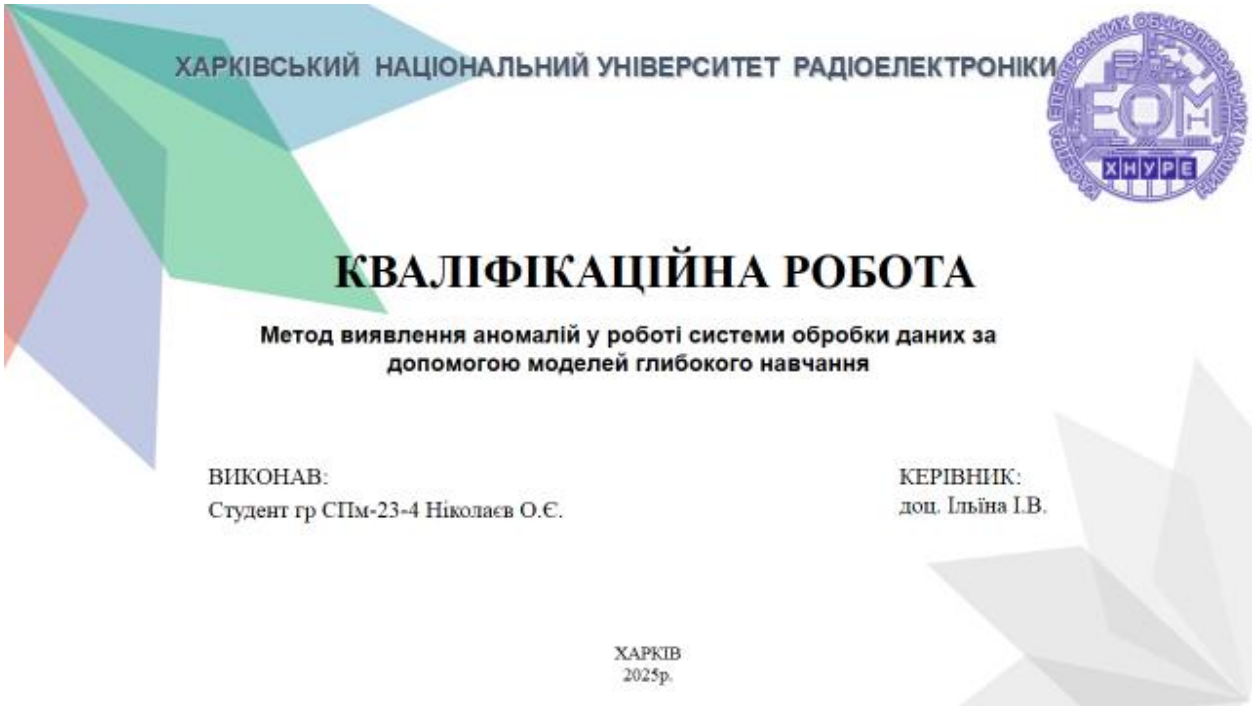
ДОДАТОК А

Графічний матеріал кваліфікаційної роботи



Актуальність дослідження

•Стрімке зростання обсягів даних у розподілених системах ускладнює ручний контроль їхньої коректності

•Виявлення аномалій є критично важливим для запобігання збоїв, втрати даних або порушень безпеки

•Традиційні методи недостатньо ефективні у динамічних середовищах із часовими залежностями

•Глибоке навчання (CNN, LSTM) дозволяє моделювати складні шаблони поведінки системи у реальному часі

•Проблематика є актуальною в контексті кібербезпеки, IT-інфраструктур та промислових IoT-систем

Мета та завдання

Мета дослідження:
Розробити та дослідити метод виявлення аномалій у роботі системи обробки даних з використанням моделей глибокого навчання, зокрема CNN та LSTM, для підвищення точності виявлення відхилень у часових рядах.

Основні завдання:

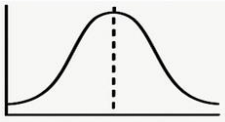
- Проаналізувати сучасні підходи до виявлення аномалій у системах обробки даних
- Вивчити можливості моделей CNN, LSTM та гібридної архітектури CNN+LSTM
- Підготувати датасет на основі часових рядів реальної системи
- Провести нормалізацію та перетворення даних у послідовності
- Побудувати моделі прогнозування та виконати гіперпараметричну оптимізацію
- Провести експерименти з виявлення аномалій та оцінити результати

3

Методи виявлення аномалій

Статистичні методи:

- порогові значення
- моделі розподілу
- кореляція

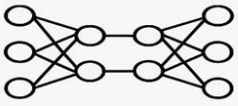


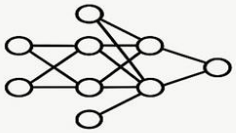
Методи машинного навчання:

- Класифікатори: SVM, дерева рішень
- Кластеризація: K-Means, DBSCAN
- Алгоритми ізоляції: Isolation Forest

Методи глибокого навчання

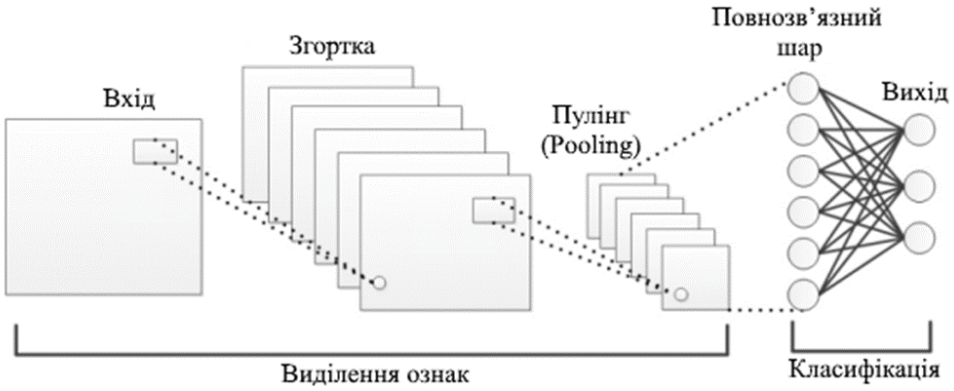
- LSTM (довготривала пам'ять) — ефективна для часових рядів
- CNN — виявлення просторових залежостей
- Гібридні моделі CNN+LSTM — поєднання просторової та часової чутливості





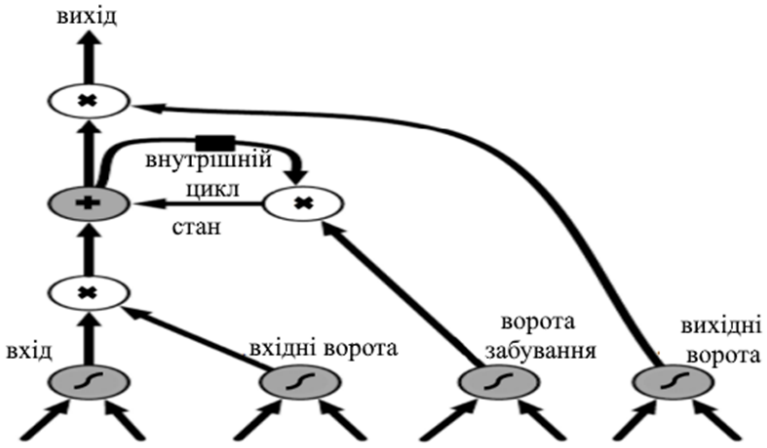
4

Архітектура CNN



5

Архітектура LSTM



6

Архітектура CNN+LSTM

7

Опис набору даних

Дані отримано з відкритого джерела (Kaggle), наданого компанією, яка реалізує хмарне рішення для обміну даними між системами підприємств.
• Система обробки даних включає **4 підсистеми**, кожна з яких відповідає за обробку певного типу даних.
• Основний набір даних — **журнали активності** підсистем з фіксацією дати, часу та API-виклику.
• Період: з **26.04.2021 по 22.03.2022**, інтервал — 1 година.
• Формат: 4,5 млн рядків, колонки — call_date_time, category.
• Кожен виклик — HTTP-запит до API: передача, оновлення або запит даних.
• Додатково: 4 окремі набори з часовими мітками **аномалій для кожної підсистеми**.

8

Опис функції groupby + size

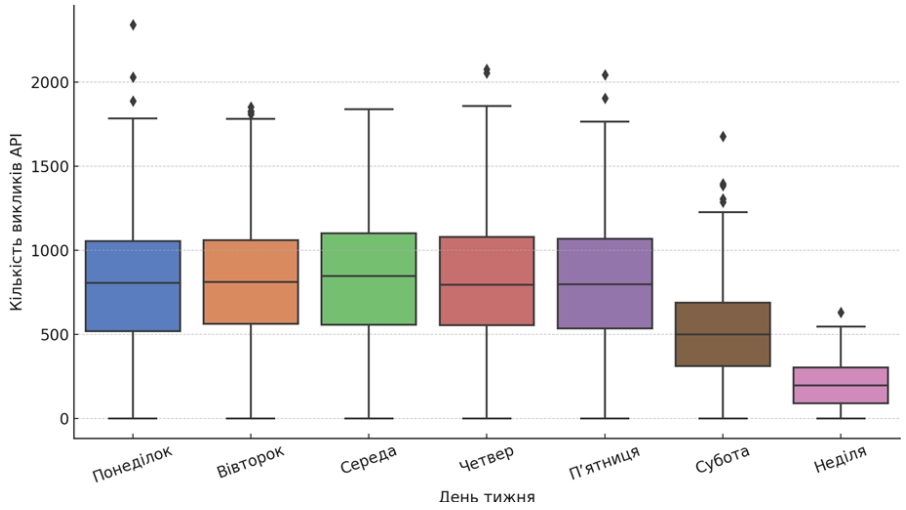
call_date_time	category
01-02-2022	Sales
01-02-2022	Items
01-02-2022	Sales
01-02-2022	Purchase
01-02-2022	Sales
01-02-2022	Items
01-03-2022	Items
01-03-2022	Sales

groupby size function

call_date_time	category	size
01-02-2022	Sales	3
01-02-2022	Items	2
01-02-2022	Purchase	1
01-03-2022	Sales	1
01-03-2022	Items	1

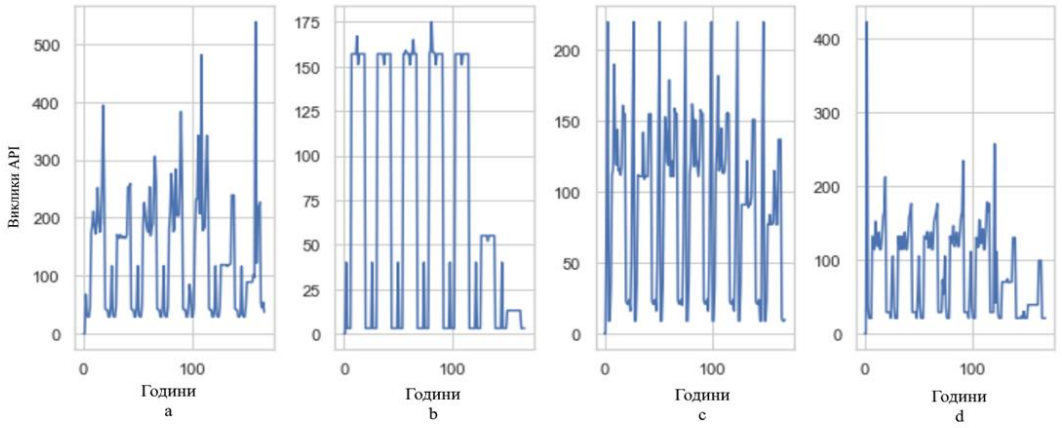
9

Сезонності за тижнем



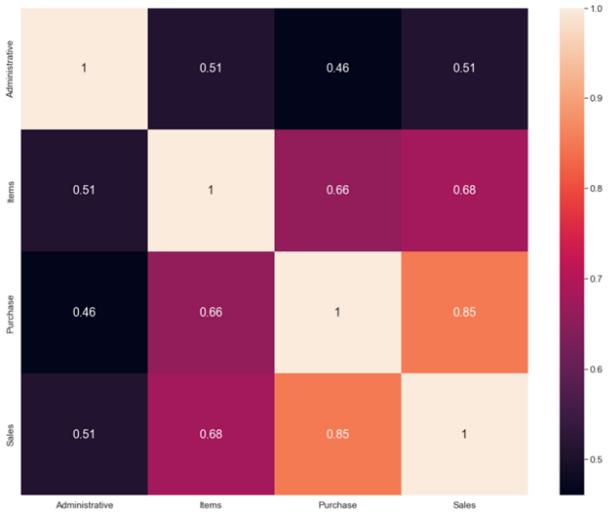
10

Лінійні графіки активності підсистем за перший тиждень



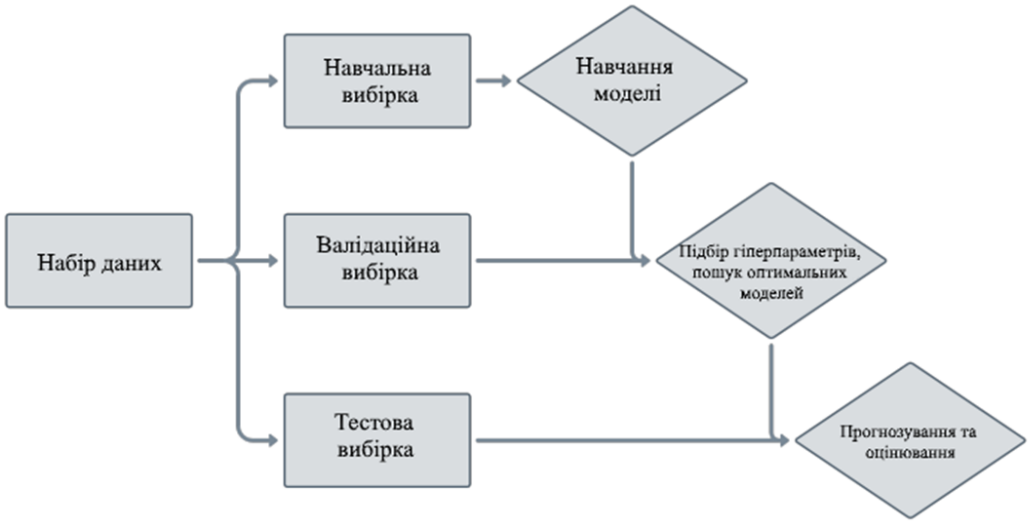
11

Взаємозв'язок між підсистемами

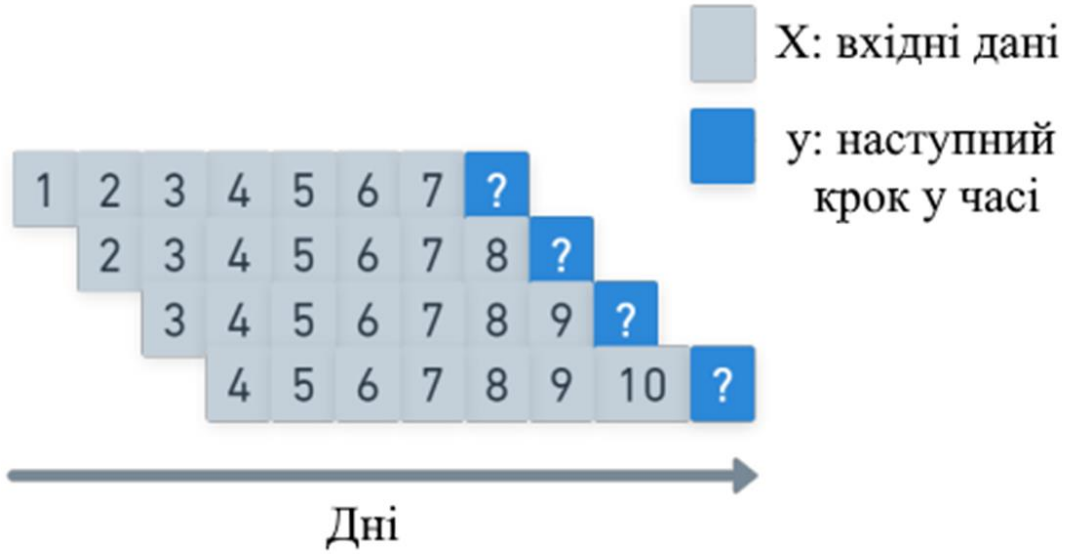


12

Розподіл даних



13

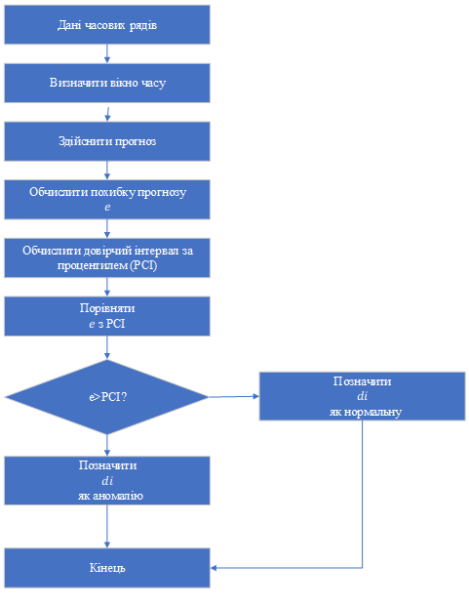


14

Перетворення часових рядів на послідовні дані з часовим вікном у 7 вхідних днів

дні дані (X)	Вихід (Прогноз y)
День 1–7	День 8, перша година
7 днів = 168 годин	169-й часовий крок
День 2–8	День 9, перша година
7 днів = 168 годин	337-й часовий крок
День 3–9	День 10, перша година
7 днів = 168 годин	504-й часовий крок

15



16

Експериментальна фаза

Складові:

1. Прогнозування наступного значення часового ряду:

- Метрики оцінки:
  - **MAE** (середня абсолютна похибка)
  - **RMSE** (корінь середньоквадратичної похибки)

2. Виявлення аномалій на основі похибки прогнозу:

- Метрики оцінки:
  - **AUC** (площа під кривою)
  - **TPR** (істинно позитивна частка)
  - **FPR** (хибнопозитивна частка)
  - **Accuracy** (точність)
  - **ROC-крива**

Базова модель:

MA (Moving Average) — використовується як еталонна модель.
Прогнозує значення на основі середнього значення за попередні 7 днів.

17

Модель LSTM

Шар (тип)	Форма виходу	Кількість параметрів
LSTM	(None, 100)	42 000
RepeatVector	(None, 1, 100)	0
LSTM	(None, 1, 100)	80 400
TimeDistributed	(None, 1, 4)	404

Модель CNN+LSTM

Шар (тип)	Форма виходу	Кількість параметрів
Conv1D	(None, 22, 64)	832
Conv1D	(None, 20, 64)	12 352
MaxPooling1D	(None, 10, 64)	0
Flatten	(None, 640)	0
RepeatVector	(None, 1, 640)	0
LSTM	(None, 1, 150)	474 600
TimeDistributed	(None, 1, 100)	15 100
TimeDistributed	(None, 1, 4)	404

18

Середнє значення MAE для 4 підсистем з історичним введенням за 1, 2, 5, 7 днів.

Метод	1 день	2 дні	5 днів	7 днів
MA	0.1618	0.1711	0.1807	0.1836
LSTM	0.1483	0.1501	0.1250	0.1296
CNN + LSTM	0.0949	0.1056	0.1241	0.1305

Середнє значення RMSE для 4 підсистем з історичним введенням за 1, 2, 5, 7 днів.

Метод	1 день	2 дні	5 днів	7 днів
MA	0.2196	0.2323	0.2432	0.2485
LSTM	0.3482	0.2659	0.2173	0.2412
CNN + LSTM	0.1824	0.1967	0.2211	0.2329

Оцінок ефективності

Метод	AUC	TPR (Чутливість)	FPR (Хибнопозитивна частка)	Точність (Accuracy)
MA	0.7639	0.5341	0.0198	0.9762
LSTM	0.7844	0.5215	0.0183	0.9774
CNN + LSTM	0.9046	0.7996	0.0178	0.9804

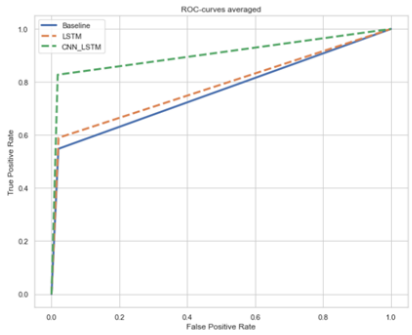


Рисунок – ROC-криві для всіх моделей

Висновки

У межах роботи реалізовано:
•Аналіз підсистем та формування часових рядів;
•Побудову моделей MA, LSTM, CNN+LSTM;
•Виявлення аномалій на основі похибки прогнозу;
•Оцінку моделей за метриками MAE, RMSE, AUC, TPR, FPR, ROC.

Ключові висновки:
1.CNN+LSTM — найефективніша модель;
2.Історичний обсяг даних впливає по-різному на точність;
3.Метод виявлення аномалій на основі PCI — ефективний;
4.Гібридна модель збалансована за точністю, стабільністю, масштабованістю;
5.Підходить для впровадження в промислові IT-системи.

Апробація результатів: Шандиба А. С., Ніколаєв О.Є Сітніков В. І Нанесення цифрових водяних знаків з використанням хаотичних карт Системи управління, навігації та зв'язку. 2024. № 4 с. 208-214