

ПОРІВНЯННЯ МЕТОДІВ ІЗ АВТОМАТИЗОВАНОГО ВИЯВЛЕННЯ АНОМАЛІЙ У МЕДИЧНІЙ СТАТИСТИЦІ

¹Вінницький національний технічний університет

Анотація

Проведено тестування двох методів пошуку аномалій із метою порівняння їхніх можливостей. Для цього було виконане тестування із пошуку аномалій на медичному датасеті для двох методів Isolation Forest та методу із використанням великих мовних моделей. Отримані результати були проаналізовані та представлені у вигляді таблиці, а також запропоновано можливі варіанти для покращення пошуку аномалій з використанням LLM у медичній статистиці.

Ключові слова: медична статистика, пошук аномалій, великі мовні моделі, LLM.

Abstract

Testing of two anomaly detection methods was conducted to compare their capabilities. For this purpose, anomaly detection was performed on a medical dataset using two methods: Isolation Forest and a method based on large language models. The obtained results were analyzed and presented in a table, and possible improvements for anomaly detection using LLMs in medical statistics were proposed.

Keywords: medical statistics, anomaly detection, large language models, LLM.

Вступ

Виявлення аномальних відхилень у медичній статистиці, що відображає динаміку захворюваності може значною мірою допомогти швидко і своєчасно реагувати на загрози епідеміологічного характеру, покращити моніторинг здоров'я населення, а також підвищити якість та ефективність управління в медичній сфері [1]. Ці аномалії також можуть сигналізувати про потенційні помилки в медичних установах чи регіональних спостереженнях. Тому забезпечення своєчасного виявлення таких відхилень є критично важливим завданням для систем охорони здоров'я, особливо в умовах обмежених ресурсів і необхідності швидкого реагування на потенційні загрози.

Метою даного дослідження є розроблення методів виявлення аномалій у медичній статистиці на основі типових закономірностей захворюваності за допомогою великих мовних моделей (Large Language Models, LLM).

Актуальність дослідження

Існуючі методи виявлення аномалій, такі як статистичні методи або алгоритми машинного навчання, орієнтовані переважно на числові дані або структуровані дані [2]. Це обмежує їх застосування в медичній сфері, де значну частину інформації становлять неструктуровані тексти. Багато методів також не враховують специфіку сезонних коливань захворюваності або регіональних особливостей, що є критично важливими для точного аналізу. Крім того, більшість класичних підходів вимагає значних обчислювальних ресурсів, мають обмежену здатність до адаптації під нові закономірності або аномалії та виявляють слабкі результати при аналізі великих обсягів текстових даних, які містять контекстуальну інформацію про динаміку захворювань [3].

Для вирішення вказаних проблем пропонується використати великі мовні моделі, які показали хорошу ефективність в аналізі та пошуку закономірностей у неструктурованих текстових даних [4]. LLM дозволяють автоматично витягувати інформацію з текстів, аналізувати її та формалізувати як типові послідовності захворюваності для певних умов. Завдяки здатності обробляти великі обсяги медичних текстів та аналізувати зміст на основі контексту, LLM можуть виділяти регіональні та сезонні закономірності захворюваності, які потім слугуватимуть еталоном для порівняння з фактичними статистичними даними. Це дозволить підвищити точність та ефективність виявлення аномалій, адаптуючись до нових закономірностей та тенденцій захворюваності.

Результати дослідження

Для проведення дослідження було вибрано два поширених методи виявлення аномалій. Перший – Isolation Forest, а другий – метод із використанням великих мовних моделей.

Isolation Forest – популярний метод, що працює за допомогою побудови випадкових дерев і був спеціально розроблений для виявлення аномалій. При його використанні, аномалії зазвичай швидше стають помітні через коротші шляхи в дереві. Цей метод добре показує себе із великими обсягами даних, швидкий та доволі легко інтерпретується. Він досить простий у використанні та може бути використаний для різноманітних типів медичних показників. До недоліків цього методу можна віднести меншу ефективність при виявленні дуже незначних або складних нелінійних залежностей та роботу тільки із числовими значеннями [5].

Аналіз із застосуванням великих мовних моделей, зокрема моделей на основі GPT, типово застосовуються для обробки текстових даних, але останнім часом їхня популярність зросла і можна зустріти їхнє використання в тому числі і для аналізу великих масивів даних і пошуку відхилень в них. У медичній сфері це додатково дозволяє інтегрувати як числові дані (наприклад, вимірювання фізіологічних показників), так і неструктуровану текстову інформацію (опис симптомів, для прикладу). До недоліків цього підходу можна віднести високі апаратні вимоги, що може бути не завжди доцільно із економічної сторони питання [6].

Тестування відбувалось на платформі Kaggle з використанням медичного датасету на 10000 рядків, що містять в собі числові та текстові значення [7]. Фрагмент такого датасету наведений на рисунку 1.

	temperature	blood_pressure	oxygen_level	symptom_description
0	37.186188	125.926768	98.507687	The patient feels well.
1	37.063958	125.253253	98.091704	No symptoms, the condition is satisfactory.
2	37.263606	119.952312	97.578218	The patient feels well.
3	36.814614	116.814811	97.952944	No symptoms, the condition is satisfactory.
4	37.875222	119.864208	99.081223	The patient feels well.

Рис. 1 – Фрагмент датасету

Частина значень цього датасету мають аномальні значення для числових і текстових даних, а інша частина – тільки для текстових. Розподіл числових значень можна побачити на рисунку 2. Для кожного запуску було підраховано оцінку F1-score, яка дозволяє краще оцінити ефективність моделі в умовах коли аномалій набагато менше в порівнянні із нормальними значеннями.

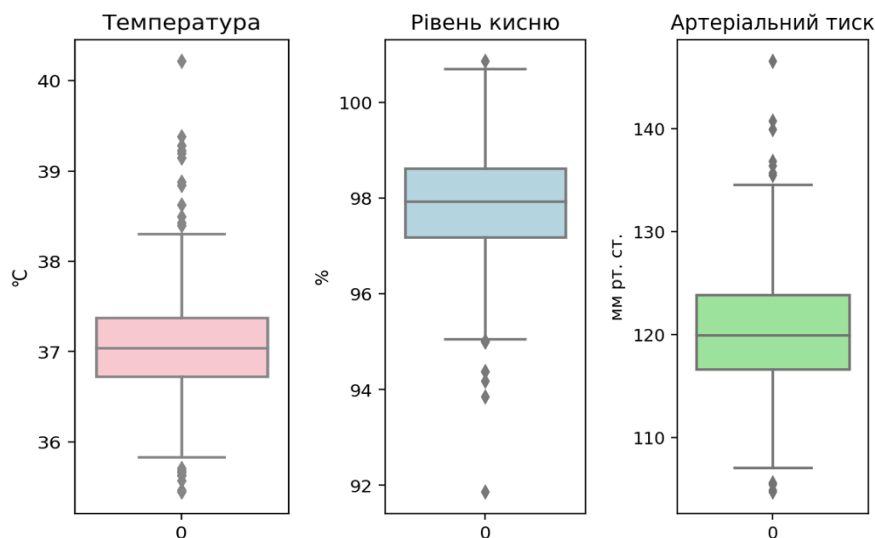


Рис. 2 – Розподіл числових даних в датасеті

Для Isolation Forest були підібрані параметри для досягнення найвищої оцінки. Параметр contamination був встановлений в значення 0.01, а також параметр random_state для відтворюваності запусків. Для оцінювання цього методу було використано тільки числові дані.

Для другого тесту було використано модель gpt-4o-mini, яка була розроблена компанією OpenAI. Вона позиціонується як невелика та доступна модель для щоденних завдань. Для цього був використаний публічний прикладний програмний інтерфейс (Application Programming Interface, API) платформи OpenAI для масової обробки запитів (Batch API). Тест проводився із використанням zero-shot підходу (без попередніх прикладів) для запитів (prompts) до моделі із використанням як числових так і текстових даних. Приклад запиту, який використовувався у тестуванні, можна побачити на рисунку 3.

```
def generate_prompt(row):  
    return f"Determine whether the following description of symptoms indicates an anomaly (i.e., abnormal) or a normal state.  
  
Temperature: {row['temperature']},  
Blood pressure: {row['blood_pressure']},  
Oxygen level: {row['oxygen_level']},  
Symptoms: \"{row['symptom_description']}\"  
  
Answer with a single word: \"anomaly\" or \"normal\"."
```

Рис. 3 – Приклад запиту (prompt)

В ході тестування, окрім запуску із стандартними налаштуваннями моделі, робились спроби у налаштуванні різних параметрів моделі для досягнення кращих результатів. Найкращий результат показав параметр temperature або його альтернатива у вигляді top_p. Цей параметр контролює варіативність при генерації відповідей. Його зниження дозволяє зробити роботу моделі більш сфокусованою та детермінованою, а підвищення приведе до більш креативних та непередбачуваних відповідей. Цей параметр може змінюватись в діапазоні від 0 до 2. За замовчуванням він має значення одиниці. Результати виконання тестів наведені в таблиці 1. З отриманих результатів можна зробити висновок, що у випадку зміни параметру temperature, поточний метод зміг визначити більше аномалій в датасеті, що вказує на те, що були враховані не тільки числові дані, а й текстові.

Таблиця 1. Результати тестування підходів для пошуку аномалій

Метод пошуку аномалій	F1-score
Isolation Forest	0.60
GPT-4o mini (temperature=1)	0.67
GPT-4o mini (temperature=0.2)	0.95

Висновки

В результаті проведеного тестування були отримані результати, які дозволяють оцінити точність визначення аномалій в медичних даних двома способами. З них можна зробити висновки, що метод із використанням Isolation Forest показує себе лише дещо гірше за метод із використанням великої мовної моделі. Хоча цей результат і стосується лише тесту із стандартними налаштуваннями. Зі зниженням параметру temperature модель GPT-4o mini почала показувати відчутно кращі результати. Це показує, що існує запас для ширших експериментів із її налаштування, який може давати кращі результати навіть без додаткового навчання чи few-shot підходу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Eze, Peter U., Nicholas Geard, Ivo Mueller, and Iadine Chades. 2023. "Anomaly Detection in Endemic Disease Surveillance Data Using Machine Learning Techniques" Healthcare 11, no. 13: 1896. <https://doi.org/10.3390/healthcare11131896>
2. Nataliia Melnykova, Roman Kulievych, Yaroslav Vycluk, Kateryna Melnykova, Volodymyr Melnykov, Anomalies Detecting in Medical Metrics Using Machine Learning Tools, Procedia Computer Science, Volume 198, 2022, Pages 718-723, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2021.12.312>

3. Samariya, D., Ma, J., Aryal, S. et al. Detection and explanation of anomalies in healthcare data. Health Inf Sci Syst 11, 20 (2023). <https://doi.org/10.1007/s13755-023-00221-2>.
4. Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario Gregori, Paola Berchiarella, Information extraction from medical case reports using OpenAI InstructGPT, Computer Methods and Programs in Biomedicine, Volume 255, 2024, 108326, ISSN 0169-2607, <https://doi.org/10.1016/j.cmpb.2024.108326>.
5. Д. О. Шмундяк, Н. С. Іжаковська, Д. О. Литвиненко, А. О. Судець, «Аналіз можливостей Python-бібліотек щодо виявлення аномальних даних у задачі прогнозування стану атмосферного повітря», на ЛІІ Науково-технічна конференція факультету інтелектуальних інформаційних технологій та автоматизації, Вінниця, 21 - 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18957/15722>
6. Li, A., Zhao, Y., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., & Mandt, S., 2024, Anomaly detection of tabular data using llms, <http://dx.doi.org/10.48550/arXiv.2406.16308>
7. Kaggle Dataset “Medical Dataset” — Бобко Б. В., 2025, <https://www.kaggle.com/datasets/bohdbanbobko/medical-dataset/data>

Бобко Богдан Володимирович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: bobko.bogdan@gmail.com

Жуков Сергій Олександрович — кандидат технічних наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: sazhukov@gmail.com

Bobko Bohdan V. – Postgraduate student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: bobko.bogdan@gmail.com

Zhukov Serhii O. – Candidate of technical sciences, associate professor of the department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com