ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

КВАЛІФІКАЦІЙНА РОБОТА

Метод виявлення аномалій у журналах API для забезпечення безпеки та надійності програмних систем

ВИКОНАВ:
Студент гр СПм-23-4 Авдєєв О. С.

КЕРІВНИК:
доц. Мартовицький В.О.

ХАРКІВ
2025р.

Актуальність дослідження

•**API** – ключовий елемент взаємодії між сервісами у сучасному програмному забезпеченні.

•**Зловмисна активність** часто маскується під звичайні запити до API.

•**Логи API** містять критичні дані, за якими можна виявити кіберзагрози.

•**Зростання кількості атак** через API вимагає автоматизованих методів моніторингу.

•**Методи машинного навчання** дозволяють ефективно аналізувати великі об’єми логів.

Мета та завдання

🎯 **Мета роботи:**
Розробка ефективного методу виявлення аномалій у журналах API для підвищення рівня безпеки та надійності програмних систем.

📌 **Основні завдання:**
1.Аналіз форматів та структури логів API.
2.Огляд існуючих підходів до виявлення аномалій.
3.Підготовка датасету з логів API для навчання моделей.
4.Розробка системи на основі методів машинного навчання.
5.Тестування системи на реальних або симульованих даних.
6.Оцінювання ефективності розробленого рішення.

3

Інформація, яка зазвичай міститься в журналах API

Інформація про лог API	Опис
Позначка часу (Timestamp)	Вказує, коли був зроблений виклик API
Кінцева точка (Endpoint)	Яка кінцева точка API була доступна
Параметри запиту (Request Parameters)	Деталізація запитаної команди
Ідентифікація користувача (User Identification)	Зазвичай ID або токени, що використовуються для ідентифікації
Дані відповіді (Response data)	Що API повернув у відповідь
Коди стану (Status Codes)	Вказує на успішність або помилку запиту
IP-адреси та інформація про пристрої	Інформація про місце та пристрій, з якого надійшов запит

4

Проблема безпеки API

❗ **Основні загрози:**

•**Невалідація запитів:** API, що не перевіряє коректність вхідних даних, стає вразливим до SQL-ін'єкцій, XSS тощо.
•**Витік конфіденційної інформації:** неправильне логування або помилки автентифікації можуть призвести до розголошення персональних даних.
•**Недостатній контроль доступу:** погано налаштована авторизація дозволяє користувачам отримувати доступ до сторонніх або критичних ресурсів.
•**Аномальна поведінка клієнтів:** автоматизовані атаки (боти, сканери) можуть залишатися непоміченими без належного моніторингу.

Наслідки:

•Втрати даних або доступ сторонніх осіб до системи.
•Порушення законодавства (GDPR, ISO/IEC 27001).
•Втрата репутації компанії.
•Фінансові збитки через зупинку сервісів або штрафи.

5

Набір даних

Короткий опис набору даних

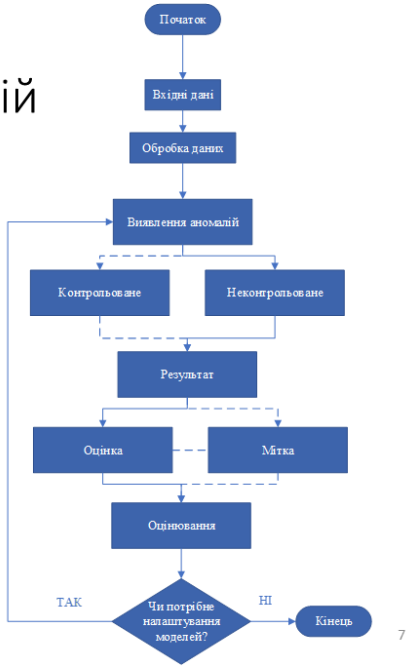
Аспект	Значення
Розмір набору даних	61000
Нормальні записи	36000
Аномалії	25000
Розподіл класів	Норма: 59%, Аномалії: 41%
Джерело	Giménez та ін. [58]
Типи даних	Змішані (категоріальні, числові, текстові)

Опис використаних стовпців журналу CSIC

Стовпець	Опис
method	HTTP-метод, що використовується (наприклад, GET, POST).
host	Домен запиту.
accept	Типи контенту, які клієнт може обробити.
content_length	Довжина запиту.
content	Фактичний вміст запиту.
url	URL (кінцева точка API), до якого було зроблено запит.
classification	Мітки, що відповідають поняттю «клас», використовуються для оцінки.

6

Схема запропонованого підходу до виявлення аномалій



Етапи підготовки

Вибір моделей для неконтрольованого виявлення аномалій

📌 **Після етапу підготовки даних було обрано чотири моделі**, що найбільш відповідають завданню аналізу API-журналів у неконтрольованому режимі:

🔍 **Критерії вибору:**
- Особливості HTTP-набору даних CSIC 2010
- Рекомендації з наукової літератури
- Попередні дослідження, релевантні тематиці безпеки API

☑ Обрані моделі:

Назва моделі	Повна назва	Короткий опис
K-means	K-means Clustering	Кластеризація за відстанню до центроїдів
GMM	Gaussian Mixture Model	Модель змішаних нормальних розподілів
IF	Isolation Forest	Виявлення аномалій шляхом ізоляції
OCSVM	One-Class Support Vector Machine	Побудова межі для нормального класу

9

Базові моделі

```
# Базова модель K-середніх
kmeans = KMeans(n_clusters=2, random_state=42)

# Базова модель GMM (Гаусівська змішана модель)
gmm = GaussianMixture(n_components=2, random_state=42)

# Базова модель Isolation Forest (ліс ізоляції)
iso_forest = IsolationForest(n_estimators=100,
contamination=0.1, random_state=42)

# Базова модель OCSVM (однокласовий SVM)
ocsvm = OneClassSVM()
```

10

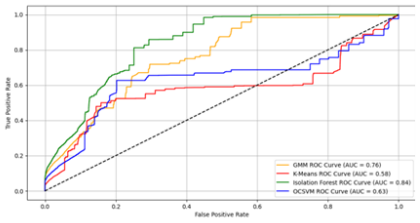
Налаштовані моделі

```
# Налаштована модель K-середніх
kmeans = KMeans(n_clusters=5, random_state=42)

# Налаштована модель GMM (Гаусівська змішана модель)
gmm = GaussianMixture(n_components=18,
random_state=42)

# Налаштована модель Isolation Forest (ліс ізоляції)
iso_forest = IsolationForest(n_estimators=300,
contamination=0.5, random_state=42)

# Налаштована модель OCSVM (однокласовий SVM)
ocsvm = OneClassSVM(kernel='rbf', nu=0.5,
gamma='auto')
```

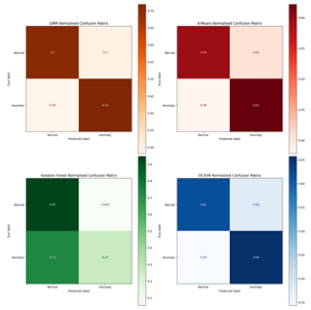
11

Показники ефективності для базових моделей

Модель	P	R	F1	AUC	A	Час (с)
GMM	0.63	0.72	0.67	0.76	0.71	0.35
K-середніх	0.52	0.62	0.56	0.58	0.61	0.12
Isolation Forest	0.78	0.27	0.40	0.84	0.67	0.18
OCSVM	0.55	0.66	0.60	0.63	0.63	32.35



Порівняння ROC-кривих для базових моделей

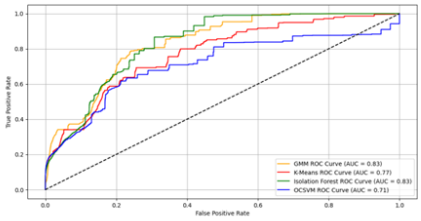


Порівняння матриць плутанини для базових моделей

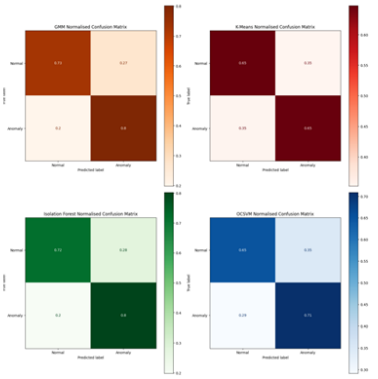
12

Показники ефективності для налаштованих моделей

Модель	P (Точність)	R (Повнота)	F1-міра	AUC	A (Точність класифікації)	Час (с)
GMM	0.66	0.79	0.71	0.83	0.74	0.75
K-середніх	0.56	0.65	0.60	0.77	0.65	0.13
Isolation Forest	0.67	0.80	0.73	0.83	0.75	0.70
OCSVM	0.59	0.71	0.64	0.71	0.67	30.84



Порівняння ROC-кривих для налаштованих моделей



Порівняння матриць плутанини для налаштованих моделей

13

Висновки

У межах кваліфікаційної роботи було реалізовано підхід до виявлення аномалій у журналах API на основі неконтрольованого машинного навчання.

- Проведено повну обробку датасету CSIC HTTP 2010, адаптованого до формату API-журналів;
- Тестовано та налаштовано 4 моделі: **K-means**, **GMM**, **Isolation Forest**, **OCSVM**;
- Оптимізація моделей (AIC, BIC, точка ліктя) значно покращила ефективність класифікації;
- Найкращі результати: **GMM (AUC = 0.83, Recall = 0.79)** та **IF (F1 = 0.73)**;
- Покращено якість кластеризації (силует, ARI, NMI).

Апробація результатів: МАРТОВИЦЬКИЙ, В., СВИРИДОВ, А., АВДЄЄВ, О., ГУДЗИНСЬКИЙ, І., & КОРОТЕЦЬКИЙ, О. (2025). ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ У API ЖУРНАЛАХ ДЛЯ ЗАБЕЗПЕЧЕННЯ БЕЗПЕКИ ТА НАДІЙНОСТІ ПРОГРАМНИХ СИСТЕМ. *Вісник Херсонського національного технічного університету*, *2*(1 (92)), 142-148.

14