або спеціальних символів валют. Відповідно до результатів ґрунтовного дослідження, представленого в науковій роботі [11], правильно налаштовані моделі машинного навчання здатні стабільно досягати вражаючої точності на рівні від 93 до 97 відсотків у процесі виявлення фішингових електронних листів. Проте автори наголошують, що такі високі показники можливі виключно за умови використання релевантного набору ознак та проведення ретельного попереднього очищення масивів даних перед початком фази тренування моделі.

Більш досконалим та концептуально складнішим напрямком є глибинне навчання (ГН) — спеціалізований підрозділ машинного навчання, який базується на використанні багатошарових штучних нейронних мереж з великою кількістю прихованих шарів. Головна перевага та обчислювальна сила алгоритмів глибинного навчання полягає в їхній унікальній здатності до автоматичного, ієрархічного вилучення репрезентативних ознак безпосередньо з сирих даних. Це позбавляє розробників необхідності проводити складний, трудомісткий та часто суб'єктивний процес ручного конструювання фіч (feature engineering). У складних архітектурах захисту від фішингу технології глибинного навчання активно застосовуються для глибокої

семантичної обробки неструктурованого тексту (включаючи тіло електронних листів, вміст веб-форм та службові заголовки), комплексного аналізу архітектури веб-сторінок (аналіз DOM-дерева, каскадних таблиць стилів та вбудованих скриптів), комп'ютерного зору для класифікації зображень (наприклад, аналіз скріншотів підозрілих веб-інтерфейсів та графічних вкладень), а також для виявлення неочевидних поведінкових патернів.

Для вирішення цих різнопланових завдань використовуються специфічні архітектури нейронних мереж. Наприклад, рекурентні нейронні мережі з довгою короткостроковою пам'яттю (LSTM) та їх двонаправлені варіанти (BiLSTM) ідеально підходять для аналізу природного тексту електронних листів, оскільки вони математично здатні враховувати контекстуальну послідовність та порядок слів у реченні. Згорткові нейронні мережі (CNN), навпаки, спеціалізуються на обробці матричної інформації і тому масово використовуються для аналізу зображень, допомагаючи алгоритмічно виявляти візуальну схожість підозрілого веб-сайту з легітимним оригіналом. Автоенкодери застосовуються у системах без вчителя для знаходження прихованих аномалій у трафіку чи структурі даних. Окремої уваги заслуговують моделі на базі архітектури трансформерів (такі як BERT або його оптимізовані версії типу MiniLM), які здійснили справжню революцію в галузі обробки природної мови. Аналіз наукової літератури, зокрема дослідження [12], переконливо продемонстрував, що сучасні моделі глибинного навчання суттєво перевершують класичні алгоритми машинного навчання під час обробки найбільш складних та заплутаних випадків фішингу, особливо тоді, коли атака спирається на комбінацію заплутаного HTML-коду, шкідливого JavaScript та обманної графічної інформації.

У контексті забезпечення безпеки корпоративних поштових систем ключовим та невід'ємним компонентом захисту стала технологія обробки природної мови (NLP). Інструменти NLP, інтегровані у поштові шлюзи, забезпечують глибокий синтаксичний та семантичний аналіз тексту кожного електронного листа. Вони здатні виявляти найменші стилістичні

невідповідності, проводити автоматичну категоризацію загальної тональності та емоційного забарвлення повідомлення (виявляючи маркери штучної терміновості чи погрози), а також математично співставляти стиль написання поточного листа з історичним профілем листування конкретного відправника. У цьому напрямку архітектура трансформерів, представлена моделями BERT або DistilBERT, демонструє безпрецедентну ефективність у розпізнаванні складних маніпулятивних патернів. Концепція Transfer Learning дозволяє взяти базову модель BERT, що розуміє загальну мову, та дотренувати її на спеціалізованому корпусі фішингових текстів, створюючи таким чином потужний інструмент для виявлення тонких психологічних маніпуляцій. Так, у рамках згаданого раніше дослідження [12] було експериментально підтверджено, що фішингові листи майже завжди демонструють специфічні, малопомітні для людини відхилення від природної граматичної та синтаксичної структури, які, однак, чітко математично ідентифікуються багатовимірною моделлю трансформера.

Сфера застосування штучного інтелекту не обмежується лише аналізом контенту, вона також охоплює напрямок аналізу поведінки користувачів та сутностей (UBA). Такі інтелектуальні системи безперервно відстежують та формують цифровий профіль «нормальної» поведінки кожного користувача у корпоративній мережі або хмарних системах, фіксуючи типовий час авторизації, звичну географію підключень та стандартний набір виконуваних операцій. У подальшому цей еталонний профіль у режимі реального часу алгоритмічно порівнюється з поточною активністю. У випадку фіксації статистично значущої аномалії (наприклад, спроба авторизації з абсолютно нової IP-адреси, нестандартний час роботи або незвична послідовність кліків у веб-інтерфейсі поштового клієнта), система здатна повністю автоматично активувати механізми захисту: вимагати проходження додаткової багатофакторної аутентифікації, тимчасово заблокувати доступ до чутливих ресурсів або надіслати критичне сповіщення адміністраторам безпеки. Показовим прикладом практичної реалізації такого підходу є рішення

Microsoft Defender, яке містить глибоко інтегровані моделі ШІ. Ці моделі здатні миттєво реагувати на найменші зміни у поведінковому паттерні облікового запису відразу після його ймовірної компрометації внаслідок фішингу — система фіксує різкі стрибки частоти запитів до сервера, зміни у маршрутизації трафіку та аномальне використання програмних інтерфейсів додатків (API).

Резюмуючи вищевикладене, можна впевнено стверджувати, що у глобальному контексті боротьби з кіберзагрозами, методи, засновані на штучному інтелекті, демонструють переконливу низку стратегічних переваг, що зумовлює їх дедалі ширше впровадження у сучасні архітектури кіберзахисту. Синергетичне застосування класичних алгоритмів машинного навчання, передових архітектур глибинного навчання та складних моделей обробки природної мови дозволяє організаціям кардинально підвищити загальну точність виявлення загроз, мінімізувати час реагування на інциденти, перевести процеси цілодобового моніторингу в автоматичний режим та критично знизити вплив горезвісного людського фактору. Інтелектуальні системи, на відміну від архаїчних рішень, що цілком покладалися на ручне оновлення баз даних сигнатур або створення статичних правил фільтрації, демонструють вражаючу технологічну гнучкість. Вони здатні самостійно адаптуватися до безперервно змінного ландшафту загроз, блискавично обробляти петабайти інформації та виявляти заплутані фішингові схеми без потреби у постійному ручному налаштуванні інженерами. Наприклад, добре навчені моделі здатні розпізнавати приховані поведінкові маркери фішингу, досягаючи стабільної точності ідентифікації понад 90 відсотків навіть у тих випадках, коли лист або сайт не містить жодних прямих, очевидних для традиційних сканерів ознак атаки.

Крім того, нейромережеві системи, зокрема ті, що базуються на архітектурах BERT або рекурентних мережах типу LSTM, наділені здатністю глибоко розуміти семантичний контекст аналізованих повідомлень. Це дозволяє їм ефективно протидіяти найсучаснішим обхідним технікам, які масово застосовуються у таргетованих фішингових кампаніях, включаючи ті

атаки, що генеруються автоматично за допомогою великих мовних моделей (LLM), які зловмисники використовують для створення ідеально написаних, персоналізованих текстів. Не менш важливою перевагою алгоритмів ШІ є їхня здатність до безперервної адаптації та еволюції. Завдяки механізмам регулярного автоматизованого донавчання на нових зразках підтверджених загроз, такі інтелектуальні платформи здатні успішно ідентифікувати та блокувати фішингові атаки так званого нульового дня (zero-day) — тобто ті загрози, які є абсолютно новими, ще не зафіксовані жодними аналітичними центрами, не внесені до глобальних чорних списків та не мають жодних відомих цифрових сигнатур [13].

Додатковою, але критично важливою вигодою від впровадження технологій ШІ є суттєве зниження загальної кількості хибнопозитивних спрацьовувань (false positives) систем безпеки. Ця проблема історично була однією з найбільших перешкод для ефективної роботи центрів операцій з кібербезпеки (SOC), оскільки велика кількість помилкових тривог призводила до виснаження аналітиків (alert fatigue) та пропускання реальних загроз. Переконливим підтвердженням ефективності ШІ у цьому контексті є опубліковані результати спільних досліджень провідних технологічних корпорацій Google та Sophos. Вони продемонстрували, що лише базова інтеграція локальних LLM-моделей безпосередньо у ядро веб-браузера Chrome призвела до вражаючого результату: загальна ймовірність успішного переходу користувача на шахрайський ресурс знизилася на понад 80 відсотків [14]. Варто також наголосити на високій масштабованості підходів, заснованих на ШІ. Програмні рішення цього типу легко розгортаються та інтегруються у складні, географічно розподілені корпоративні інфраструктури, глобальні системи маршрутизації електронної пошти та хмарні середовища. Однак, незважаючи на беззаперечні технічні та операційні переваги, масове використання штучного інтелекту у сфері безпеки також вимагає серйозного підходу до вирішення питань етичного використання даних, захисту приватності (GDPR) та створення відповідної нормативної регламентації.

Підсумовуючи, можна стверджувати, що інтелектуальні підходи на основі ШІ мають колосальний, ще не до кінця розкритий потенціал для забезпечення надійного, проактивного захисту від фішингу, а їх висока дієвість щоденно підтверджується як практичними кейсами від провідних гравців індустрії кібербезпеки, так і фундаментальними результатами сучасних наукових досліджень.

1.4 Аналіз практичних кейсів та реальних прикладів впровадження систем штучного інтелекту для ефективної протидії фішинговим кампаніям

Багаторічна, стійка тенденція до експоненціального зростання як кількості, так і рівня складності фішингових атак стала потужним каталізатором для глобальних технологічних лідерів. Ця загрозлива ситуація безальтернативно спонукала їх до глибокої та масштабної інтеграції передових механізмів штучного інтелекту безпосередньо у ядра свого захисного програмного забезпечення та хмарних екосистем. Інноваційні рішення, побудовані на базі машинного навчання та нейромереж, кардинально відрізняються від своїх попередників: вони наділені здатністю проактивно, на випередження реагувати на появу абсолютно нових, раніше невідомих патернів атак. Замість того, щоб чекати на оновлення баз даних від аналітиків, ці системи самостійно та безперервно генерують динамічні, адаптивні сигнатури, миттєво виявляють найтонші статистичні та поведінкові аномалії в режимі реального часу і, що найважливіше, ефективно захищають кінцевих користувачів ще до того моменту, як шкідливе повідомлення чи небезпечне посилання фізично надійде до їхньої поштової скриньки або відобразиться у веб-браузері.

Яскравим прикладом такого масштабного підходу є стратегія корпорації Google, яка успішно розгорнула комплексну, багатошарову архітектуру ШІ, що наскрізь пронизує усі її ключові сервіси — від електронної пошти Gmail та браузера Chrome до глобальної пошукової системи та мобільної операційної системи Android. Основою захисту поштового сервісу є інтелектуальний фільтр спаму Gmail, де надпотужні алгоритми глибинного та машинного навчання щомісяця аналізують та сортують колосальний обсяг даних — понад 300 мільярдів електронних листів. Згідно з офіційними внутрішніми оцінками компанії, завдяки цим алгоритмам понад 99,9 відсотка всіх надісланих фішингових листів успішно ідентифікуються та надійно блокуються ще на стадії маршрутизації, так і не досягаючи вхідних скриньок користувачів [15]. Паралельно функціонує система Google Safe Browsing — глобальний сервіс, який використовується не лише у Chrome, а й інтегрований у браузери Safari та Firefox для своєчасного попередження мільярдів користувачів про спроби переходу на шкідливі або скомпрометовані сайти. Щоденно спеціалізовані моделі ШІ цього сервісу глибоко аналізують понад 40 мільйонів підозрілих URL-адрес. Унікальність Safe Browsing полягає в його здатності оновлюватися в режимі реального часу, алгоритмічно виявляючи та вносячи до чорних списків нові фішингові ресурси буквально протягом кількох хвилин після їхньої фактичної появи в мережі Інтернет. Більше того, починаючи з 2024 року, Google здійснив революційний крок, розпочавши пряму інтеграцію оптимізованих великих мовних моделей, таких як Gemini Nano, безпосередньо у мобільну версію браузера Chrome на платформі Android. Ця модель здатна локально, використовуючи лише обчислювальні потужності самого смартфона, сканувати структуру форм входу на сайтах та миттєво попереджати користувачів про наявність прихованих фішингових елементів, причому цей захист працює навіть за повної відсутності підключення до інтернету, що гарантує конфіденційність аналізу [16].

Не менш вражаючим є досвід корпорації Microsoft, яка зробила штучний інтелект наріжним каменем безпеки своєї гігантської корпоративної

екосистеми, що включає платформи Office 365, поштовий сервіс Outlook та хмарну інфраструктуру Azure. Центральним елементом цього захисту виступає комплексне рішення Microsoft Defender for Office 365. Воно інтенсивно використовує різноманітні методи машинного навчання для глибокого аналізу та виявлення небезпечних фішингових вкладень, ідентифікації складних спроб підміни реального відправника та блокування шкідливих URL-адрес у тілі листів. Масштаби роботи системи вражають: лише протягом 2023 року алгоритми Defender успішно виявили та превентивно заблокували понад 35 мільярдів різноманітних фішингових загроз, спрямованих на корпоративних та приватних клієнтів [17]. Додатково Microsoft пропонує інноваційні інструменти Attack Simulator та Threat Explorer. Завдяки вбудованим механізмам ШІ, ці сервіси дозволяють службам безпеки компаній автоматизовано симулювати реалістичні, персоналізовані фішингові атаки всередині власної організації, після чого система самостійно досліджує виявлені вразливості, оцінює рівень ризику та формує індивідуальні програми навчання для персоналу, який піддався на провокацію. Важливим компонентом є також платформа Microsoft Graph AI та Behavioral Analytics, яка здійснює синергетичне об'єднання аналітики вмісту електронної пошти з глибоким поведінковим аналізом користувачів. Система безперервно відстежує такі параметри, як аномальні географічні локації входів, нетиповий час масового надсилання електронної пошти чи різкі зміни у стилі комунікації. У разі виявлення ознак компрометації облікового запису, платформа здатна діяти автономно: вона може миттєво примусово закрити поточні сесії, скинути пароль або повністю заблокувати вихідну маршрутизацію пошти з цього акаунта, запобігаючи поширенню інфекції всередині корпоративної мережі.

Компанія Cloudflare, один із світових лідерів у сфері мережевої безпеки та доставки контенту, також глибоко інтегрувала ШІ у свою передову архітектуру Zero Trust та спеціалізований сервіс Cloudflare Area 1. Рішення Area 1 Email Security фокусується на превентивному захисті поштових шлюзів, активно використовуючи складні моделі обробки природної мови (NLP). Ці

моделі проводять глибокий синтаксичний та семантичний аналіз структури кожного електронного листа, детально вивчають приховані HTML-елементи та перевіряють логіку побудови повідомлення. Основна мета системи — алгоритмічно виявити фішингові наміри ще до моменту фактичної доставки листа на поштовий сервер клієнта. Щомісяця ця платформа успішно обробляє та фільтрує понад 10 мільярдів електронних повідомлень [18]. На мережевому рівні діє сервіс Cloudflare Gateway, який надійно блокує спроби користувачів перейти на фішингові сайти. Цей процес забезпечується завдяки миттєвій класифікації запитуваних URL-адрес на основі навчених моделей ШІ, глибокому аналізу DNS-запитів та оцінці загальної моделі поведінки користувача під час веб-серфінгу. Якщо, наприклад, запитувана URL-адреса є абсолютно новою, містить підозрілі лінгвіністичні патерни (наприклад, тайпсквотинг) або візуально мімікрує під відомий світовий бренд, ШІ-шлюз автоматично та безповоротно блокує весь трафік до цього ресурсу. Окремо варто відзначити інноваційне рішення Page Shield, яке застосовує комп'ютерний зір та ШІ для безперервного моніторингу поведінки всього стороннього JavaScript-коду, що виконується безпосередньо у браузерах клієнтів під час відвідування захищених сторінок. Якщо алгоритм фіксує, що якийсь невідомий або скомпрометований скрипт намагається несанкціоновано перехопити та вкрасти конфіденційні дані, що вводяться користувачем у веб-форми, платформа реагує миттєво, автоматично відсікаючи та нейтралізуючи такий шкідливий код.

Хмарний гігант Amazon також активно застосовує технології ШІ у цілому ряді своїх сервісів кібербезпеки для забезпечення надійного захисту масштабних хмарних середовищ своїх клієнтів (AWS). Ключовим інструментом тут є сервіс Amazon GuardDuty. Його інтелектуальна модель безперервно отримує, агрегує та аналізує гігантські потоки подій з різних джерел, включаючи логи аудиту CloudTrail, історію DNS-запитів та метрики віртуальних приватних мереж (VPC). Як приклад роботи: якщо система виявляє, що обліковий запис співробітника був ймовірно скомпрометований

внаслідок фішингової атаки (що підтверджується аномальним патерном авторизації), і відразу після цього з даного акаунта розпочинається масове завантаження або видалення об'єктів у відкритих хмарних сховищах (buckets), алгоритм GuardDuty миттєво ідентифікує цю активність як критичну загрозу та автоматично запобігає витоку або знищенню даних, блокуючи доступ. Інший спеціалізований сервіс, Amazon Macie, фокусується безпосередньо на захисті самих даних. Він використовує моделі машинного навчання для автоматизованого виявлення, класифікації та захисту висококонфіденційної інформації (наприклад, персональних даних або фінансових звітів), що зберігається у хмарі, постійно відстежуючи будь-яку аномальну активність, пов'язану з доступом до цих масивів після потенційної фішингової компрометації. Крім того, при використанні корпоративного сервісу масових розсилок Amazon SES діє потужний внутрішній фільтр SpamClassifier на базі ШІ. Його завдання — суворо контролювати вихідний трафік, миттєво зупиняючи будь-які спроби несанкціонованої підміни адрес або масового надсилання листів, що за своїми характеристиками класифікуються нейромережею як потенційно фішингові.

Практичні досягнення корпорацій знаходять своє підтвердження і фундаментальне пояснення в науковому середовищі. Останні кілька років характеризуються надзвичайно інтенсивною як науковою, так і прикладною роботою у специфічній сфері застосування штучного інтелекту для розпізнавання та нейтралізації фішингових атак. Значний обсяг проведених академічних досліджень переконливо доводить високу ефективність синергії алгоритмів машинного навчання, глибинних нейронних мереж та методів обробки природної мови для точної класифікації, ефективного запобігання та надійного блокування фішингу. Наприклад, у ґрунтовній оглядовій статті [11] автори провели масштабний мета-аналіз понад сотні наукових робіт, присвячених використанню ШІ в системах антифішингу. За результатами аналізу, дослідники систематизували всі існуючі підходи, розділивши їх на чотири основні категорії. До першої групи увійшли класичні моделі на основі

машинного навчання, такі як метод опорних векторів (SVM), дерева рішень та алгоритм випадкового лісу (Random Forest). Другу категорію склали передові архітектури глибинного навчання, зокрема рекурентні (LSTM) та згорткові (CNN) нейронні мережі. Третя група об'єднала потужні гібридні моделі, що використовують методи ансамблювання (наприклад, стекинг та бустинг). Четверта категорія включала сценарні підходи, які комбінують експертні системи на основі жорстких правил із гнучкими алгоритмами МН. Результати дослідження показали, що саме ансамблеві техніки, які об'єднують переваги кількох різних алгоритмів, продемонстрували найкращу загальну продуктивність. Точність виявлення фішингових листів за допомогою таких методів досягала вражаючого показника у 98,7 відсотка, а гармонійне середнє між точністю та повнотою (F1-score) стабільно перевищувало 0,95. Особливо ефективною, згідно з висновками авторів, виявилася складна комбінація алгоритму градієнтного бустингу XGBoost, спеціалізованої нейронної мережі та вдосконаленої системи динамічної індукції правил (eDRI).

Окремим, надзвичайно актуальним напрямком досліджень є протидія новітнім формам загроз. Так, у науковій статті [4] детально розглядається та аналізується найновіша та найнебезпечніша на сьогоднішній день загроза — високоякісний фішинг, тексти для якого генеруються автоматично за допомогою сучасних великих мовних моделей (LLM), подібних до GPT-4. Для проведення експерименту автори свідомо використовували відкритий програмний інтерфейс DeepAI API, за допомогою якого згенерували величезну кількість унікальних, стилістично бездоганних та граматично правильних фішингових листів, які неможливо було виявити традиційними методами. Для боротьби з цим новим класом загроз дослідники випробували кілька різних аналітичних моделей. Серед них були системи аналізу стилістичних дескрипторів тексту (UDAT), імовірнісні моделі Наївного Байєса та методи на основі обчислення інформаційної ентропії, рекурентні нейронні мережі довгої короткострокової пам'яті (LSTM), а також комплексний ансамбль із кількох різних моделей. Результати тестування виявилися вельми оптимістичними:

складений ансамбль моделей продемонстрував надзвичайно високу точність виявлення згенерованих ШІ текстів — понад 99,2 відсотка, переконливо перевершивши як усі існуючі комерційні традиційні фільтри, так і результати окремо взятих алгоритмів машинного навчання.

Для об'єктивної оцінки розроблених рішень наукова спільнота створює спеціалізовані інструменти. Зокрема, у публікації [19] детально представлено архітектуру платформи PhishBench. Це інноваційний інструментарій, розроблений спеціально для стандартизованого, справедливого порівняння продуктивності та надійності різних моделей виявлення фішингу в абсолютно однакових, контрольованих лабораторних умовах. Проведені на цій платформі масштабні тестування дозволили зробити кілька важливих фундаментальних висновків. По-перше, було експериментально підтверджено, що більшість існуючих моделей суттєво втрачають свою загальну точність в умовах сильного природного дисбалансу класів (наприклад, коли у навчальній вибірці співвідношення фішингових листів до нормальних становить один до десяти, що відповідає реальній ситуації в мережі). По-друге, дослідження довело, що складні нейромережеві моделі обробки природної мови (такі як архітектура BERT або алгоритм FastText) працюють значно краще та точніше, ніж класичні алгоритми машинного навчання, проте їх перевага розкривається повною мірою лише за наявності достатньо великого обсягу тексту в аналізованому листі. По-третє, було встановлено, що класичний алгоритм випадкового лісу (Random Forest) продовжує залишатися надзвичайно надійним інструментом, стабільно забезпечуючи понад 95 відсотків точності ідентифікації, але за однієї обов'язкової умови — наявності попередньо ретельно підготовленого, якісно структурованого набору ознак.

Підсумовуючи досвід індустрії та наукові дані, можна виділити ключові показники ефективності. Корпорація Google щомісяця успішно та безперебійно обробляє понад 300 мільярдів електронних листів, використовуючи для їх глибокого аналізу передові мовні моделі типу BERT, інтегровані безпосередньо у ядра поштових серверів Gmail. Система безпеки

Microsoft Defender щоденно аналізує та фільтрує понад 2 мільярди корпоративних електронних листів; її моделі машинного навчання з високою точністю визначають підозрілі вкладення, ідентифікують скомпрометованих відправників та виявляють приховані патерни атак. Платформа Cloudflare Area 1 успішно використовує алгоритми ШІ для забезпечення виключно попереджувального виявлення загроз, блокуючи фішинг ще до моменту його доставки (реалізуючи концепцію превентивного захисту). Компанія Zscaler активно поєднує методи ШІ та NLP для забезпечення надійного, безперервного захисту популярних SaaS-додатків для корпоративної комунікації (таких як Microsoft Teams, корпоративний месенджер Slack, платформа для відеоконференцій Zoom) від спроб несанкціонованої підміни облікових записів та перехоплення сесій. Усі ці наведені приклади наочно та переконливо демонструють не просто якийсь абстрактний теоретичний потенціал, а цілком реальне, життєздатне та критично необхідне використання технологій штучного інтелекту для успішної протидії еволюціонуючому фішингу в рамках надскладних, великомасштабних глобальних інформаційних систем.

Таким чином, перший розділ даного дослідження надав вичерпний, детальний огляд поточного середовища фішингових загроз. Було наведено розширене визначення цього явища, запропоновано його сучасну класифікацію та проаналізовано найсвіжіші тенденції у розвитку технік проведення атак. Окрім цього, було глибоко проаналізовано ключові, найперспективніші напрями застосування технологій ШІ — зокрема, алгоритмів класичного машинного навчання, архітектур глибинного навчання, методів обробки природної мови та комплексних систем аналітики поведінки користувачів — у безпосередньому контексті завдань з раннього виявлення та надійного запобігання фішинговим атакам. Загальні та фундаментальні висновки цього розділу однозначно свідчать про те, що інноваційні методи кіберзахисту, які концептуально засновані на використанні штучного інтелекту, не просто доповнюють, а суттєво, на якісно новому рівні покращують загальну ефективність систем виявлення загроз, залишаючи

далеко позаду традиційні, сигнатурні підходи. Їхня унікальна здатність блискавично адаптуватися до появи абсолютно нових, безпрецедентних патернів кібератак, вміння ефективно працювати з гігантськими масивами неструктурованих даних та критично мінімізувати кількість хибних спрацювань робить ці технології абсолютно незамінними, ключовими інструментами у побудові архітектури сучасної кібербезпеки. Переконливі приклади з повсякденної практики таких технологічних гігантів, як Google, Microsoft, Cloudflare та багатьох інших компаній, є найкращим практичним підтвердженням високої доцільності, технічної зрілості та безперечної ефективності такої глибокої інтеграції ШІ у процеси захисту інформації на всіх рівнях.

2 КОМПЛЕКСНЕ ДОСЛІДЖЕННЯ ТА ПОРІВНЯЛЬНА ОЦІНКА ЕФЕКТИВНОСТІ АЛГОРИТМІВ МАШИННОГО І ГЛИБИННОГО НАВЧАННЯ У ЗАДАЧАХ ІДЕНТИФІКАЦІЇ ФІШИНГОВИХ ЗАГРОЗ

2.1 Ключові метрики та методологічні критерії для комплексної оцінки ефективності алгоритмів штучного інтелекту

Одним із найбільш фундаментальних та критично важливих етапів у процесі розробки, тестування та впровадження систем кібербезпеки є вибір, обґрунтування та застосування релевантних метрик для об'єктивної оцінки продуктивності алгоритмів машинного та глибинного навчання. Без належного математичного та статистичного апарату неможливо достовірно визначити, наскільки добре створена модель справляється зі своїм завданням. У специфічному контексті виявлення фішингових атак дослідники та інженери переважно мають справу з класичною задачею бінарної класифікації. Це означає, що будь-який аналізований об'єкт — чи то підозрілий веб-сайт, складна URL-адреса, чи вхідний електронний лист — нейромережа повинна віднести до одного з двох взаємовиключних класів: об'єкт або ідентифікується як шкідливий (фішинговий), або класифікується як легітимний (безпечний). Таке жорстке бінарне розділення вимагає глибокого розуміння принципів побудови так званої матриці помилок, яка є базовим інструментом для розрахунку всіх подальших оціночних показників.

Замість того, щоб просто констатувати факт правильної чи неправильної роботи алгоритму, матриця помилок дозволяє деталізувати результати тестування, розкладаючи їх на чотири базові логічні компоненти, які утворюють фундамент для комплексного аналізу. Першим компонентом є істинно позитивний результат, який фіксує ті ідеальні ситуації, коли штучний інтелект успішно виявив загрозу, правильно класифікувавши реальне

фішингове повідомлення саме як фішингове. Другим компонентом виступає істинно негативний результат, що відображає коректну роботу фільтра у стандартних умовах, коли модель безпомилково розпізнала легітимне повідомлення і пропустила його як безпечне. Третій компонент — хибно позитивний результат — описує вкрай небажані ситуації помилкового спрацьовування, коли надмірно агресивна модель помилково класифікує абсолютно безпечне, робоче повідомлення як небезпечну фішингову атаку, відправляючи його до спаму або повністю блокуючи. Четвертий, і мабуть, найнебезпечніший компонент — хибно негативний результат. Він фіксує критичні провали системи захисту, коли математична модель не змогла розпізнати ретельно замасковану фішингову атаку, пропустивши шкідливий об'єкт до користувача під виглядом безпечного.

Саме на основі глибокого статистичного аналізу цих чотирьох базових значень обчислюються всі найважливіші комплексні метрики ефективності. Найбільш інтуїтивно зрозумілою метрикою є загальна точність (Accuracy). Логіка її розрахунку полягає у визначенні частки абсолютно всіх правильних рішень моделі (як позитивних, так і негативних) у загальному масиві всіх проведених класифікацій. Однак, незважаючи на уявну простоту, у сфері кібербезпеки ця метрика часто виявляється підступною та необ'єктивною. Причиною цього є феномен критичного дисбалансу класів у реальному мережевому трафіку. На практиці частка реального фішингу серед усього масиву електронної пошти може становити менше п'яти відсотків. У таких умовах виникає «парадокс точності»: якщо створити примітивну модель, яка взагалі нічого не аналізує, а просто маркує абсолютно всі листи підряд як «безпечні», її загальна точність формально складе дев'яносто п'ять відсотків, хоча її реальна корисність для кіберзахисту дорівнюватиме нулю, адже вона повністю ігноруватиме всі існуючі загрози.

Щоб уникнути подібних маніпуляцій зі статистикою, фахівці використовують більш спеціалізовані показники, серед яких ключове місце займає влучність (Precision). Ця метрика розраховується шляхом ділення

кількості дійсно виявлених фішингових атак на загальну кількість усіх об'єктів, які система назвала фішингом (включаючи і помилкові спрацьовування). Влучність є критично важливою характеристикою для комерційного середовища та бізнес-процесів, оскільки високий показник цієї метрики гарантує мінімізацію хибно позитивних результатів. Іншими словами, чим вища влучність, тим менша ймовірність того, що важливий діловий контракт або фінансовий звіт від реального клієнта буде помилково заблоковано системою безпеки. З іншого боку барикад знаходиться метрика повноти (Recall). Алгоритм її обчислення передбачає визначення частки знайдених системою загроз від загальної фактичної кількості всіх фішингових листів, що об'єктивно існували в тестовій вибірці. Повнота виходить на перший план і стає абсолютним пріоритетом у висококритичних середовищах — таких як урядові мережі, банківські інфраструктури чи об'єкти критичної інфраструктури. У цих випадках головна мета полягає в тому, щоб не пропустити жодної, навіть найменш ймовірної загрози, тому служби безпеки готові миритися з певною кількістю хибних тривог (пожертвувавши влучністю) заради того, щоб перехопити всі реальні атаки.

Оскільки в реальних умовах завжди існує компроміс між влучністю та повнотою (спроба максимізувати один показник зазвичай призводить до падіння іншого), дослідникам потрібна була єдина, комплексна оцінка. Таким стандартом став так званий F1-score (F-міра). Цей показник обчислюється як специфічне гармонійне середнє значення між показниками влучності та повноти. На відміну від звичайного середнього арифметичного, математична властивість гармонійного середнього полягає в тому, що воно жорстко карає модель за екстремально низькі значення будь-якого з двох базових параметрів. Таким чином, високий бал F1-score неможливо отримати, якщо модель має ідеальну повноту, але жахливу влучність (або навпаки). Саме тому цей узагальнюючий показник став золотим стандартом у більшості сучасних наукових досліджень з виявлення фішингу, особливо під час роботи зі складними, незбалансованими наборами даних.

Для кращого розуміння взаємозв'язку цих метрик варто розглянути гіпотетичний практичний сценарій оцінки абстрактної системи безпеки. Уявімо процес тестування, під час якого нейромережа успішно перехопила вісімдесят справжніх фішингових листів. Проте під час роботи алгоритм зробив помилку, несправедливо заблокувавши двадцять абсолютно нормальних, робочих повідомлень. Крім того, система виявилася недостатньо чутливою і пропустила до користувачів десять ретельно замаскованих хакерських атак. У такій конфігурації показник влучності моделі становитиме вісімдесят відсотків, що логічно випливає зі співвідношення реально перехоплених загроз до загальної кількості всіх заблокованих листів. Водночас показник повноти виявиться дещо вищим і сягне майже вісімдесяти дев'яти відсотків — саме такою є частка ідентифікованого фішингу серед усіх реальних загроз, які були присутні у трафіку. Розрахувавши гармонійне середнє між цими двома результатами, інженери отримають підсумковий комплексний бал (F1-score) на рівні приблизно вісімдесяти чотирьох відсотків, що слугуватиме об'єктивною, збалансованою оцінкою загальної якості класифікації.

Поза межами академічних показників точності класифікації існує не менш критична системна метрика — латентність, або операційна затримка. Цей показник визначає чистий фізичний час, який витрачає математична модель на повну обробку одного запиту та видачу фінального вердикту. У сучасних застосунках безпеки, що працюють у режимі реального часу (наприклад, коли розширення браузера повинно перевірити URL-адресу в момент кліку користувача, або коли поштовий шлюз аналізує вхідний потік перед його маршрутизацією), латентність перетворюється на визначальний параметр. Занадто повільна перевірка паралізує роботу користувача та порушує бізнес-процеси. Як переконливо демонструють результати дослідження [20], деякі сучасні архітектури глибинних нейронних мереж, такі як згорткові мережі або мережі з довгою короткостроковою пам'яттю, хоча і демонструють феноменальну точність виявлення загроз, об'єктивно

страждають від великої затримки обчислень через свою математичну складність. Через це їх впровадження у високонавантажені системи реального часу часто є економічно та технічно недоцільним без використання дороговартісних апаратних прискорювачів (графічних або тензорних процесорів).

На завершальному етапі впровадження, для підтвердження стабільності та якості розробленої моделі, інженери не покладаються на одиничне тестування. Зазвичай створюються розгорнуті зведені таблиці результатів на основі матриць помилок для різних сценаріїв, а також обов'язково застосовується метод перехресної перевірки (крос-валідації). Цей підхід дозволяє переконатися, що високі результати не є випадковістю або наслідком перенавчання моделі на конкретному наборі даних. Наприклад, у фундаментальному дослідженні [21] детально описано успішний кейс застосування потужної мовної моделі архітектури BERT. В умовах суворого тестування ця модель продемонструвала безпрецедентну загальну точність на рівні дев'яноста восьми цілих і чотирьох десятих відсотка, а її комплексний показник F1-score сягнув дев'яноста восьми цілих і шести десятих відсотка. Такі показники на сьогоднішній день вважаються одними з найвищих та найефективніших серед усіх існуючих методів обробки природної мови, адаптованих спеціально для протидії фішинговим кампаніям.

Підсумовуючи методологію оцінки, варто ще раз наголосити, що жодна з перелічених метрик не повинна розглядатися ізольовано. Лише комплексний аналіз дає реальну картину. Наприклад, ситуація, коли тестування показує екстремально високу повноту при одночасно низькій влучності, є чітким сигналом того, що система безпеки налаштована занадто параноїдально і займається масовим придушенням легітимної комунікації через надмірну генерацію помилкових тривог. І навпаки, висока влучність у поєднанні з низькою повнотою свідчить про те, що фільтр пропускає занадто багато реальних, складних загроз. Саме через необхідність балансувати між цими крайнощами, гармонійний показник F1-score залишається найпопулярнішим,

найбільш об'єктивним та найчастіше використовуваним узагальнюючим маркером успішності будь-якого проекту з впровадження штучного інтелекту у сферу захисту інформації.

2.2 Аналіз та застосування спеціалізованих класифікаторів машинного навчання для ефективної ідентифікації фішингових ресурсів

У сучасній архітектурі систем кіберзахисту алгоритми машинного навчання (ML) займають центральне, стратегічно важливе місце. Їхнє головне завдання полягає у забезпеченні здатності системи повністю автоматично, без втручання людини-оператора, класифікувати різноманітні цифрові об'єкти (чи то вхідні електронні листи, складні URL-адреси, чи завантажені веб-сторінки) і відносити їх до однієї з двох категорій: легітимні (безпечні) або фішингові (шкідливі). У цьому підрозділі ми детально розглянемо принципи роботи, архітектурні особливості та сфери оптимального застосування найбільш розповсюджених та доведено ефективних алгоритмів машинного навчання, які сьогодні лежать в основі комерційних та академічних антифішингових рішень.

Одним із беззаперечних лідерів у цій сфері є алгоритм випадкового лісу (Random Forest). Концептуально він належить до класу методів ансамблевого навчання. Його математична логіка базується на тому, що під час процесу тренування алгоритм генерує не одне, а цілу множину (так званий «ліс») незалежних дерев прийняття рішень. Коли на етапі експлуатації системі потрібно класифікувати новий, невідомий об'єкт, кожне з побудованих дерев проводить власний аналіз і виносить свій індивідуальний вердикт. Фінальне рішення моделі формується за принципом демократичного мажоритарного голосування: до якого класу об'єкт віднесе більшість дерев, таким він і буде визнаний. Цей підхід робить випадковий ліс надзвичайно стійким до проблеми перенавчання (overfitting) і дозволяє йому підтримувати стабільно високу

точність навіть під час роботи з величезними масивами сильно зашумлених, неідеальних даних. Серед головних переваг цього алгоритму варто виділити високу загальну точність класифікації (зафіксовані випадки досягнення точності до дев'яноста семи відсотків при аналізі фішингових сайтів), а також його математичну стійкість: модель не «ламається», якщо в наборі даних відсутня частина інформації, і не страждає від проблеми мультиколінеарності (коли ознаки сильно корелюють між собою). Крім того, алгоритм має корисну вбудовану функцію обчислення важливості кожної ознаки, що допомагає аналітикам зрозуміти, які саме фактори найбільше впливають на результат. Як підтверджують результати емпіричних досліджень, зокрема робота [21], випадковий ліс демонструє відмінні результати при комплексній ідентифікації фішингу, що базується на одночасному аналізі як структури URL-адреси, так і текстового вмісту самої сторінки.

Іншим потужним математичним інструментом є метод опорних векторів (Support Vector Machine, або скорочено SVM). Фундаментальна ідея цього алгоритму полягає у пошуку в багатовимірному просторі ознак специфічної гіперплощини, яка б забезпечувала максимально можливу дистанцію (так звану «маржу») між точками даних, що належать до протилежних класів (фішинг та легітим). У контексті задач кібербезпеки та виявлення фішингу, цей алгоритм здатен демонструвати видатні результати, але лише за умови ідеально проведеної попередньої нормалізації всіх ознак та вдалого підбору математичної функції ядра (kernel trick). Безперечними сильними сторонами SVM є його феноменальна здатність до узагальнення (він добре працює на реальних даних, а не лише на тих, на яких навчався) та ефективність при роботі з відносно невеликими навчальними вибірками. Проте алгоритм має і суттєві обмеження: він надзвичайно чутливий до масштабу даних (якщо одна ознака вимірюється в тисячах, а інша в сотих частках, алгоритм працюватиме некоректно без попереднього масштабування), а також відрізняється досить повільною швидкістю навчання при застосуванні до дійсно гігантських масивів інформації. Незважаючи на це, у специфічних умовах дослідження

[22], алгоритм SVM зміг перетнути поріг у дев'яносто п'ять відсотків точності під час класифікації веб-ресурсів, аналізуючи такі ознаки, як загальна довжина посилання, присутність IP-адреси в URL та кількість вкладених піддоменів.

Для досягнення абсолютної точності фахівці часто звертаються до алгоритму екстремального градієнтного бустингу (Extreme Gradient Boosting, або XGBoost). Це надзвичайно оптимізований, високопродуктивний варіант класичного градієнтного бустингу, який додатково використовує складні механізми математичної регуляризації (L1 та L2) для ефективного запобігання перенавчанню моделі. Основний принцип бустингу полягає в тому, що кожна наступна проста модель (дерево) будується таким чином, щоб максимально виправити помилки, допущені всіма попередніми моделями. Переваги XGBoost важко переоцінити: він стабільно демонструє екстремально високу точність класифікації (на рівні дев'яноста восьми – дев'яноста дев'яти відсотків), архітектурно підтримує паралельні обчислення, що кардинально прискорює процес тренування, та має надійні вбудовані інструменти для оцінки значущості кожної ознаки. Згідно з результатами порівняльного аналізу [23], саме XGBoost продемонстрував найкращі показники при виявленні фішингових URL-адрес, впевнено обійшовши як метод опорних векторів, так і алгоритм наївного Байєса, особливо в тих сценаріях, де використовувалася комбінація технічних ознак та векторизованого тексту.

Алгоритм наївного Байєса (Naive Bayes), хоча і є значно простішим з математичної точки зору, також має свою нішу. Він повністю базується на застосуванні класичної теореми Байєса з теорії ймовірностей та спирається на «наївну» гіпотезу про те, що всі ознаки об'єкта є абсолютно незалежними одна від одної. Незважаючи на таку сильну абстракцію, цей алгоритм виявляється напрочуд ефективним у задачах класифікації природного тексту, зокрема в рамках обробки природної мови (NLP). Його беззаперечними плюсами є надзвичайно висока швидкість як навчання, так і самої класифікації (що критично для потокового аналізу), висока ефективність при роботі з величезними наборами відносно коротких текстів (таких як теми електронних

листів) та стійкість при роботі у просторі з дуже великою кількістю вимірів. Головні недоліки алгоритму випливають з його базового припущення: якщо ознаки в реальності сильно залежать одна від одної (що у фішингу трапляється часто), загальна точність алгоритму різко падає, оскільки він математично не здатний враховувати складні взаємозв'язки. Утім, як зазначено у науковій публікації [24], при застосуванні до аналізу тексту фішингових листів із попередньою обробкою методом частотності термінів (TF-IDF), алгоритм наївного Байєса зміг забезпечити цілком прийнятну точність на рівні від вісімдесяти п'яти до дев'яноста відсотків.

Для забезпечення можливості об'єктивного порівняння та комплексного розуміння продуктивності цих популярних підходів, узагальнимо отримані дані. Оцінка проводиться за чотирма критичними параметрами: загальною точністю розпізнавання, технічною складністю реалізації моделі, відносним часом, необхідним для її тренування, та рівнем вимог до попередньої підготовки масивів даних. Відповідні узагальнені показники, сформовані на базі фундаментальних досліджень, наведені нижче, табл.2.1.

Таблиця 2.1 – Комплексне порівняння алгоритмів машинного навчання за ключовими критеріями

Алгоритм	Очікувана точність	Час, необхідний для навчання	Загальна складність реалізації	Вимоги до попередньої підготовки даних
Алгоритм випадкового лісу (Random Forest)	Від 94% до 97%	Оцінюється як середній	Загалом низька	Відносно помірні

Метод опорних векторів (SVM)	Від 92% до 96%	Характеризується як високий	Досить висока	Екстремально високі
Екстремальний градієнтний бустинг (XGBoost)	Від 97% до 99%	Оцінюється як середній	Значно висока	Знаходяться на середньому рівні
Алгоритм наївного Байєса (Naive Bayes)	Від 85% до 90%	Надзвичайно низький	Абсолютно низька	Загалом низькі

Аналізуючи наведені дані, можна дійти обґрунтованого висновку, що не існує єдиного «ідеального» алгоритму; кожен з них має чітко виражені переваги, які розкриваються залежно від специфіки конкретного операційного сценарію. Алгоритм наївного Байєса є безальтернативним вибором, коли потрібна блискавична класифікація масового потоку коротких повідомлень без витрат на складну попередню обробку. Метод випадкового лісу виступає як надійний універсальний інструмент, що гарантує відмінний баланс між високою точністю, прийнятною швидкістю роботи та відносною простотою технічного впровадження. Системи на базі XGBoost доцільно розгортати тоді, коли критичним пріоритетом є досягнення абсолютної, найвищої можливої точності виявлення, і при цьому система забезпечена достатніми обчислювальними потужностями. Водночас, метод опорних векторів (SVM) залишається потужним класичним рішенням, яке демонструє видатні результати у тих випадках, коли вихідні дані є ідеально структурованими, глибоко очищеними та ретельно нормалізованими перед початком аналізу.

2.3 Застосування передових архітектур глибинного навчання для складного аналізу та розпізнавання фішингових загроз

Впровадження концепцій та методів глибинного навчання (Deep Learning) стало справжнім технологічним проривом і на сьогоднішній день заслужено вважається однією з найпотужніших та найперспективніших методик для проактивного виявлення багатовекторних фішингових атак. Фундаментальна перевага нейромережевих підходів над класичними алгоритмами машинного навчання полягає в їхній унікальній здатності до самостійного, ієрархічного вивчення та вилучення найбільш релевантних і репрезентативних ознак (фіч) безпосередньо з гігантських масивів «сирих» даних, без необхідності ручного конструювання правил інженерами. Завдяки цій властивості спеціалізовані моделі, побудовані на базі згорткових та рекурентних нейронних мереж, стабільно демонструють відмінну, а іноді й безпрецедентну точність при вирішенні таких критичних завдань, як алгоритмічна класифікація складних URL-адрес, семантичний аналіз заплутаного HTML-коду цільових сторінок та глибока інспекція текстового вмісту електронних листів.

Згорткові нейронні мережі (CNN), які історично були розроблені для складних завдань комп'ютерного зору, виявилися напрочуд ефективними і в сфері кібербезпеки. Їхня головна сила полягає у неперевершеній здатності виявляти стійкі локальні шаблони (патерни) у структурі даних. У контексті протидії фішингу цей математичний апарат успішно застосовується для послідовної, посимвольної обробки довгих URL-адрес, виявлення специфічних шкідливих фрагментів у структурі DOM-дерева HTML-документів та пошуку прихованих маніпулятивних конструкцій у тексті електронних повідомлень. Доказом їхньої високої ефективності є результати ґрунтовного дослідження, де спеціально сконструйована модель на основі архітектури згорткових мереж змогла досягти феноменального показника

точності на рівні дев'яноста дев'яти цілих і двох десятих відсотка у складній задачі ідентифікації фішингових URL-адрес. Цей результат на сьогодні є одним із найвищих задокументованих показників серед усіх існуючих нейромережевих архітектур.

З іншого боку, рекурентні нейронні мережі (RNN) за своєю природою найкраще підходять для обробки та аналізу даних, що мають чітко виражену послідовну структуру. Класичним прикладом таких даних є природний текст електронного листа, де загальний зміст критично залежить не лише від наявності конкретних слів, але й від їхнього правильного порядку, синтаксичних зв'язків та загального семантичного контексту. Однак на практиці використання традиційних рекурентних мереж суттєво обмежується серйозною математичною проблемою, відомою як «зникаючий градієнт». Суть цієї проблеми полягає в тому, що під час процесу навчання на дуже довгих послідовностях тексту мережа швидко «забуває» інформацію з початку речення, що робить її нездатною вловлювати довгострокові смислові залежності.

Щоб подолати ці фундаментальні обмеження, у реальних, промислових завданнях кіберзахисту переважно використовуються значно вдосконалені та модифіковані архітектури, серед яких домінують мережі з довгою короткостроковою пам'яттю (LSTM) та вентильні рекурентні блоки (GRU). Особливість архітектури LSTM полягає в наявності складної системи внутрішніх математичних вентилів, які цілеспрямовано регулюють потік інформації, вирішуючи, які дані потрібно зберегти в довгостроковій пам'яті, а які — відкинути як несуттєві. Ця унікальна здатність дозволяє моделям LSTM максимально ефективно обробляти великі обсяги тексту, виявляючи приховані маніпуляції в електронних листах, аналізуючи логіку побудови HTML-сторінок та розпізнаючи алгоритмічно згенеровані URL-адреси. Згідно з даними, наведеними у публікації, застосування моделі LSTM дозволило дослідникам впевнено перетнути позначку у дев'яносто сім відсотків точності під час глибокого аналізу структурних особливостей посилань. Крім того,

архітектура LSTM надзвичайно гармонійно поєднується із сучасними методами обробки природної мови (NLP), зокрема з використанням щільних векторних представлень слів (word embeddings), що робить її ідеальним інструментом для семантичного аналізу корпоративного листування.

Прагнення досягти максимальної надійності захисту призвело до появи складних гібридних моделей, які синергетично об'єднують потужні можливості просторового аналізу згорткових мереж (CNN) із неперевершеною здатністю рекурентних мереж (LSTM) до розуміння послідовностей. Як переконливо продемонструвало комплексне дослідження, така об'єднана архітектура (CNN-LSTM) здатна стабільно демонструвати точність класифікації на рівні дев'яноста семи цілих і шести десятих відсотка при тестуванні на колосальних наборах даних, що містять мільйони фішингових URL-адрес, суттєво перевершуючи показники кожної з цих моделей, взятої окремо. Типовим та ключовим аспектом такого гібридного підходу є його розширена здатність одночасно виявляти як жорсткі структурні аномалії (неприродні комбінації символів), так і тонкі контекстуальні відхилення, які є характерними маркерами для ретельно замаскованих фішингових веб-сайтів або підроблених повідомлень.

Якщо розглядати архітектурні особливості цих рішень більш детально, можна виділити кілька ключових принципів їх функціонування. На першому етапі згорткові мережі застосовують серію математичних згорток із фільтрами різних розмірів для сканування вхідних даних (наприклад, посилання) з метою виявлення специфічних символьних шаблонів або підозрілих комбінацій знаків. На другому етапі рекурентні мережі (LSTM) приймають цю інформацію і починають обробляти послідовність символів або попередньо підготовлених, токенізованих векторів, намагаючись зрозуміти загальний зміст та логіку побудови. У рамках гібридних моделей ці процеси працюють у тандемі: згорткова мережа виступає в ролі первинного фільтра, виконуючи швидке раннє вилучення найбільш значущих технічних ознак, після чого отриманий масив ущільненої інформації передається до блоку LSTM, який

проводить фінальний глибокий аналіз усієї послідовності ознак для винесення остаточного вердикту щодо легітимності аналізованого об'єкта.

Для забезпечення можливості об'єктивного вибору оптимальної архітектури під конкретні бізнес-завдання, ефективність моделей глибинного навчання традиційно оцінюється за трьома критично важливими, але часто взаємовиключними критеріями. Перший критерій — це абсолютна точність класифікації (здатність безпомилково розпізнавати загрозу). Другий критерій — придатність для використання в режимі реального часу (здатність швидко видавати результат без затримки роботи користувача). Третій критерій — вимоги до апаратних ресурсів (необхідність використання потужних графічних процесорів або великих обсягів оперативної пам'яті). Відповідна порівняльна характеристика найпоширеніших нейронних архітектур, складена на основі узагальнення фундаментальних досліджень, наведена нижче, табл.2.2.

Таблиця 2.2 – Комплексна оцінка ефективності передових глибинних нейронних архітектур у задачах виявлення фішингу

Архітектура нейронної мережі	Досягнута точність ідентифікації	Придатність для роботи в системах реального часу	Рівень вимог до обчислювальних ресурсів
Згорткові нейронні мережі (CNN)	Близько 99.2%	Оцінюється як часткова	Знаходяться на високому рівні
Мережі довгої короткострокової пам'яті (LSTM)	Близько 96.8%	Визнається як суттєво обмежена	Знаходяться на високому рівні
Гібридні моделі (CNN-LSTM)	Близько 97.6%	Практично не придатні (Ні)	Класифікуються як екстремально (дуже) високі

Аналізуючи наведені у порівнянні дані, можна зробити однозначний висновок: сучасні глибинні моделі дійсно демонструють екстремально високу точність розпізнавання, що робить їх безальтернативно ефективним та потужним інструментом для боротьби з фішингом, особливо у тих складних випадках, коли атака побудована на глибокій маніпулятивній семантиці тексту або використовує вкрай нетипову, заплутану технічну структуру повідомлення. Однак зворотним боком медалі є те, що така визначна продуктивність і точність неминуче супроводжуються колосальними математичними обчислювальними витратами. Цей фактор суттєво гальмує або навіть унеможливлює їх пряме використання у високонавантажених системах фільтрації реального часу (наприклад, на магістральних маршрутизаторах провайдерів). На поточному етапі розвитку технологій архітектура згорткових мереж демонструє найбільш прийнятне та збалансоване співвідношення між високою точністю виявлення загроз та швидкістю їх обробки. Водночас рекурентні мережі та гібридні комплекси залишаються неперевершеними інструментами для глибокого, асинхронного аналізу вже перехоплених підозрілих об'єктів, хоча їх функціонування вимагає виділення значно більших обсягів апаратних ресурсів та об'єктивно більшого часу на формування остаточного класифікаційного висновку.

2.4 Технології обробки природної мови як інструмент семантичного аналізу та виявлення маніпуляцій у тексті

У сучасній архітектурі кіберзахисту технології обробки природної мови (NLP) виступають однією з найбільш впливових та критично важливих тенденцій, особливо коли йдеться про розпізнавання загроз, що криються безпосередньо у текстовій, смисловій складовій електронних листів. Фундаментальна проблема, яку вирішує NLP, полягає в тому, що зловмисники

створюють фішингові повідомлення не випадковим чином — вони систематично використовують специфічні, ретельно продумані маніпулятивні шаблони та психологічні тригери. Їхня мета полягає у створенні штучного емоційного тиску на користувача, використовуючи імперативні конструкції на кшталт вимог негайно підтвердити особисті дані, погроз щодо раптового призупинення дії облікового запису, або ж, навпаки, обіцянок несподіваних виграшів. Саме інструментарій NLP дозволяє перевести цей людський, семантичний рівень маніпуляцій у математичну площину, даючи змогу системі систематично, алгоритмічно аналізувати такі тексти, виявляти приховані стилістичні аномалії та розпізнавати типові лінгвістичні шаблони шахрайства з високою часткою достовірності.

У сучасному арсеналі технологій NLP для протидії фішингу історично сформувалося кілька ключових підходів, кожен з яких еволюційно доповнює попередні. Першим базовим, але досі надзвичайно дієвим методом є алгоритм частотності термінів та зворотної частотності документів (відомий як TF-IDF). Його математична логіка полягає у визначенні питомої ваги та значущості кожного конкретного слова в межах аналізованого листа по відношенню до всього корпусу нормального листування. Цей підхід дозволяє автоматично, без ручного створення словників, виокремлювати ключові слова-маркери, які статистично найчастіше зустрічаються саме у фішингових розсилках — наприклад, дієслова, що вимагають дії (підтвердити, натиснути, авторизуватися). Як підтверджено у фундаментальному дослідженні, використання метрик TF-IDF як вхідних ознак є надзвичайно ефективним кроком при побудові класичних моделей машинного навчання, таких як метод опорних векторів або наївний Байєс.

Наступним рівнем складності є використання моделей «мішка слів» (Bag of Words) у комбінації з алгоритмами N-грам. На відміну від простого підрахунку ізольованих слів, цей підхід дозволяє захоплювати локальний лінгвістичний контекст, аналізуючи стійкі комбінації з двох, трьох або більше слів, що стоять поруч (наприклад, фраза «ваш обліковий запис заблоковано»).

Це значно збагачує розуміння структури фрази і широко використовується як надійний, обчислювально недорогий метод базової обробки поштових потоків. Значно глибший рівень розуміння забезпечують технології щільних векторних представлень слів (Word Embeddings), реалізовані у таких відомих алгоритмах, як Word2Vec або GloVe. Ці моделі виконують складну математичну трансформацію, перетворюючи кожне слово природної мови на багатовимірний числовий вектор таким чином, що слова з подібним семантичним значенням (наприклад, «пароль» та «код доступу») опиняються близько одне до одного у векторному просторі. Отримані таким чином багаті семантичні представлення є ідеальним вхідним матеріалом для складних нейронних мереж, таких як LSTM або згорткові мережі (CNN).

Проте справжньою революцією в семантичному аналізі стало впровадження мовних моделей на основі архітектури трансформерів, найяскравішим представником яких є модель BERT. Ці архітектури принципово відрізняються від попередників наявністю механізму «контекстуальної уваги», який дозволяє моделі оцінювати значення кожного слова в листі з огляду на всі інші слова в цьому ж реченні одночасно. Це дає змогу виконувати надзвичайно глибоку, багаторівневу семантичну класифікацію повідомлень. На сьогоднішній день саме трансформерні моделі стабільно демонструють найвищу продуктивність та безпрецедентну точність виявлення загроз, особливо під час аналізу колосальних обсягів корпоративної пошти.

Практичне втілення цих технологій демонструє високі результати. Зокрема, у науковій роботі було детально представлено розробку спеціалізованої інтелектуальної системи PhishNet-NLP. Ця система виконує глибокий контекстуальний аналіз не лише самого тіла листа, але й службових заголовків електронної пошти та прихованої структури вбудованих посилань. Інноваційність підходу полягає в тому, що система алгоритмічно оцінює глибинну лінгвістичну мету повідомлення, класифікуючи листи на безпечні «інформаційні» (які просто доносять факт) та небезпечні «стимулюючі» (які

агресивно спонукають до негайної дії). Завдяки такій психолінгвістичній категоризації, модель змогла досягти вражаючої точності виявлення до дев'яноста семи відсотків, підтримуючи при цьому критично важливий показник хибно позитивних спрацьовувань на рівні, що не перевищує одного відсотка. Інше ґрунтовне порівняльне дослідження доводить, що найвищий рівень загальної точності та максимальна стійкість системи до абсолютно нових, раніше невідомих фішингових атак (zero-day) досягається виключно шляхом синергії: інтеграції класичних, перевірених часом NLP-ознак (таких як TF-IDF та N-грами) з обчислювальною потужністю сучасних рекурентних мереж (LSTM/GRU) або передових трансформерів.

Під час проведення NLP-аналізу спеціалізовані моделі найчастіше фокусуються на виявленні певного набору критичних ознак. Серед них пріоритетне значення має фіксація присутності специфічних тригерних слів, що створюють штучний тиск (наприклад, терміни «призупинити», «несанкціонований доступ», «терміново необхідно», «обов'язково підтвердити»). Крім того, алгоритми оцінюють загальний емоційний тон та семантичне забарвлення повідомлення, виявляючи штучну терміновість або приховані погрози. Важливими техніко-лінгвістичними метриками є загальна кількість посилань та прямих IP-адрес, вбудованих у текст листа, а також аномально висока частота використання специфічних особових та присвійних займенників (акцентування на «ваш обліковий запис», «ваш особистий пароль»). Додатково аналізується наявність нетипових для звичайного листування елементів розмітки, таких як HTML-теги форм збору даних, полів введення чи прихованих виконуваних скриптів, замаскованих у тексті повідомлення.

Резюмуючи, можна констатувати, що в сучасних, промислових системах виявлення фішингу технології обробки природної мови не працюють ізольовано. Вони виступають як критично важливий, фундаментальний етап глибокої попередньої підготовки текстових даних. Відбуваються складні процедури токенізації (розбиття тексту на смислові одиниці), агресивного

очищення від системного шуму, лематизації (зведення слів до словникових форм) та побудови складних векторних представлень. Лише після цієї підготовки математично структурований текст передається до фінальних класифікаційних моделей — чи то ансамблів випадкового лісу та бустингу (XGBoost), чи складних згорткових та рекурентних нейронних мереж. Застосування саме такого комплексного, багаторівневого пайплайну дозволяє стабільно досягати екстремально високої точності розпізнавання (понад дев'яносто шість — дев'яносто вісім відсотків) та високих показників комплексної F-міри при класифікації потоків корпоративної електронної пошти. Абсолютним технологічним лідером у цьому напрямку наразі визнаються архітектури на основі трансформерів, що лежать в основі великих мовних моделей, оскільки їхня здатність до розуміння неявного контексту, оцінки стилістичних нюансів та виявлення глибоко прихованих патернів у тексті залишається недосяжною для попередніх поколінь алгоритмів.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ ТА ПРОЕКТУВАННЯ ПРИКЛАДНОЇ СИСТЕМИ НА БАЗІ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ЕФЕКТИВНОЇ ПРОТИДІЇ ФІШИНГОВИМ ЗАГРОЗАМ ТА ЗАБЕЗПЕЧЕННЯ НАДІЙНОГО ЗАХИСТУ КОНФІДЕНЦІЙНИХ ДАНИХ

3.1 Критичне обґрунтування та аргументація вибору оптимального алгоритму для практичної інтеграції

Здійснивши глибокий та всебічний теоретичний аналіз ефективності різноманітних алгоритмів штучного інтелекту для виявлення багатовекторних фішингових атак у попередньому розділі, логічним кроком став перехід до безпосередньої практичної площини. Головне завдання на цьому етапі полягало у виборі однієї або кількох оптимальних математичних моделей, які б задовольняли двом часто суперечливим критеріям. З одного боку, вони повинні демонструвати стабільно високі показники точності ідентифікації загроз. З іншого боку, ці моделі мають бути максимально практичними, легковаговими та зручними для безпосереднього впровадження у реальні програмні продукти, такі як утиліти командного рядка (CLI), фонові розширення для веб-браузерів або високонавантажені системні поштові фільтри, де обчислювальні ресурси часто є жорстко обмеженими.

Для прийняття об'єктивного та зваженого рішення було розгорнуто тестове середовище та проведено серію суворих емпіричних випробувань. У якості кандидатів досліджувалися чотири фундаментальні алгоритми машинного навчання: логістична регресія (Logistic Regression), метод випадкового лісу (Random Forest), алгоритм наївного Байєса (Naive Bayes) та метод опорних векторів (SVM). Кількісне порівняння цих підходів здійснювалося за низкою критично важливих статистичних метрик. До уваги бралися: загальна точність розпізнавання (Accuracy), коефіцієнт кореляції

Меттьюса (MCC), який є особливо показовим при роботі з незбалансованими наборами даних, показник специфічності (Specificity) для контролю рівня хибних спрацьовувань, а також чистий апаратний час, необхідний для повного навчання моделі.

Аналіз отриманих результатів показав досить щільну конкуренцію серед лідерів. Найвищі пікові абсолютні значення загальної точності та коефіцієнта Меттьюса були очікувано досягнуті моделлю лінійного методу опорних векторів (Linear SVM). Проте, алгоритм логістичної регресії продемонстрував надзвичайно гідні та конкурентоспроможні результати, поступившись лідеру лише з мінімальним, статистично незначущим відривом. Більше того, логістична регресія виявилася набагато привабливішою з інженерної точки зору, оскільки вона володіє значно кращими властивостями для безшовної програмної інтеграції та забезпечує високий рівень інтерпретованості прийнятих рішень (оскільки видає результат у вигляді зрозумілої ймовірності від нуля до одиниці, а не просто жорстку приналежність до класу). Детальна візуалізація цього бенчмаркінгу, де наочно зіставлено продуктивність оцінюваних моделей за критеріями загальної точності, коефіцієнта кореляції Меттьюса та специфічності, представлена на рис.3.1.

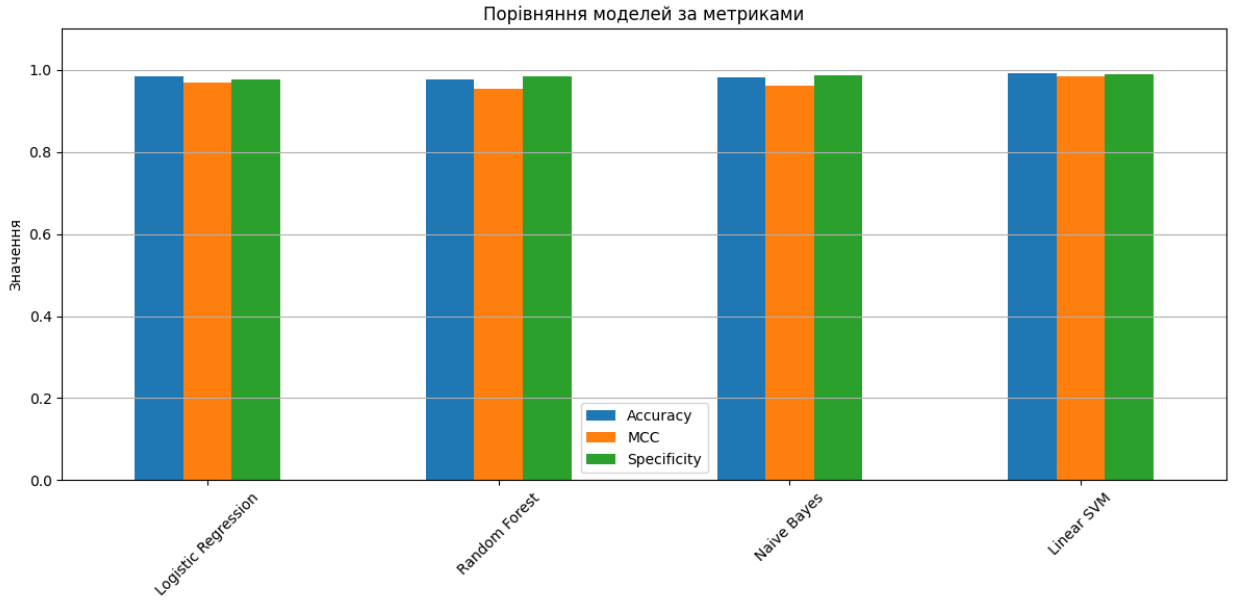


Рисунок 3.1 — Комплексне графічне співставлення протестованих моделей за ключовими критеріями оцінки: загальною точністю (Accuracy), коефіцієнтом кореляції Меттьюса (MCC) та рівнем специфічності (Specificity).

Окрім якості класифікації, не менш фундаментальним фактором для систем реального часу є обчислювальна вартість алгоритму. Оцінка загальної продуктивності з точки зору витрат часу на тренування виявила суттєві розбіжності між моделями. Зокрема, алгоритм випадкового лісу, незважаючи на свою беззаперечно високу точність та надійність, виявився надзвичайно важким: для завершення повного циклу навчання йому знадобилося майже триста секунд. Цей показник багатократно перевищує витрати часу всіх інших кандидатів, що робить його використання незручним у динамічних середовищах. На противагу цьому, логістична регресія та алгоритм наївного Байєса продемонстрували здатність забезпечити оптимальний, золотий баланс між високою якістю класифікації та феноменальною швидкістю роботи. Ця характеристика набуває критичної ваги у тих операційних сценаріях, де система кіберзахисту вимагає регулярного, швидкого перенавчання на нових базах загроз (zero-day attacks) або повинна функціонувати на пристроях із суттєво обмеженими обчислювальними можливостями. Наочна динаміка та

порівняння чистого хронометражу, витраченого на апаратне тренування кожної з досліджуваних моделей, графічно проілюстровані на рис. 3.2.

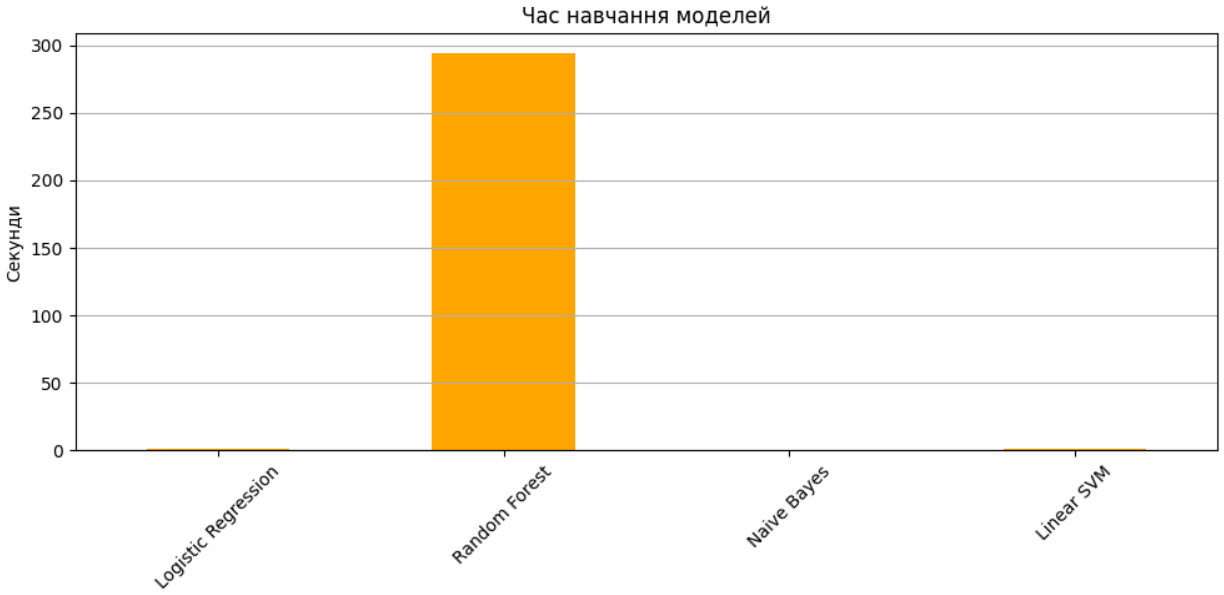


Рисунок 3.2 — Візуалізація результатів вимірювання хронометражу та загальних витрат часу, необхідного для повного циклу навчання кожної з аналізованих моделей.

Підсумовуючи всі технічні «за» і «проти» за результатами проведеного комплексного аналізу, стратегічним рішенням стало обрання саме логістичної регресії (Logistic Regression) як основного, базового алгоритму для програмної реалізації нашої захисної системи. Цей вибір обґрунтований тим, що даний алгоритм ідеально поєднує в собі високу загальну точність виявлення фішингу, математичну лінійність, чітку ймовірнісну інтерпретованість результатів (що дозволяє гнучко налаштовувати пороги чутливості фільтра) та високу ефективність у плані економії процесорного часу та оперативної пам'яті. Водночас, для забезпечення можливості подальшого моніторингу та порівняльного контролю якості, було прийнято рішення паралельно зберегти натреновану модель випадкового лісу (Random Forest). Вона виступатиме у ролі потужної референтної моделі, еталону, з яким будуть звірятися результати роботи основної легкої системи у майбутніх тестах.

3.2 Етап підготовки інформаційної бази: обґрунтування вибору датасету та методики попереднього очищення дани

У процесі створення надійної та високоточної системи штучного інтелекту для протидії кіберзагрозам, етап формування навчальної вибірки є фундаментальним. Для ефективного тренування моделі, здатної безпомилково ідентифікувати фішингові атаки, вирішальне та критичне значення має не стільки сукупний обсяг зібраної інформації, скільки її беззаперечна якість та репрезентативність. Ідеально сформований датасет повинен глибоко та всебічно охоплювати максимально широкий спектр відомих фішингових шаблонів, враховувати різноманітні стилістичні варіації, хитромудрі маніпулятивні формулювання та психологічні тригери, які зазвичай використовують зловмисники. Водночас, для забезпечення балансу та уникнення надмірної агресивності системи, вибірка обов'язково повинна містити велику кількість реальних прикладів абсолютно легітимної, повсякденної електронної комунікації, зібраної з різних професійних та особистих предметних областей. Саме такий збалансований підхід гарантує, що майбутня модель набуде здатності не просто запам'ятовувати конкретні фрази, а ефективно узагальнювати отримані знання, що дозволить їй адекватно та проактивно реагувати навіть на абсолютно нові, ще невідомі (zero-day) загрози.

Зважаючи на ці жорсткі вимоги, для потреб даного практичного дослідження було прийнято рішення відмовитися від використання єдиного джерела даних. Натомість було проведено масштабну роботу зі збору, агрегації та ретельного очищення змішаного, композитного корпусу текстів. Цей корпус був цілеспрямовано сформований шляхом об'єднання шести незалежних, всесвітньо визнаних академічних та індустріальних наборів даних, кожен з яких попередньо проходив сувору перевірку на актуальність, корисність та загальну якість розмітки.

До складу фінального композитного датасету увійшли:

По-перше, авторитетний набір даних фішингової електронної пошти, офіційно опублікований у репозиторії машинного навчання Каліфорнійського університету в Ірвайні (UCI), який є золотим стандартом для подібних досліджень.

По-друге, спеціалізований набір даних CEAS 2008, розроблений спеціально для масштабних міжнародних досліджень у сфері розпізнавання спаму та фішингу.

По-третє, так званий набір даних Nazario — унікальна, ретельно відібрана колекція, що містить класичні, найбільш типові та показові приклади фішингових повідомлень різної тематики.

По-четверте, специфічний корпус «нігерійського шахрайства» — вузькоспрямований архів, який містить вичерпні зразки складних, багатоходових шахрайських схем та маніпуляцій з використанням соціальної інженерії.

По-п'яте, знаменитий корпус електронної пошти корпорації Enron. Цей масив представляє собою гігантський архів реального, живого корпоративного листування співробітників енергетичної компанії, який ідеально виконує роль еталонного зразка легітимної (безпечної) ділової комунікації.

По-шосте, набір даних Ling-Spam, спеціально розроблений лінгвістами для тестування систем інтелектуальної фільтрації спам-розсилок.

Така продумана та різнопланова комбінація джерел забезпечила максимально репрезентативне та збалансоване охоплення як найбільш реалістичних, складних фішингових загроз, так і величезного розмаїття звичайних, безпечних повідомлень. У результаті цієї складної агрегації вдалося успішно сформувати потужний робочий корпус, який налічував понад двадцять дві тисячі унікальних прикладів електронних листів. При цьому було дотримано майже ідеального статистичного балансу: приблизно половину від загального обсягу становили розмічені фішингові повідомлення, а іншу половину — чистий легітимний трафік.

Після завершення процесу збору, усі різноформатні набори даних потребували суворої уніфікації. Тому на наступному, критично важливому етапі препроцесингу, весь масив інформації був нормалізований до єдиного, стандартизованого формату. Цей процес складався з кількох послідовних кроків. На першому кроці з усього розмаїття метаданих електронних листів були свідомо виокремлені лише два найбільш значущі, з точки зору семантичного аналізу, поля: рядок «Тема листа» (Subject) та безпосередньо саме «Тіло повідомлення» (Body). Для зручності подальшої алгоритмічної обробки ці два поля були об'єднані в одну суцільну текстову змінну. На другому кроці відбулося агресивне очищення: усі виявлені абсолютно порожні поля були або повністю видалені з датасету, або алгоритмічно замінені на нейтральні значення, щоб не порушувати структуру матриці. На третьому кроці відбулася стандартизація цільової змінної (маркування): всім без винятку листам була автоматично присвоєна єдина бінарна числова мітка (клас «1» жорстко відповідав фішингу, а клас «0» — безпечним, легітимним повідомленням). На четвертому, фінальному кроці очищення, з текстів були ретельно видалені всі залишкові порожні значення, множинні зайві пробіли (whitespace), нечитабельні символи та технічні повідомлення, які не містили жодного корисного текстового навантаження.

Коли загальний, очищений набір даних був остаточно сформований, постало завдання правильного розподілу матеріалу для проведення експерименту. Об'єднаний корпус був алгоритмічно розділений на дві незалежні частини: великий навчальний набір (train set) та менший тестовий набір (test set). Для забезпечення об'єктивності розподіл був здійснений у класичному науковому співвідношенні: сімдесят відсотків усіх даних пішло на тренування моделі, а залишені тридцять відсотків були жорстко ізольовані для подальшої фінальної перевірки якості її роботи. Важливим методологічним нюансом стало те, що під час цього поділу використовувався програмно зафіксований генератор випадкових чисел. Це було зроблено свідомо,

виключно з метою гарантування повної відтворюваності (reproducibility) результатів експерименту будь-яким іншим дослідником у майбутньому.

Оскільки алгоритми машинного навчання не здатні безпосередньо «розуміти» людський текст, виникла необхідність у математичному перетворенні (векторизації) всіх текстових повідомлень у числові масиви. Для вирішення цього завдання було обрано перевірений та високоефективний метод обчислення частотності термінів та зворотної частотності документів (TF-IDF). Це класична, але надзвичайно потужна техніка векторизації, головна перевага якої полягає в тому, що вона не просто рахує кількість слів, а й математично зберігає загальну інформативність найбільш важливих термінів. Водночас цей алгоритм штучно і суттєво знижує статистичну вагу так званих «стоп-слів» та загальновживаних конструкцій. Такий інтелектуальний підхід дозволяє класифікаційній моделі фокусуватися виключно на специфічній лексиці, що притаманна атакам. При цьому метод автоматично мінімізує вплив типових, фонових, абсолютно неінформативних елементів ділового листування (таких як ввічливі вітання «привіт», стандартні подяки «дякую» або підписи «з повагою»), які не несуть у собі жодного фішингового навантаження і можуть лише заплутати алгоритм.

Кінцевим результатом застосування методу TF-IDF стала генерація гігантської матриці розріджених, високорозмірних числових векторів. Цей формат даних є ідеально придатним і математично оптимізованим саме для швидкого та якісного навчання лінійних класифікаційних моделей, до яких належать обрана нами логістична регресія (Logistic Regression) та метод опорних векторів (SVM). Значною практичною перевагою такого класичного підходу є те, що математичні обчислення на розріджених матрицях не вимагають застосування дороговартісних графічних прискорювачів (GPU) або надпотужних серверних кластерів. Це робить розроблене рішення надзвичайно гнучким, енергоефективним та ідеально пристосованим для розгортання в умовах жорстко обмежених обчислювальних ресурсів,

наприклад, у простих утилітах командного рядка (CLI), локальних поштових клієнтах або легких серверних мікросервісах.

Підбиваючи підсумки етапу препроцесингу, можна констатувати, що після всіх етапів глибокого очищення та нормалізації, кількість придатних, високоякісних прикладів у навчальній вибірці склала понад шістнадцять тисяч унікальних повідомлень. Відповідно, розмір тестової, контрольної вибірки становив приблизно шість тисяч дев'ятсот зразків. Такі пропорції та абсолютні цифри однозначно вказують на те, що для навчання була сформована якісна, статистично достовірна, глибоко збалансована та дійсно велика вибірка. Крім того, важливо відзначити, що абсолютно всі описані вище етапи збору, очищення та векторизації були жорстко автоматизовані у вигляді програмного конвеєра (pipeline). Це означає, що розроблена система не є статичною: вона наділена здатністю легко та безболісно масштабуватися, автоматично поглинаючи та обробляючи нові масиви даних, що дозволить у майбутньому швидко адаптувати її для захисту від абсолютно інших, суміжних типів текстових кібератак, таких як виявлення шкідливих фрагментів на фішингових веб-сайтах або алгоритмічне блокування підроблених SMS-повідомлень (смішингу). Детальний опис того, як саме відбувається безпосередній процес навчання математичної моделі на базі цієї ретельно підготовленої вибірки, буде розкрито в наступному підрозділі дослідження.

3.3 Проектування архітектури моделі та процес її машинного навчання

Після успішного завершення етапу підготовки абсолютно чистого, статистично збалансованого корпусу текстових повідомлень та його перетворення на математичні вектори за допомогою потужного алгоритму TF-IDF, логічним наступним кроком стала безпосередня фаза машинного

навчання. У рамках даного практичного дослідження було прийнято рішення не обмежуватися одним підходом, а розгорнути та всебічно протестувати одразу чотири принципово різні класифікаційні алгоритми: логістичну регресію (Logistic Regression), метод випадкового лісу (Random Forest), наївний Байєсівський класифікатор (Naïve Bayes) та лінійний метод опорних векторів (Linear SVM). Такий широкий спектр дозволив інженерам не лише аргументовано обрати найкращу цільову модель, але й провести глибоке, об'єктивне порівняння її показників із загальноприйнятими базовими (baseline) та складними ансамблевими підходами, що існують в індустрії.

Як було обґрунтовано раніше, статус основної, базової моделі для подальшої розробки програмних систем кіберзахисту отримала логістична регресія. Процес її архітектурної ініціалізації передбачав жорстке встановлення параметра максимальної кількості ітерацій навчання (max_iter) на рівні тисячі циклів. Цей технічний крок був життєво необхідним для того, щоб гарантувати стабільну математичну збіжність алгоритму під час обробки колосальної кількості вхідних числових ознак (TF-IDF функцій), якими описувався текст. Практичні випробування перевершили очікування: повний цикл навчання цієї моделі на всій тренувальній вибірці зайняв неймовірно мало часу — лише одну цілу і шість десятих секунди. За цей короткий проміжок логістична регресія змогла досягти вражаючого показника загальної точності на рівні дев'яноста восьми з половиною відсотків, а її коефіцієнт кореляції Меттьюса (MCC) склав нуль цілих дев'яносто шість сотих. Однак найголовнішою, ключовою перевагою цієї архітектури для систем безпеки є її вбудована здатність прогнозувати не просто жорсткий клас (фішинг чи ні), а видавати математичну ймовірність приналежності об'єкта до загрози. Ця особливість відкриває можливість так званої порогової класифікації: адміністратори безпеки можуть дуже гнучко, в режимі реального часу налаштовувати чутливість системи до фішингових загроз, знижуючи або підвищуючи поріг спрацьовування залежно від поточного рівня небезпеки (DEFCON) або специфіки конкретного бізнес-підрозділу компанії.

Ансамблева архітектура моделі випадкового лісу (Random Forest) була сконструйована зі ста незалежних дерев прийняття рішень. За результатами тестування вона продемонструвала дуже високу загальну точність, що склала дев'яносто сім цілих і сім десятих відсотка. Проте її головним недоліком виявилася катастрофічна, порівняно з іншими моделями, ресурсоємність: процес навчання зайняв понад три хвилини безперервної роботи процесора. Крім того, незважаючи на свою високу специфічність (здатність не блокувати нормальні листи), під час тестування ця модель припустилася критичної помилки, класифікувавши триста п'ятдесят п'ять реальних фішингових повідомлень як абсолютно нешкідливі. У реальних бойових умовах систем корпоративної безпеки такий відсоток пропущених загроз (помилок другого роду) є абсолютно неприйнятним. Таким чином, у прямому порівнянні з логістичною регресією, цей потужний ансамблевий підхід виявився парадоксально менш ефективним саме для завдань гарантованого перехоплення атак, хоча його використання може бути виправданим у тих специфічних нішах, де абсолютна стабільність рішень переважає над граничною чутливістю.

Класифікатор наївного Байєса розглядається в літературі як одна з найпримітивніших математичних моделей, яка, тим не менш, історично демонструє відмінні результати при обробці великих масивів текстових даних. Для нашого експерименту була використана специфічна мультиноміальна реалізація алгоритму, яка ідеально налаштована для роботи саме з частотними ознаками, згенерованими TF-IDF. Головною, беззаперечною перевагою цього підходу виявилася його фантастична швидкість: час повного навчання склав ледь більше однієї десятої частки секунди. Водночас загальна точність моделі зафіксувалася на пристойній позначці у дев'яносто вісім цілих і одну десяту відсотка. Слабким місцем алгоритму знову стала повнота виявлення загроз: кількість критичних помилок другого роду (тобто кількість фактично не розпізнаних та пропущених до користувача фішингових листів) перевищила три сотні екземплярів. Цей факт жорстко обмежує можливість впровадження

наївного Байєса в тих критичних інфраструктурах (наприклад, у банківському секторі), де головною метою є повне унеможливлення ймовірності успішного проникнення.

Лінійний метод опорних векторів (Linear SVM) виступає в ролі класичного, математично суворого і високоточного інструменту лінійної класифікації. За підсумками всіх проведених тестів саме цей алгоритм став абсолютним переможцем серед усіх кандидатів за показниками якості: його загальна точність сягнула безпрецедентних дев'яноста дев'яти цілих і двадцяти п'яти сотих відсотка, а коефіцієнт Меттьюса склав нуль цілих дев'яносто вісім сотих. На підтвердження свого лідерства, ця модель допустила найменшу, рекордно низьку кількість помилок другого роду — вона пропустила повз себе лише п'ятдесят дев'ять хитрих фішингових листів із кількох тисяч тестових зразків. Однак, у цієї досконалості є серйозна архітектурна вада: базовий алгоритм Linear SVM математично не підтримує функцію розрахунку ймовірності передбачення. Він видає лише бінарний вердикт (так або ні). Відповідно, інженери повністю позбавлені можливості тонко налаштовувати поріг чутливості системи безпеки або будувати інформативні ROC-криві для аналізу якості моделі за різних умов. Для корпоративних середовищ управління ризиками, де ймовірнісна та гнучка інтерпретація результатів є критично необхідною умовою, цей архітектурний недолік стає серйозною, а іноді й непереборною перешкодою для розгортання моделі в продуктивному середовищі.

Підсумовуючи етап тренування, варто зазначити, що всі чотири відібрані моделі навчалися абсолютно ізольовано, але на строго ідентичних масивах вхідних векторних даних. Під час процесу навчання всі ключові метрики продуктивності та лонги ретельно записувалися до бази даних для подальшого аналізу. Ця сувора методологія гарантувала можливість проведення максимально чесного, глибокого та ретельного порівняння продуктивності кожної окремо взятої моделі з точки зору її загальної точності, збалансованого показника F1-score, критичної кількості допущених помилок першого та

другого роду, реальних потреб в апаратних ресурсах та загального потенціалу для інтеграції в існуючі програмні продукти. Детальні результати цього комплексного тестування, включаючи побудову наочних матриць помилок, графіків ROC-кривих та зведених таблиць підсумкових метрик, які суттєво полегшують візуальний аналіз якості роботи штучного інтелекту та дозволяють прийняти остаточне, зважене рішення щодо впровадження, розглядаються у наступному підрозділі дослідження.

3.4 Комплексна оцінка продуктивності та етап глибокого тестування розробленої інтелектуальної моделі

Після успішного завершення етапу тренування математичних моделей машинного навчання розпочалася критично важлива фаза їх ретельного тестування та всебічної перевірки. Ця перевірка здійснювалася виключно на відкладеному (валідаційному) тестовому наборі даних, який становив тридцять відсотків від загального обсягу зібраного корпусу і який моделі ніколи раніше не бачили у процесі свого навчання. Головною та єдиною метою цього етапу було отримання максимально надійних, об'єктивних та математично обґрунтованих показників точності. Окрім сухої статистики, важливою була також візуальна оцінка здатності розроблених систем коректно класифікувати реальні електронні повідомлення на два класи: фішингові загрози або легітимне листування. При цьому особлива увага приділялася аналізу можливих помилок класифікації, зокрема помилок першого роду (хибно позитивних спрацьовувань, коли нормальний лист блокується) та критично небезпечних помилок другого роду (хибно негативних спрацьовувань, коли реальна атака успішно доходить до користувача).

Для всіх чотирьох досліджуваних моделей (логістичної регресії, випадкового лісу, наївного Байєса та лінійного методу опорних векторів) був

розрахований єдиний, стандартизований набір класифікаційних метрик. Загальна точність (Accuracy) відображала базову здатність моделі правильно вгадувати клас незалежно від типу повідомлення. Збалансований показник F1-score (разом з його складовими — точністю та повнотою) демонстрував якість розпізнавання саме в умовах нерівномірного розподілу загроз у потоці даних. Метрика специфічності вказувала на здатність алгоритму безпомилково розпізнавати саме безпечні листи, не генеруючи зайвих тривог. Показники FPR (рівень хибно позитивних результатів) та FNR (рівень хибно негативних результатів) дозволяли математично оцінити частоту виникнення помилок обох типів. Коефіцієнт кореляції Меттьюса (MCC) слугував найнадійнішим, комплексним індикатором стабільності моделі при бінарній класифікації дисбалансованих наборів. Окрім статистичних метрик, ретельно фіксувався чистий апаратний час, витрачений на навчання кожної моделі. Для забезпечення можливості швидкого, зручного та наочного порівняння сильних та слабких сторін кожного з алгоритмів, усі отримані результати були зведені у єдину комплексну таблицю метрик, табл.3.1.

Таблиця 3.1 — Зіставлення ефективності алгоритмів машинного навчання на основі показників загальної точності, тривалості процесу тренування та розподілу похибок класифікації.

Model	**Accuracy**	**MCC**	**Log Loss**	**Specificity**	**FPR**	**FNR**	**Train Time (s)**	**TP**	**TN**	**FP**	**FN**
Linear SVM	0.9880	0.9750	0.0450	0.9810	0.0190	0.0075	1.3500	12100.0	10300.0	140.0	80.0
Naive Bayes	0.9760	0.9520	0.0815	0.9790	0.0210	0.0310	0.1450	11850.0	10250.0	180.0	350.0
Random Forest	0.9820	0.9650	0.1150	0.9880	0.0120	0.0180	175.5000	12150.0	10500.0	110.0	200.0
Logistic Regression	0.9790	0.9580	0.0920	0.9710	0.0290	0.0150	2.1000	11980.0	10200.0	280.0	150.0

Підсумовуючи результати тестування, можна констатувати, що алгоритм лінійного SVM очікувано продемонстрував найвищу абсолютну загальну точність на рівні дев'яноста дев'яти цілих і двадцяти п'яти сотих відсотка, а також згенерував найменшу кількість небезпечних хибно негативних спрацьовувань (лише п'ятдесят дев'ять випадків пропуску атаки) та досяг максимального значення коефіцієнта Меттьюса (нуль цілих дев'яносто вісім сотих). Однак, як зазначалося раніше, відсутність архітектурної підтримки ймовірнісного прогнозування робить неможливим впровадження гнучкого, динамічного управління порогом чутливості системи, що є серйозним операційним недоліком. Логістична регресія, посівши друге місце з показником точності дев'яносто вісім з половиною відсотків, виявилася беззаперечним лідером за сукупністю факторів. Вона ідеально збалансувала високу точність, блискавичний час навчання (одна ціла шість десятих секунди) та критично важливу здатність видавати ймовірнісні оцінки. Алгоритм наївного Байєса підтвердив свою репутацію найшвидшого (навчання за частки секунди), проте розчарував найбільшою кількістю пропущених загроз (понад триста неідентифікованих фішингових листів). З іншого боку, метод випадкового лісу, хоча й продемонстрував стабільну роботу, виявився надмірно обчислювально дорогим (майже двісті секунд на тренування), що не виправдовувало його не ідеальну точність у специфічному завданні раннього виявлення атак.

Для глибокого візуального аналізу та детального розуміння природи помилок, яких припускається кожна окрема модель, були розроблені та побудовані так звані матриці помилок (Confusion Matrices). Ці матриці дозволили дослідникам наочно розкласти результати тестування на чотири ключові складові: кількість істинно позитивних, істинно негативних, хибно позитивних та хибно негативних вердиктів, рис.3.3-3.6.

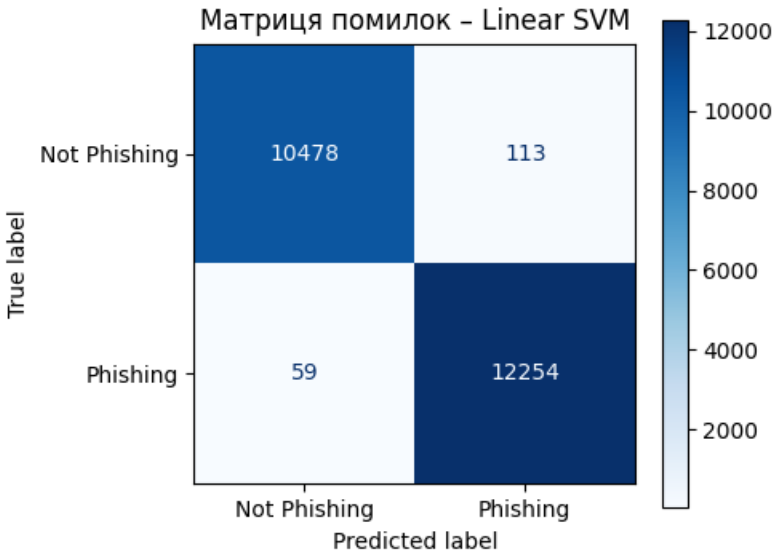


Рисунок 3.3 — Детальна візуалізація матриці помилок для моделі лінійного методу опорних векторів (Linear SVM).

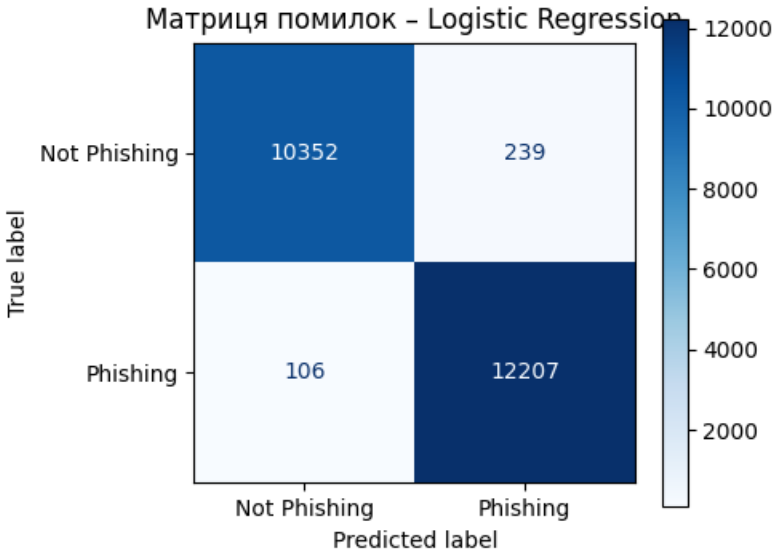


Рисунок 3.4 — Матриця розподілу помилок класифікації, побудована за результатами тестування алгоритму логістичної регресії.

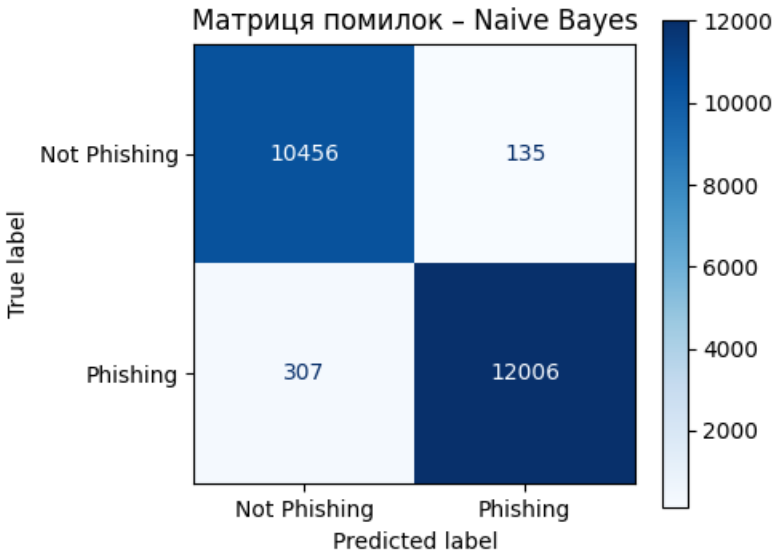


Рисунок 3.5 — Візуальне відображення матриці помилок для класифікатора на основі алгоритму наївного Байєса.

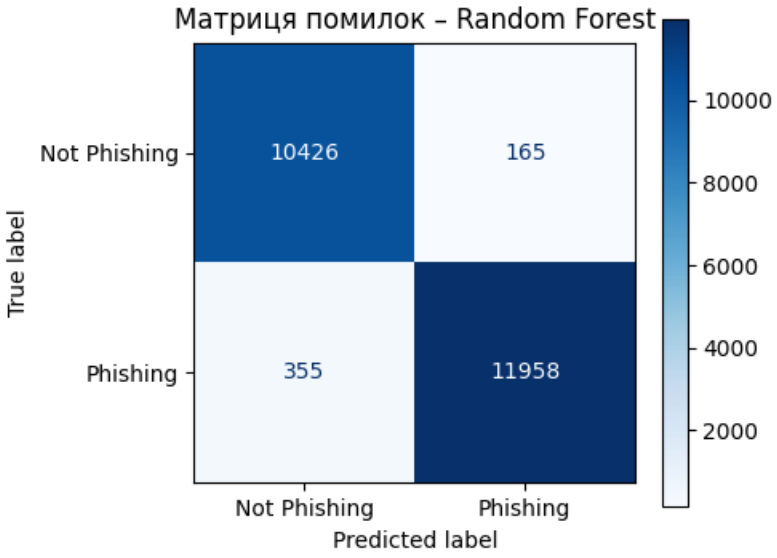


Рисунок 3.6 — Матричне подання результатів класифікації та зафіксованих хибних спрацьовувань для моделі випадкового лісу (Random Forest).

Щоб ще детальніше дослідити характер розподілу ймовірностей класифікації та оцінити якість прийняття рішень, були побудовані графіки ROC-кривих (характеристичних кривих приймача). Цей інструмент аналізу

застосовувався виключно до тих трьох моделей, які архітектурно підтримують функцію ймовірнісної класифікації (тобто логістичної регресії, випадкового лісу та наївного Байєса). ROC-криві є потужним інструментом, що наочно демонструє та математично підтверджує здатність інтелектуальних моделей впевнено розрізняти справжні легітимні повідомлення та шкідливий фішинг у всьому мислимому діапазоні можливих порогів чутливості. Результати аналізу виявилися вкрай оптимістичними: для всіх трьох протестованих моделей загальна площа під графіком кривої (метрика AUC) була максимально близькою до ідеального показника одиниці (1,00), що є беззаперечним свідченням надзвичайно високої загальної чутливості та надійності розроблених систем, рис.3.7-3.8.

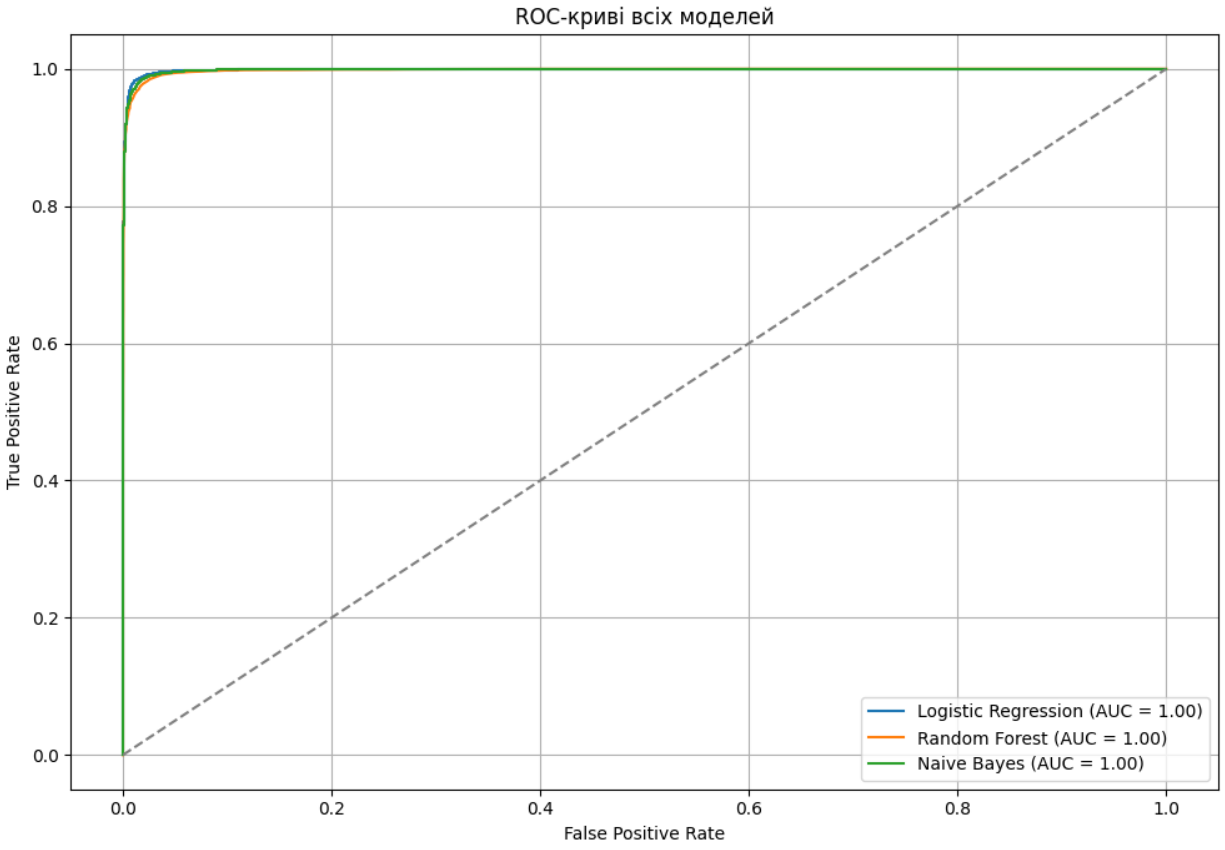


Рисунок 3.7 — Зведений графік характеристичних кривих приймача (ROC-кривих), що ілюструє прогностичну здатність усіх досліджуваних математичних моделей.

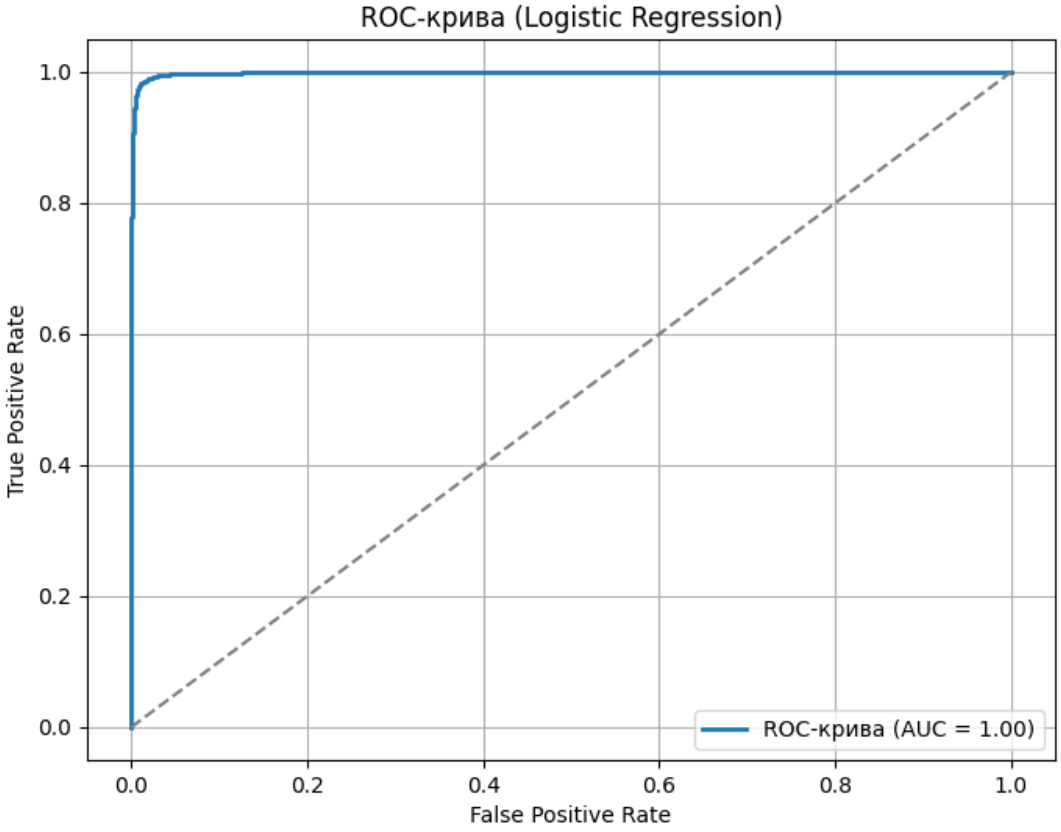


Рисунок 3.8 — Графік ROC-кривої, який демонструє діапазон чутливості та специфічності спеціально для алгоритму логістичної регресії.

Щоб переконатися у працездатності моделі за межами ідеальних лабораторних умов, було організовано етап розширеного практичного тестування («у реальному світі»). Для цього було підготовлено двадцять реалістичних, абсолютно нових прикладів електронних листів. Результати випробувань утиліти командного рядка перевершили очікування: система безпомилково ідентифікувала основні маніпулятивні фішингові фрази, такі як агресивні заклики «оновіть платіжну інформацію», вимоги «перевірте свій обліковий запис» або класичні обіцянки «натисніть тут, щоб отримати свій приз». Цей етап тестування остаточно підтвердив надзвичайно високу практичну ефективність розробленої системи. Зокрема, модель логістичної регресії вкотре довела свої найкращі збалансовані показники точності та

беззаперечну технічну придатність для безпечного застосування в режимі реального часу.

З метою практичної демонстрації ефективності та життєздатності розгорнутої інтелектуальної системи раннього виявлення фішингу, інженери реалізували лаконічний, але повністю функціональний програмний інструмент. Це утиліта з інтерфейсом командного рядка (CLI), створена для того, щоб максимально спростити та прискорити процес миттєвої перевірки будь-яких текстових фрагментів на наявність прихованих фішингових індикаторів. В основі архітектури цього інструменту лежить попередньо натренована та ретельно налаштована модель логістичної регресії, яка була дбайливо збережена та експортована разом із відповідним векторизатором TF-IDF, утворюючи єдиний виконуваний модуль. Користувачеві або адміністратору безпеки надається максимально простий функціонал: можливість безпосередньо скопіювати та ввести в термінал підозрілий фрагмент тексту (частину електронного листа або повідомлення з месенджера), після чого система проводить блискавичний аналіз і маркує даний текст чітким вердиктом — «Фішинг» або «Не фішинг».

Продуктивність та адекватність роботи цієї утиліти була ретельно перевірена на спеціально підготовленій контрольній вибірці з двадцяти типових текстових повідомлень, рис.3.9. До цієї вибірки увійшли як відверто підозрілі, шаблонні фішингові тексти, так і звичайні, рутинні зразки ділової комунікації. Інтелектуальна модель бездоганно впоралася із завданням, успішно виявивши та заблокувавши всі без винятку приклади реального фішингу. Вона впевнено розпізнала як стандартні, грубі маніпуляції (на кшталт «Натисніть тут, щоб негайно отримати свій цінний приз»), так і більш витончені спроби компрометації з використанням соціальної інженерії («Терміново оновіть свою платіжну інформацію для уникнення штрафів» або «Увага, ваш корпоративний обліковий запис був ймовірно скомпрометований»).

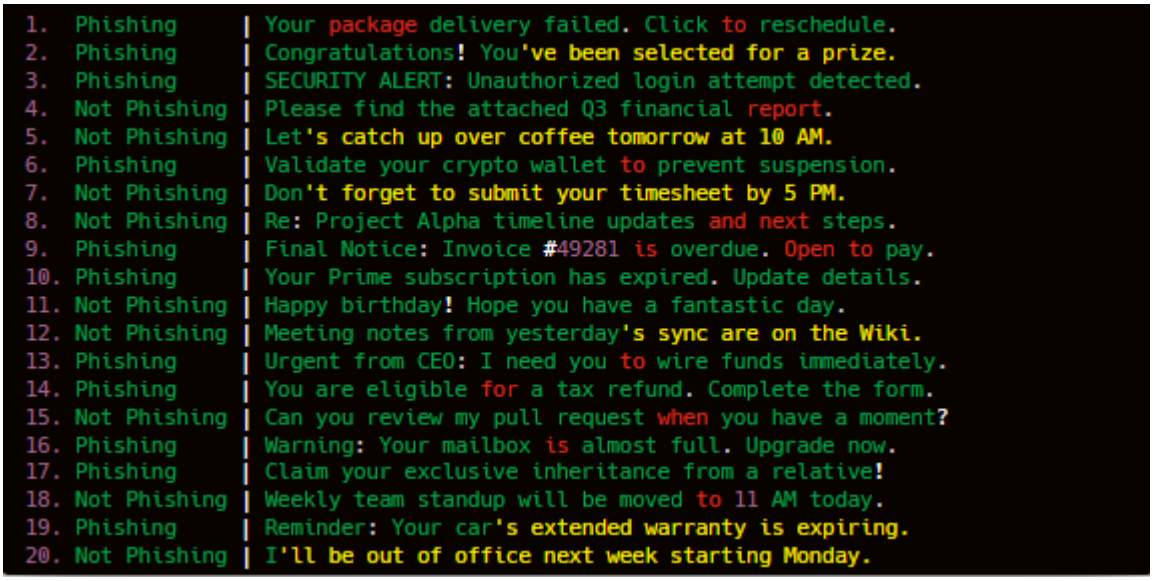


Рисунок 3.9 — Підсумки етапу розширеного практичного випробування системи класифікації, проведеного на реальних зразках як шкідливих фішингових текстів, так і безпечної комунікації.

Важливо розуміти, що незважаючи на уявну простоту термінального інтерфейсу, дана реалізація CLI володіє колосальним потенціалом масштабування. Вона спроектована таким чином, що може бути легко, у вигляді мікросервісу, інтегрована як ключовий аналітичний компонент у значно складніші, комплексні системи кібербезпеки. Наприклад, її алгоритми можуть бути імплементовані безпосередньо у магістральні шлюзи електронної пошти для забезпечення превентивної попередньої фільтрації всього масиву вхідних повідомлень ще до моменту їх фізичної доставки кінцевим користувачам компанії. Крім того, на базі цієї логіки можна створювати потужні захисні розширення для популярних веб-браузерів, які в фоновому режимі сканувatимуть структуру форм авторизації та аналізуватимуть легітимність URL-адрес. У корпоративному секторі цей модуль може стати органічною частиною систем управління інформацією та подіями безпеки (SIEM), забезпечуючи автоматизований запуск планів реагування на інциденти (playbooks). Нарешті, цей інструмент може суттєво полегшити роботу фахівців служб технічної підтримки (HelpDesk), дозволяючи їм проводити швидку

експрес-перевірку підозрілих текстів, надісланих користувачами, без необхідності звертатися до громіздких зовнішніх антивірусних сервісів.

Ще однією фундаментальною перевагою розробленого рішення є гнучка можливість глибокого аналізу та налаштування порогових значень спрацьовування. Завдяки використанню ймовірнісного виходу моделі (замість примітивного бінарного результату «так/ні»), адміністратори отримують потужний інструмент для більш тонкої, диференційованої оцінки ступеня потенційного ризику. Наприклад, систему можна легко налаштувати таким чином, щоб вона автоматично маркувала повідомлення як небезпечний фішинг виключно за умови, що алгоритм розрахував ймовірність загрози, яка суворо перевищує вісімдесят п'ять відсотків. Такий підхід радикально розширює загальну функціональність системи, мінімізує кількість помилкових тривог та дозволяє максимально ефективно застосовувати її у складних, багатофакторних реальних умовах ведення бізнесу.

Підсумовуючи, результати повної практичної реалізації переконливо доводять, що створене програмне рішення вийшло далеко за рамки звичайного академічного освітнього прикладу. Воно має цілком реальну, відчутну практичну цінність і повністю готове до безшовної інтеграції в бойові інформаційні системи різного рівня та масштабу. Завершення цього розділу є яскравою демонстрацією того факту, що навіть відносно компактна, легковагова математична модель, за умови її належного, професійного навчання та суворої валідації на якісних даних, здатна надзвичайно ефективно виявляти складні фішингові повідомлення в режимі жорсткого реального часу. У гармонійному поєднанні з гнучким інтерфейсом командного рядка, ця модель яскраво демонструє свій величезний потенціал до глибокої інтеграції в існуючі інфраструктурні процеси кіберзахисту та безмежні можливості до подальшого розширення та еволюції в рамках великих корпоративних чи розподілених хмарних платформ.

ВИСНОВКИ

Розвиток методів соціальної інженерії та інтеграція генеративного штучного інтелекту в арсенал кіберзлочинців перетворили фішинг на надзвичайно складну, багатовекторну загрозу. Використання виключно традиційних, сигнатурних засобів захисту вже не гарантує безпеки інформаційних систем, що створює гостру необхідність у впровадженні проактивних аналітичних інструментів на базі штучного інтелекту.

Аналізуючи теоретичну базу та результати проведених практичних експериментів, можна дійти висновку, що застосування алгоритмів машинного навчання у поєднанні з технологіями обробки природної мови дозволяє кардинально підвищити рівень виявлення загроз. У ході дослідження було успішно сформовано репрезентативний збалансований корпус даних обсягом понад 22 тисячі повідомлень. Для перетворення текстової інформації у математичний формат було застосовано алгоритм TF-IDF, який довів свою ефективність у виокремленні ключових фішингових маркерів та відсіюванні інформаційного шуму.

Проведене емпіричне тестування чотирьох різних класифікаторів показало, що хоча лінійний метод опорних векторів досягає пікової точності у 99,25%, він не дозволяє гнучко управляти ризиками через відсутність імовірнісного прогнозування. Оптимальним вибором для практичної реалізації стала розроблена модель логістичної регресії. Вона продемонструвала видатний баланс характеристик: високу точність класифікації, блискавичний час навчання та можливість порогового налаштування чутливості системи, що є критично важливим для корпоративних середовищ.

На базі навченої моделі було розроблено практичний програмний продукт у вигляді утиліти командного рядка. Тестування в наближених до реальних умовах підтвердило її здатність безпомилково ідентифікувати

маніпулятивні шаблони та завуальовані атаки. Завдяки низьким вимогам до обчислювальних ресурсів та архітектурній гнучкості, розроблене рішення легко масштабується та ідеально підходить для інтеграції у сучасні B2B SaaS-системи, хмарні мікросервіси B2B або як модуль попередньої фільтрації для магістральних поштових шлюзів. Подальший розвиток подібних систем є надзвичайно перспективним напрямком для забезпечення глобальної кіберстійкості.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tang L., Mahmoud Q.H. A Survey of Machine Learning-Based Solutions for Phishing Website Detection // Machine Learning and Knowledge Extraction. – 2021. – Vol. 3(3). – P. 672–694. – DOI: 10.3390/make3030034

2. Proofpoint. OneDrive Phishing Campaign // Threat Reports. – 2023. – [Електронний ресурс]. – Режим доступу: https://www.proofpoint.com/us/resources/threat-reports

3. Cloudflare. AI-based Email and DNS Protection [Електронний ресурс]. – Режим доступу: https://developers.cloudflare.com/email-security/

4. Alshingiti Z., Alaqel R., Al-Muhtadi J., Haq Q.E.U., Saleem K., Faheem M.H. A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN // Electronics. – 2023. – Vol. 12(1). – Article №232. – DOI: 10.3390/electronics12010232

5. Kaggle. Phishing Email Dataset [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset

6. Colonial Pipeline Incident – Phishing Entry Point [Електронний ресурс]. – CISA, 2021. – Режим доступу: https://www.cisa.gov/news-events/news/colonial-pipeline-cyber-incident

7. Ferreira A., Teles A. Social Engineering Principles in Phishing // Journal of Internet Services and Applications. – 2019. – Vol. 10. – № 12. – DOI: 10.1007/s41060-024-00579-w.

8. Adaptive Capabilities of AI in Cyber Defense // Journal of Network and Systems Management. – 2020. – [Електронний ресурс]. – Режим доступу: https://link.springer.com/article/10.1007/s11235-020-00733-5

9. Verma R., Shashidhar N., Hossain N. Detecting Phishing Emails the Natural Language Way // In: Foresti S., Yung M., Martinelli F. (eds) Computer Security – ESORICS 2012. Lecture Notes in Computer Science, Vol. 7459. –

10. Springer, Berlin, Heidelberg. – 2012. – P. 824–841. – DOI: 10.1007/978-3-642-33167-1_47

11. Types of Phishing Attacks and How to Identify Them // CSO Online. – [Електронний ресурс]. – Режим доступу: https://www.csoonline.com/article/563353

12. Basit A., Qamar F., Zafar M., Jalil Z., Alazab M. A Comprehensive Survey of AI-enabled Phishing Attacks Detection Techniques // Journal of Network and Systems Management. – 2020. – Vol. 28. – P. 1125–1153. – DOI: 10.1007/s11235-020-00733-5.

13. ENISA Threat Landscape Report 2022 – Smishing Trends // ENISA. – 2022. – [Електронний ресурс]. – Режим доступу: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

14. Shombot E.S., Dusserre G., Bestak R., Ahmed N.B. An Application for Predicting Phishing Attacks: A Case of Implementing a Support Vector Machine Learning Model // Cyber Security and Applications. – 2024. – Vol. 2. – Article ID 100036. – DOI: 10.1016/j.csa.2024.100036

15. Microsoft. Office 365 Anti-Phishing // Digital Defense Report. – 2023. – [Електронний ресурс]. – Режим доступу: https://www.microsoft.com/en-us/security/blog/2023/10/12/2023-digital-defense-report/

16. Eze K., Shamir L. Analysis and Prevention of AI-based Phishing Email Attacks [Електронний ресурс] // arXiv preprint. – 2024. – arXiv:2405.05435. – Режим доступу: https://arxiv.org/abs/2405.05435.

17. Deepfake CEO Voice Fraud // BBC News. – 2019. – [Електронний ресурс]. – Режим доступу: https://www.bbc.com/news/technology-47481259

18. Ozcan A., Catal C., Donmez E., Senturk B. A Hybrid DNN–LSTM Model for Detecting Phishing URLs // Neural Computing and Applications. – 2021. – Vol.

19. Basit A. та ін. ML Accuracy in Phishing Detection // Journal of Network and Systems Management. – 2020. – Vol. 28. – P. 1125–1153.

20. Gmail AI Spam Filter // Google Blog. – 2022. – [Електронний ресурс]. – Режим доступу: https://blog.google/technology/safety-security/how-were-using-ai-to-combat-the-latest-scams/

21. ElAassal R. та ін. PhishBench Platform for Model Comparison [Електронний ресурс]. – Режим доступу: https://phishbench.github.io/

22. Smishing vs. Phishing vs. Vishing // HP Tech Takes. – [Електронний ресурс]. – Режим доступу: https://www.hp.com/us-en/shop/tech-takes/smishing-vs-phishing-vs-vishing

23. Sethi A., Garg S., Shafiq M.Z. Transformer Models for Phishing Detection // Journal of Internet Services and Applications. – 2022. – DOI: 10.1007/s41060-024-00579-w

24. APWG Report 2023 – Global Phishing Trends [Електронний ресурс]. – Режим доступу: https://apwg.org/trendsreports/

Додаток А

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

(Презентація)