ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.62 **DOI: 10.20998/2411-0558.2025.01.09**

Б. С. ЛИСОВ, асп.*,* НТУУ "КПІ ім. І. Сікорського", Київ, Україна,
В. Г. ГУСЬКОВА, PhD, доцент*,* НТУУ "КПІ ім. І. Сікорського", Київ, Україна,
Т. І. ПРОСЯНКІНА-ЖАРОВА, канд. екон. наук, доц., Інститут телекомунікацій та глобального інформаційного простору НАН України,
А. А. ХАЛИГОВ, асп.*,* Інститут телекомунікацій та глобального інформаційного простору НАН України

ІНФОРМАЙНА ТЕХНОЛОГІЯ КЛАСИФІКАЇ ДАНИХ МОНІТОРИНГУ СИСТЕМ І МЕРЕЖ ПЕРЕДАЧІ ДАНИХ ДЛЯ ВИЯВЛЕННЯ КІБЕРЗАГРОЗ

На сьогоднішній день безпека об’єктів критичної інфраструктури (КІ) є ключовим викликом у сучасному цифровому середовищі. Зростаюча складність інформаційних систем і методів кібератак потребує розробки ефективних підходів для виявлення та класифікації загроз. У даній роботі представлено методологію класифікації типів загроз для об’єктів критичної інфраструктури на основі мережевих даних. Запропоновано використання машинного навчання та методів аналізу аномалій для виявлення шкідливої активності. Проаналізовано сучасні алгоритми класифікації загроз та нейронні мережі. Отримані результати демонструють ефективність запропонованого підходу у виявленні та ідентифікації потенційних загроз, що дозволяє підвищити рівень кібербезпеки КІ.. Іл.: 2. Табл. 3. Бібліогр.: 16 назв.

Ключові слова: критична інфраструктура, класифікація загроз, машинне навчання, аналіз мережевих даних, кібербезпека

Вступ. Критична інфраструктура (КІ) складає основу сучасного суспільства, охоплюючи ключові сфери, такі як енергетика, транспорт, телекомунікації, охорона здоров’я, водопостачання та інше. Від безперебійного функціонування цих систем залежить стабільність суспільства, економічний розвиток та національна безпека. Будь-які збої або порушення в роботі КІ можуть призвести до значних негативних наслідків, включаючи фінансові втрати, порушення громадського порядку та навіть загрозу життю людей.

© Б.С. Лисов, В.Г. Гуськова, Т.І. Просянкіна-Жарова, А.А. Халигов, 2025

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Однак, із розвитком цифрових технологій та зростаючою взаємозалежністю інфраструктурних об’єктів збільшується ймовірність фізичних та кібернетичних загроз. Кібератаки на системи управління КІ можуть не лише порушувати їхню роботу, але й спричиняти катастрофічні наслідки, такі як відключення електроенергії в масштабах цілих регіонів, транспортні колапси або компрометація систем життєзабезпечення. У зв’язку з цим питання кібербезпеки КІ стає пріоритетним напрямом у сфері інформаційних технологій та національної безпеки.

Актуальність проблеми підсилюється стрімким поширенням високотехнологічних атак, що використовують методи штучного інтелекту, машинного навчання та автоматизованих кіберзасобів. Зловмисники застосовують складні алгоритми для обходу традиційних систем безпеки, роблячи кібератаки дедалі складнішими для виявлення та нейтралізації. Водночас, зростання обсягів мережевого трафіку та великих даних створює додаткові виклики для аналітичних систем безпеки, що вимагає розробки нових методів аналізу та класифікації загроз.

Ефективне виявлення, аналіз та класифікація загроз для об’єктів критичної інфраструктури є невід’ємною частиною стратегій кіберзахисту. Використання сучасних методів обробки мережевих даних, зокрема алгоритмів машинного навчання та аналізу аномалій, дозволяє підвищити рівень захисту КІ, забезпечуючи виявлення потенційних атак у режимі реального часу.

Мета. Метою даного дослідження є розробка та впровадження ефективної методології класифікації загроз для об’єктів критичної інфраструктури на основі мережевих даних. Дослідження спрямоване на виявлення найбільш релевантних підходів та алгоритмів, які дозволяють аналізувати аномалії в мережевому трафіку, визначати типи атак і прогнозувати їхній розвиток. Основна увага приділяється застосуванню методів машинного навчання та нейронним мережам, для покращення точності та швидкості виявлення загроз. Запропонована методологія має забезпечити можливість своєчасного реагування на потенційні атаки та мінімізації ризиків для критичної інфраструктури.

Задача дослідження постає у розробці ефективних підходів до використання мережевих даних для виявлення та класифікації загроз

критичної інфраструктури, зокрема ідентифікації аномалій, типів атак та відповідних механізмів реагування [1 – 3].

Поточна задача є надзвичайно складною та вимагає вирішення кількох взаємопов'язаних завдань. Природа мережевих даних потребує роботи з багатокритеріальними часовими рядами, де різні атрибути є часово залежними та взаємопов'язаними. Запропонована методологія для вирішення задачі класифікації та прогнозування кіберзагроз, спрямованих на об'єкти критичної інфраструктури, включає три основні етапи:

- *Попередня обробка даних* – усунення пропущених значень, аналіз та обробка аномалій, нормалізація ознак і зменшення розмірності для забезпечення якості даних перед подальшим аналізом.
- *Класифікація загроз* – розробка та валідація моделей для ідентифікації та класифікації загроз на основі їхніх характеристик та ключових індикаторів.
- *Практичне застосування результатів* – використання отриманих класифікованих даних для розробки рекомендацій щодо реагування на загрози та мінімізації їхнього впливу на критичну інфраструктуру.

Кожен із цих етапів спрямований на ретельну обробку та аналіз мережевих даних із застосуванням сучасних методів машинного навчання та статистичного аналізу для забезпечення точного виявлення та прогнозування загроз.

Крок 0. Загальний алгоритм аналізу даних часового ряду.

Оскільки у задачах кібербезпеки для критичної інфраструктури дані постійно оновлюються та динамічно змінюються з часом, було запропоновано застосування одночасно кількох підходів до аналізу та класифікації. Це зумовлено тим, що один метод може ефективно працювати з певним набором даних, але виявлятися менш придатним для інших. Таким чином, мета нашого дослідження полягає у створенні адаптивної та автоматизованої системи, яка зможе адаптуватись до змін у даних, забезпечуючи високу якість аналізу та прогнозування [2 – 4].

Нижче наведено загальний алгоритм аналізу даних для класифікації кіберзагроз.

Крок 1: Попередній аналіз даних. На цьому етапі проводиться підготовка та первинне дослідження даних з метою забезпечення їхньої якості, релевантності та придатності для подальшого моделювання.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Попередній аналіз даних є фундаментальним етапом, який закладає основу для успішної реалізації подальших методів аналізу, прогнозування та класифікації [5].

Основні завдання цього етапу включають:

Крок 1.1: Перевірка даних на пропуски. Перевірка даних на пропуски є важливим етапом попереднього аналізу, який дозволяє оцінити якість даних і виявити проблеми, пов’язані з відсутністю значень у наборах даних. Пропуски можуть впливати на результати аналізу та моделювання, тому їх виявлення та обробка є критично важливими.

Для оцінки кількості пропусків у даних було використано спеціалізовані функції, які дозволяють ефективно виявляти відсутні значення. Наприклад, у Python для цього застосовуються методи df.isnull().sum() або df.info(), які надають інформацію про кількість пропусків у кожному стовпці, а в середовищі R використовується функція sum(is.na(data)), що дозволяє підрахувати загальну кількість пропущених значень у наборі даних. Результати розрахунку кількості пропущених значень представлені на рис. 2.

Аналіз набору даних "Cyber Security Attacks" з Kaggle показав наявність значних пропусків у кількох атрибутах (до 20,067 значень). Для візуалізації розподілу пропусків можна використовувати теплові карти (sns.heatmap у Python). Змінні з понад 15% пропусків доцільно вилучити, оскільки вони можуть спотворювати аналіз.

Оскільки ми використовуємо такі підходи як Random Forest, XGBoost та SVM для аналізу кіберзагроз, важливо правильно підготувати дані, зокрема перетворити категорійні змінні у числовий формат. Алгоритми машинного навчання, такі як ці три моделі, працюють лише з числовими значеннями, тому такі атрибути, як Protocol, Packet Type, Traffic Type, Severity Level та Attack Type, потрібно закодувати, щоб вони могли бути враховані моделлю.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

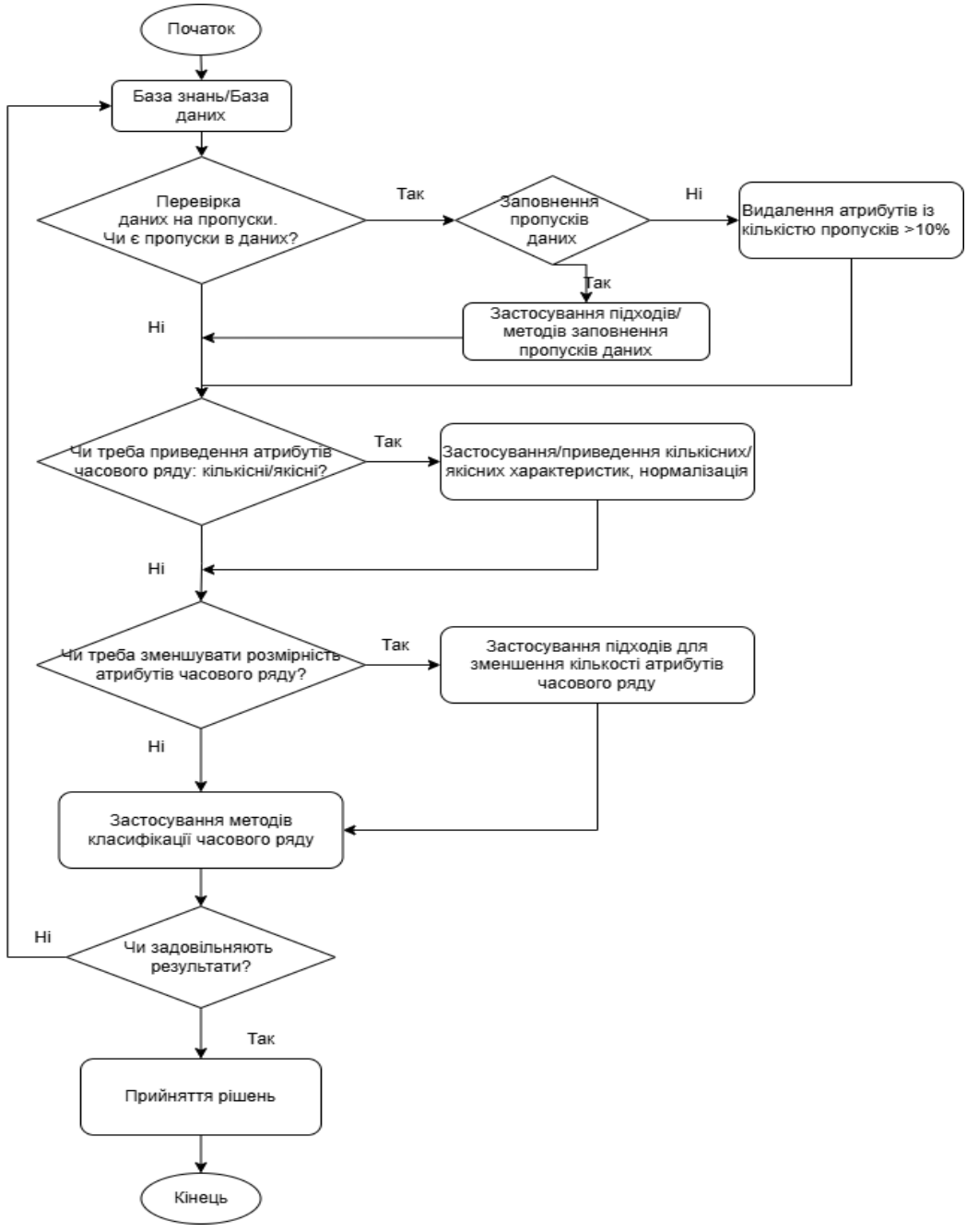

Рис. 1. Загальний аналіз класифікації кіберзагроз

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

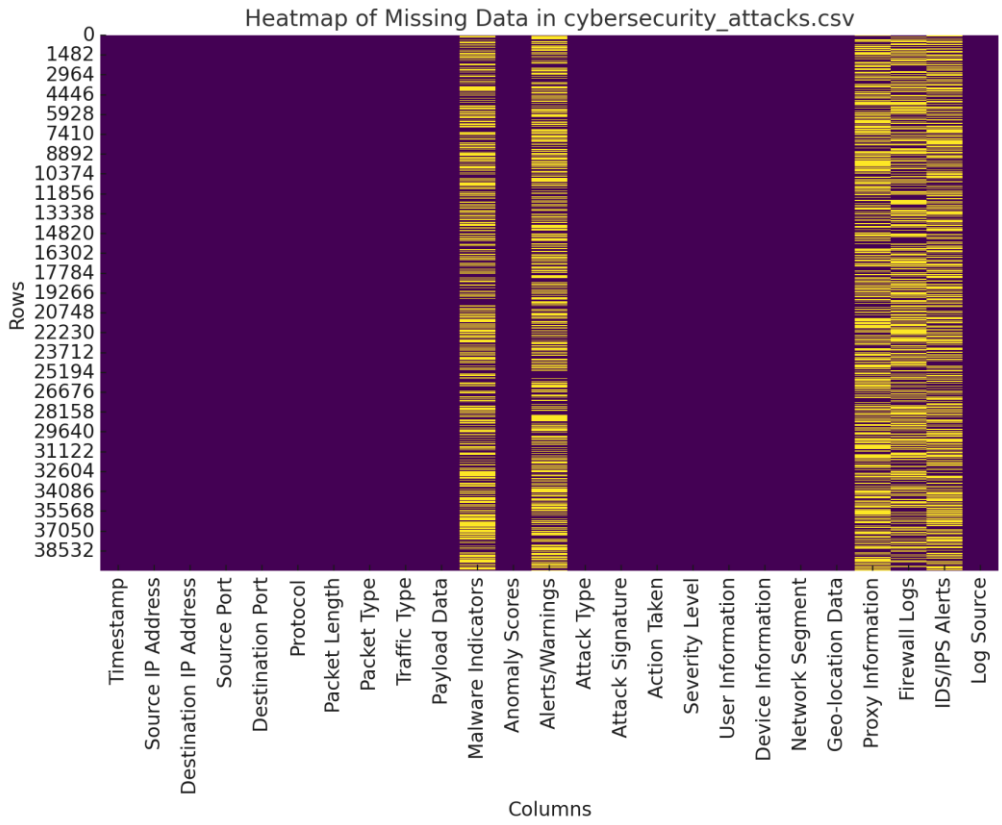

Рис. 2. Пропуски даних в початковому наборі даних

Крок 1.2: Зменшення кількості ознак. Зменшення кількості ознак (feature reduction) є критичним етапом підготовки даних, оскільки дозволяє оптимізувати процес аналізу та моделювання шляхом вибору найбільш інформативних змінних або трансформації існуючих. Цей підхід забезпечує підвищення ефективності моделей за рахунок зниження обчислювальної складності, зменшення часу обробки даних і зниження ризику перенавчання. У контексті кібербезпеки, де набори даних можуть містити сотні або тисячі атрибутів, зменшення розмірності є особливо важливим. Воно дозволяє сконцентруватися на найбільш значущих характеристиках, які впливають на виявлення загроз і аномалій, забезпечуючи більш точне й ефективне моделювання, навіть у високорозмірних або складних середовищах [6].

Для отримання якісних результатів розглядається комбінація підходів, а саме застосування декількох методів: фільтраційний метод для видалення нерелевантних ознак.

Фільтраційний метод базується на статистичному аналізі кожної ознаки щодо її значущості для цільової змінної. В якості цільової змінної беремо – "Attack Type". Для кожної ознаки X_i з набору $\{X_1, X_2, \ldots, X_n\}$ обчислюється значення міри значущості S_i, що показує, наскільки ця ознака впливає на цільову змінну y. Загальна формула:

$$S_i = f(X_i, y),\ i = 1, 2, \ldots, n, \tag{1}$$

де f – це критерій значущості, який залежить від типу ознаки X_i та природи змінної y. Нижче розглянуто основні критерії значущості для зменшення кількості атрибутів початкового набору даних.

1. **Коефіцієнт кореляції.** Використовується для числових ознак та числової цільової змінної:

$$S_i = \text{corr}(X_i, y) = \frac{\text{cov}(X_i, y)}{\sigma(X_i) \cdot \sigma(y)}, \tag{2}$$

де:

$\text{cov}(X_i, y)$ – коваріація між X_i та y;

$\sigma(X_i) \cdot \sigma(y)$ – стандартні відхилення.

2. **Інформаційний приріст** Використовується для категорійних або змішаних змінних:

$$S_i = \text{IG}(X_i, y) = H(y) - H(y|X_i), \tag{3}$$

де:

$H(y)$ – ентропія цільової змінної y,

$H(y|X_i)$ умовна ентропія y за умови X_i.

3. **Статистичний тест χ2** Використовується для дискретних змінних:

$$S_i = \chi^2(X_i, y) = \sum_k \frac{(O_k - E_k)^2}{E_k}, \tag{4}$$

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

де:

O_k – спостережене значення,

E_k – очікуване значення.

2. Вибір релевантних ознак:

Після обчислення S_i для всіх ознак обираються ті, що мають значення значущості, яке перевищує встановлений поріг τ:

$$\chi_{filtered}=\{X_i|S_i \geq \tau\}. \quad (5)$$

τ – це поріг значущості, який може бути встановлений вручну на основі попереднього аналізу або визначений автоматично (наприклад, шляхом відсікання нижніх q% ознак). Результатом є зменшений набір ознак який використовується для подальшого аналізу чи моделювання [7, 8].

За кожним методом відбору ознак було обчислено коефіцієнт значущості, який дозволяє визначити, наскільки кожна ознака впливає на цільову змінну. Це дало можливість зробити висновок про кількість атрибутів, які можна вважати статистично значущими для подальшого аналізу. Підсумкові результати оцінки представлені у табл. 1 (атрибути відсортовані за рівнем значущості).

Для прийняття рішення щодо включення ознаки до наступного етапу аналізу було застосовано бінарну класифікацію:

- Якщо значення метрики для ознаки відповідало встановленому порогу, їй присвоювалося 1 (включається в подальший аналіз).
- Якщо значення не досягало порогового рівня, ознака отримувала 0 (відсікається з подальшого аналізу).

Далі було підраховано суму балів для кожної ознаки за всіма методами.

- Ознаки, які набрали вищу кількість балів, було визнано релевантними, і вони були включені до подальшого моделювання.
- Ознаки з низькою сумою балів вважалися незначущими та виключалися з наступних етапів аналізу.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Таблиця 1.

Відбір релевантних ознак для подальшого аналізу

Атрибути	**Значення**
Source Port	Так
Destination Port	Так
Payload Data	Так
User Information	Так
Source IP Address	Так
Timestamp	Так
Device Information	Так
Destination IP Address	Так
Geo-location Data	Так
Packet Length	Так
Anomaly Scores	Так
Action Taken	Так
Packet Type	Ні
Severity Level	Ні
Protocol	Ні
Attack Signature	Ні
Network Segment	Ні
Traffic Type	Ні
Log Source	Ні

Такий підхід дозволяє об’єктивно відібрати найважливіші ознаки, використовуючи поєднання декількох методів оцінки значущості. Це допомагає підвищити ефективність моделей, зменшуючи розмірність простору ознак без втрати критичної інформації.

Крок 2. Підхід до класифікації загроз.

Оскільки в нашому наборі даних є атрибут Attack Type (мітки класів), ми застосовуємо контрольоване навчання (Supervised Learning). Контрольоване навчання – це тип машинного навчання, в якому алгоритм отримує набір пар "вхідні дані – очікуваний результат" та намагається знайти залежності між ознаками та мітками класів.

Маємо набір даних $D=\{(X_1,Y_1), (X_2,Y_2), \ldots, (X_n,Y_n)\}$, де:

X_i – набір вхідних ознак (атрибутів).

Y_i – відповідна мітка класу або числове значення (цільова змінна).

Алгоритм будує функцію $f(x)\rightarrow Y$, яка дозволяє передбачати мітку Y для нових даних.

Цільова змінна Attack Type включає наступні типи загроз:

- **Malware** – Шкідливе програмне забезпечення (віруси, трояни, шпигунські програми).
- **DDoS** – Розподілена атака на відмову в обслуговуванні (перевантаження мережевих ресурсів).
- **Intrusion** – Вторгнення (несанкціонований доступ або злом системи).

Для ефективної класифікації кіберзагроз важливо правильно підібрати алгоритми, які зможуть врахувати особливості набору даних, типи атак та складність їхньої взаємодії. Різні підходи мають свої переваги та недоліки, тому порівняння кількох методів допоможе визначити найкращий варіант для вирішення поставленої задачі.

В рамках даної роботи було використано: Деревоподібні моделі (Random Forest, XGBoost) які добре працюють на великих наборах даних і є стійкими до шуму та можуть визначати важливість ознак. А також метод опорних векторів (SVM), який є ефективним для точного розділення класів, особливо якщо вони мають чіткі межі. Цей підхід добре працює у випадках, коли важливо знайти оптимальні роздільні гіперплощини [9, 10].

2.1. Random Forest є ансамблевим методом машинного навчання, який використовує техніку побудови множини дерев рішень (Decision Trees) і комбінує їх результати для покращення загальної продуктивності моделі. Основна ідея методу полягає в створенні ансамблю "випадкових" дерев, які працюють разом для прийняття точних і стійких рішень.

Нехай початковий набір даних складається з N об’єктів і p ознак:

$$D=\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}, \quad (6)$$

де $x_i \in R^p$, y_i — відповідна мітка класу.

Вибір випадкових ознак:

На кожному вузлі дерева вибирається випадкова підмножина ознак mmm (зазвичай $m = \sqrt{p}$) для пошуку оптимального розділення.

Комбінація результатів:

Для класифікації:

$$\hat{y} = majority\_vote\{T_1(x), T_2(x), \ldots, T_k(x)\}, \quad (7)$$

де $T_j(x)$ – прогноз j-го дерева.

Для регресії:

$$\hat{y} = \frac{1}{k}\sum_{j=1}^{k} T_j(x). \quad (8)$$

У таблиці 2 наведені результати застосування методу до початкового набору даних.

Таблиця 2.

Характеристика застосування Random Forest

Клас	Точність	Recall
0	63%	67%
1	66%	64%
2	63%	62%

Результати навчання моделі Random Forest показали загальну точність 64.3%, що вказує на недостатньо ефективну класифікацію атак. Розподіл точності між класами Attack Type є приблизно рівномірним, проте загальна якість прогнозування залишається середньою.

2.2. Gradient Boosting є методом ансамблювання, який використовує ітеративний підхід для побудови сильної моделі шляхом послідовного навчання слабких моделей (наприклад, дерев рішень). Він базується на мінімізації похибки через градієнтний спуск.

Задача навчання полягає в пошуку апроксимації для цільової функції $f^*(X)$ шляхом ітеративного додавання нових базових моделей $h_m(x)$, які зменшують похибку попередньої моделі:

$$F_m(x) = F_{m\text{-}1}(X) + \gamma_m h_m(X), \quad (9)$$

де:

$F_m(x)$ – прогноз моделі на mmm-ій ітерації,

$F_{m\text{-}1}(X)$ – прогноз попередньої моделі,

$h_m(X)$– нова база (слабка модель), яка мінімізує залишкову помилку,

γ_m – коефіцієнт навчання (learning rate), який контролює внесок кожного нового дерева.

2. Мінімізація втрат через градієнтний спуск

На кожній ітерації додається нова база $h_m(X)$, яка навчена на залишкових помилках попередньої моделі. Мінімізується функція втрат $L(y, F(X))$.

$$h_m = argmin_h \sum_{i=1}^{N} \mathrm{L}(y_i, \mathrm{F}_{m\text{-}1}(X_i) + h(X_i)). \quad (10)$$

Градієнт втрат обчислюється як:

$$g_i = -\frac{\partial L\left(y_i, F(X_i)\right)}{\partial F(X_i)} \quad (11)$$

Тобто модель намагається наблизити негативний градієнт функції втрат, що дає напрямок, у якому потрібно рухатися для зменшення помилки [11 – 13].

У табл. 3 наведено результати застосування двох методів класифікації – Random Forest та Gradient Boosting – до початкового набору даних.

Таблиця 3.

Характеристика застосування Gradient Boosting

Модель	**Precision (середнє)**	**Recall (середнє)**	**F1-score (середнє)**
Random Forest	64%	64%	64%
Gradient Boosting	63%	63%	63%

Основні статистичні характеристики моделей: Random Forest показав середню точність 64.3%, що є трохи вищим результатом порівняно з Gradient Boosting (62.9%). Значення Precision, Recall та F1-score для обох моделей майже однакові (64% для Random Forest та 63% для Gradient Boosting), що вказує на схожі характеристики класифікації.

2.3. Метод опорних векторів

Support Vector Machine (SVM) шукає оптимальну гіперплощину, яка розділяє класи у багатовимірному просторі з максимальною відстанню між класами. Цей метод добре працює з невеликими та середніми за обсягом наборами даних, а також зі складними структурами даних завдяки використанню ядерних функцій (kernel functions). В рамках поточного експерименту було обрано лінійне ядро (kernel='linear'), що є більш швидким, але менш ефективне для нелінійних залежностей [14 – 15].

Основне рівняння гіперплощини. Гіперплощина в N-вимірному просторі задається рівнянням:

$$f(x) = \omega^T x + b = 0, \tag{12}$$

де:

x – вектор ознак (вхідні дані),

ω^T – вектор ваг моделі (нормаль до гіперплощини),

b – зміщення (bias).

Для класифікації потрібно знайти таку оптимальну гіперплощину, яка максимально віддаляє точки двох класів [16].

Умова класифікації. Припустимо, що мітки класів $y \in \{-1,+1\}$, тоді гіперплощина розділяє їх за правилом:

$$y_i(\omega^T x_i + b) \geq 1, \forall i. \tag{13}$$

Це означає, що:

Якщо $y_i = +1$, то $(\omega^T x_i + b) \geq 1$;

Якщо $y_i = -1$, то $(\omega^T x_i + b) \leq -1$

Оптимізаційна задача (максимізація зазору). SVM шукає максимальний зазор (margin) між класами. Відстань між підтримуючими гіперплощинами (support vectors) дорівнює:

$$\frac{2}{\|\omega\|}. \tag{14}$$

Оптимізаційна задача записується як задача мінімізації:

$$\min_{w,b} \frac{1}{2}\|\omega\|^2, \tag{15}$$

за умови:

$$y_i(\omega^T x_i + b) \geq 1, \forall i. \tag{16}$$

Ця задача є опуклою задачею квадратичного програмування (QP).

За результатами експерименту SVM показав результат який дорівнює 62.7%.

Результати дослідження. За результатами тестування трьох моделей машинного навчання – SVM, Random Forest та Gradient Boosting, найбільш ефективними виявилися Random Forest і Gradient Boosting, які показали точність, повноту та F1-міру на рівні 63%. Метод SVM із лінійним ядром продемонстрував трохи гірші результати, що свідчить про те, що дані, ймовірно, мають нелінійний характер і потребують використання більш гнучких підходів.

Результати тестування показали, що загальні методи класифікації, такі як SVM, Random Forest та Gradient Boosting, продемонстрували приблизно однакову ефективність, досягнувши рівня точності близько 62-63%. Це є досить непоганим результатом, враховуючи складність задачі, але все ще недостатнім для високоточної класифікації.

Обговорення. Результати тестування показали, що загальні методи класифікації, такі як SVM, Random Forest та Gradient Boosting, продемонстрували приблизно однакову ефективність на рівні 62 – 63%. Це свідчить про стабільність застосованих підходів, але водночас вказує на необхідність подальшої оптимізації. Незважаючи на достатньо непогані результати, для покращення точності класифікації необхідно розглянути альтернативні підходи.

Одним із ключових напрямків удосконалення є розширення набору ознак. Поточні характеристики, такі як номери портів і довжина пакета, можуть не повністю відображати природу атак. Включення додаткових параметрів, таких як протокол, тип трафіку, статистичні особливості пакетів та поведінкові особливості джерел трафіку, може суттєво покращити результати. Також варто приділити увагу оптимізації гіперпараметрів, оскільки правильний вибір параметрів для кожної моделі може суттєво вплинути на продуктивність. Наприклад, для Random Forest можна змінювати кількість дерев і їхню глибину, для Gradient Boosting – швидкість навчання та кількість слабких моделей, а для SVM – тип ядра та параметри регуляризації.

Окрім цього, варто розглянути альтернативні методи машинного навчання, які можуть краще адаптуватися до складної структури даних. Глибокі нейронні мережі здатні виявляти складні закономірності у великих обсягах трафіку, а ансамблеві методи можуть поєднувати сильні сторони різних класифікаторів. Також важливим є підхід до виявлення аномалій, де можна застосувати методи, такі як One-Class SVM або Autoencoders, які здатні виявляти невідомі або рідкісні типи атак.

Ще одним напрямком вдосконалення є балансування класів у навчальній вибірці. Якщо певні типи атак зустрічаються значно рідше за інші, модель може їх ігнорувати. Використання технік збалансованого навчання, таких як SMOTE (Synthetic Minority Over-sampling Technique) або вагові коефіцієнти у функції втрат, може суттєво покращити здатність моделі розпізнавати рідкісні атаки. Додатково важливо проводити фільтрацію даних, щоб виключити шумові записи, які можуть спотворювати процес навчання.

У практичному застосуванні такі моделі можуть бути інтегровані у системи виявлення вторгнень (IDS/IPS), SIEM-системи для аналізу подій безпеки, а також використовуватися для автоматичного моніторингу мережевого трафіку. Виявлення атак у реальному часі дозволить суттєво підвищити рівень кібербезпеки та своєчасно реагувати на загрози.

Висновки. У даній роботі розглянуто проблему виявлення та класифікації загроз для об’єктів критичної інфраструктури на основі аналізу мережевих даних. Зростаюча складність інформаційних систем і методів кібератак робить актуальним використання методів машинного навчання та аналізу аномалій для виявлення загроз у режимі реального часу.

У дослідженні було проведено аналіз сучасних підходів до класифікації загроз, серед яких розглянуто деревоподібні моделі (Random Forest, Gradient Boosting) та метод опорних векторів (SVM). В якості вхідних даних використовувалися мережеві параметри, такі як номери портів, довжина пакету, тип трафіку, рівень загрози та інші.

Результати експериментального дослідження показали, що методи Random Forest та Gradient Boosting продемонстрували найкращу точність (64%), тоді як SVM із лінійним ядром досяг рівня 62,7%. Це свідчить про ефективність ансамблевих методів у задачах класифікації мережевого

трафіку. Усі три підходи продемонстрували схожий рівень продуктивності, що підтверджує можливість їх використання для автоматизованого аналізу кіберзагроз.

Запропонована методологія включає основні етапи: попередню обробку даних та класифікацію загроз. Використання машинного навчання дозволяє не лише покращити точність ідентифікації атак, але й забезпечити оперативне виявлення загроз, що є критично важливим для захисту критичної інфраструктури. Отримані результати можуть бути застосовані у системах моніторингу кібербезпеки, SIEM-платформах та системах виявлення вторгнень (IDS/IPS).

Таким чином, проведене дослідження підтверджує ефективність застосування методів машинного навчання у сфері кібербезпеки, а отримані результати можуть слугувати основою для подальшої автоматизації процесів аналізу мережевого трафіку та виявлення потенційних загроз.

Список літератури:

1. Tsay R. Analysis of Financial Time Series / R. Tsay. – New York: John Wiley & Sons, 2010. – 715 p.
2. Forecasting Methods in Decision Support Systems / [S.Dovhyi, P.Bidyuk, O.Trofymchuk, O.Savenkov]. – Kyiv: Azimut-Ukraine, 2011. – 608 p.
3. Dovhyi S. Decision Support Systems Based on Statistical and Probabilistic Methods / S. Dovhyi, P. Bidyuk, O. Trofymchuk. – Kyiv: Logos, 2014. – 419 p.
4. Harvey A. Econometric Analysis of Time Series / A.Harvey – Cambridge, MA: MIT Press, 1990. – 402 p.
5. Harvey A. Forecasting Structural Time Series Models and the Kalman Filter / A.Harvey – Cambridge: Cambridge University Press, 1994. – 554 p.
6. Engle R. Handbook of Econometrics / R. Engle, D. McFadden. – New York: Elsevier Science, 1994. – 1044 p.
7. Tjostheim D. Some Doubly Stochastic Time Series Models / D. Tjostheim // Journal of Time Series Analysis. – 1986. – Vol.7, № 1. – P. 51–72.
8. Terasvirta T. Aspects of Modeling Nonlinear Time Series / T. Terasvirta, T. Terasvirta, C.W.J.Granger // In Handbook of Econometrics. – 1994. – Vol. 13, Issue 1. – P. 2919–2957.
9. Tsay R. Testing and Modeling Threshold Autoregressive Processes / R. Tsay // Journal of the American Statistical Association. – 1989. – Vol. 84., № 3. – P.231–240.
10. Gallant A. Nonlinear Statistical Models / A. Gallant. – New York: John Wiley & Sons, 1987. – 622 p.
11. Hinich M. Testing for Dependence in the Input to a Linear Time Series Model / M. Hinich // Journal of Nonparametric Statistics, 1996. – Vol. 2. – P. 205–221.
12. Brooks C. Bicorrelations and Cross-Bicorrelations as Non-Linearity Tests and Tools for Exchange Rate Forecasting / C. Brooks, M. Hinich // Journal of Forecasting. – 2001. – Vol. 20. – P.181–196.
13. Tsay R. Model Checking via Parametric Bootstraps in Time Series Analysis / R. Tsay // Applied Statistics. – 1989. – Vol. 41, № 1. – P. 1-15.

14. Chen R. Functional Coefficient Autoregressive Models / R. Chen, R. Tsay // Journal of the American Statistical Association. – 1993. – Vol. 88. – P. 298–308.
15. Lewis P. Semi-Multivariate Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS) / P. Lewis, L. Stevens // Journal of the American Statistical Association. – 1991. – Vol. 86. – P. 864–877.
16. Friedman J. Multivariate Adaptive Regression Splines / J. Friedman // Annals of Statistics. – 1991. – Vol. 19. – P. 1–141.

References:

1. Tsay R. Analysis of Financial Time Series / R. Tsay. – New York: John Wiley & Sons, 2010. – 715 p.
2. Forecasting Methods in Decision Support Systems / [S.Dovhyi, P.Bidyuk, O.Trofymchuk, O.Savenkov]. – Kyiv: Azimut-Ukraine, 2011. – 608 p.
3. Dovhyi S. Decision Support Systems Based on Statistical and Probabilistic Methods / S. Dovhyi, P. Bidyuk, O. Trofymchuk. – Kyiv: Logos, 2014. – 419 p.
4. Harvey A. Econometric Analysis of Time Series / A.Harvey – Cambridge, MA: MIT Press, 1990. – 402 p.
5. Harvey A. Forecasting Structural Time Series Models and the Kalman Filter / A.Harvey – Cambridge: Cambridge University Press, 1994. – 554 p.
6. Engle R. Handbook of Econometrics / R. Engle, D. McFadden. – New York: Elsevier Science, 1994. – 1044 p.
7. Tjostheim D. Some Doubly Stochastic Time Series Models / D. Tjostheim // Journal of Time Series Analysis. – 1986. – Vol.7, № 1. – P. 51–72.
8. Terasvirta T. Aspects of Modeling Nonlinear Time Series / T. Terasvirta, T. Terasvirta, C.W.J.Granger // In Handbook of Econometrics. – 1994. – Vol. 13, Issue 1. – P. 2919–2957.
9. Tsay R. Testing and Modeling Threshold Autoregressive Processes / R. Tsay // Journal of the American Statistical Association. – 1989. – Vol. 84., № 3. – P.231–240.
10. Gallant A. Nonlinear Statistical Models / A. Gallant. – New York: John Wiley & Sons, 1987. – 622 p.
11. Hinich M. Testing for Dependence in the Input to a Linear Time Series Model / M. Hinich // Journal of Nonparametric Statistics, 1996. – Vol. 2. – P. 205–221.
12. Brooks C. Bicorrelations and Cross-Bicorrelations as Non-Linearity Tests and Tools for Exchange Rate Forecasting / C. Brooks, M. Hinich // Journal of Forecasting. – 2001. – Vol. 20. – P.181–196.
13. Tsay R. Model Checking via Parametric Bootstraps in Time Series Analysis / R. Tsay // Applied Statistics. – 1989. – Vol. 41, № 1. – P. 1-15.
14. Chen R. Functional Coefficient Autoregressive Models / R. Chen, R. Tsay // Journal of the American Statistical Association. – 1993. – Vol. 88. – P. 298–308.
15. Lewis P. Semi-Multivariate Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS) / P. Lewis, L. Stevens // Journal of the American Statistical Association. – 1991. – Vol. 86. – P. 864–877.
16. Friedman J. Multivariate Adaptive Regression Splines / J. Friedman // Annals of Statistics. – 1991. – Vol. 19. – P. 1–141.

Статтю представив д.т.н., проф. Національного технічного університету "Харківський політехнічний інститут" С.Ю. Леонов.

Надійшла (received) 11.05.2025

Lysov Bogdan, PhD Student
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
av. Peremohy, 37, Kyiv, Ukraine, 03056
Tel.: +38093-218-76-69, E-mail: bogukraine@gmail.com

Huskova Vira, PhD, Associate Professor
Department of Artificial Intelligence,
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
av. Peremohy, 37, Kyiv, Ukraine, 03056
Tel.: +38095-626-65-80, E-mail: huskova.vira@lll.kpi.ua
ORCID ID: 0000-0001-7637-201X

Prosiankina-Zharova Tetiana, Dr. Sci. Tech., Associate Professor
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
av. Peremohy, 37, Kyiv, Ukraine, 03056
Tel.: +38097-521-15-08, E-mail: tpruman@gmail.com

Khalygov Artem, PhD Student
Institute of Telecommunications and Global Information Space,
National Academy of Sciences of Ukraine,
13, Chokolivskyi Boulevard, Kyiv 186, 03186, Ukraine
Tel.: +38050-049-29-03, E-mail: khalygovartem@gmail.com

УДК 004.62

Інформаійна технологія класифікаї даних моніторингу систем і мереж передачі даних для виявлення кіберзагроз Лисов Б., Гуськова В.Г., Просянкіна-Жарова Т., Халигов А. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2025. – № 1 (13). – С. 129 – 147.

На сьогоднішній день безпека об’єктів критичної інфраструктури (КІ) є ключовим викликом у сучасному цифровому середовищі. Зростаюча складність інформаційних систем і методів кібератак потребує розробки ефективних підходів для виявлення та класифікації загроз. У даній роботі представлено методологію класифікації типів загроз для об’єктів критичної інфраструктури на основі мережевих даних. Запропоновано використання машинного навчання та методів аналізу аномалій для виявлення шкідливої активності. Проаналізовано сучасні алгоритми класифікації загроз та нейронні мережі. Отримані результати демонструють ефективність запропонованого підходу у виявленні та ідентифікації потенційних загроз, що дозволяє підвищити рівень кібербезпеки КІ.. Бібліогр.: 16 назв.

Ключові слова: критична інфраструктура, класифікація загроз, машинне навчання, аналіз мережевих даних, кібербезпека.

UDC 004.62

Information technology for the classification of data monitoring systems and data transmission networks for cyber threat detection/ Lysob B., Huskova V., Prosyankina-Zharova T., Khalyhov A. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2025. – № 1 (13). – P. 129 – 147.

Nowadays, the security of critical infrastructure (CI) is a key challenge in the modern digital environment. The growing complexity of information systems and cyberattack methods necessitates the development of effective approaches for threat detection and classification. This paper presents a methodology for classifying types of threats for critical infrastructure based on network data. It proposes the use of machine learning and anomaly detection methods to identify malicious activities. Modern threat classification algorithms and neural networks are analyzed. The obtained results demonstrate the effectiveness of the proposed approach in detecting and identifying potential threats, thereby enhancing the cybersecurity level of critical infrastructure.

Key words: critical infrastructure, threat classification, machine learning, network data analysis, cybersecurity.