

ВИЯВЛЕННЯ СТРУКТУРНИХ І СЕМАНТИЧНИХ АНОМАЛІЙ У МЕДИЧНИХ ЗАПИСАХ ЗАСОБАМИ ГЕТЕРОГЕННИХ ГРАФОВИХ НЕЙРОННИХ МЕРЕЖ

¹Вінницький національний технічний університет

Анотація

Роботу присвячено задачі виявлення аномалій у записах системи охорони здоров'я України. Запропоновано метод побудови гетерогенного графа медичних даних із вузлами типів: пацієнт, ICD-10, ICP-2, АСНІ, заклад та направлення. На датасеті (281 167 записів, 192 799 пацієнтів, 149 закладів) побудовано граф з понад 5 млн ребер та виявлено два рівні аномалій: структурний (пацієнти зі структурно нетиповим профілем зв'язності у графі, виявлені GAE) та семантичний (незвичайні поєднання кодів, виявлені HetGNN). Коефіцієнт Спірмена між рейтингами аномальності моделей становив $\rho = -0,010$, що свідчить про практично відсутню рангову узгодженість результатів GAE та HetGNN і вказує на виявлення ними різних аспектів аномальності (перетин top-100 аномалій порожній, Jaccard = 0).

Ключові слова: виявлення аномалій, електронні медичні записи, гетерогенні графові нейронні мережі.

Abstract

The paper addresses anomaly detection in Ukraine's electronic health records (EHR). A heterogeneous graph of medical data is constructed with node types: patient, ICD-10, ICP-2, ACHI, facility, and referral. Applied to a dataset (281,167 records, 192,799 patients, 149 facilities), the graph contains over 5 million edges. Two anomaly levels were identified: structural (patients structurally atypical in their graph-connectivity profile, detected by GAE), and semantic (unusual code co-occurrences, detected by HetGNN). The Spearman correlation coefficient between the models' anomaly rankings was $\rho = -0.010$, indicating virtually no rank-order agreement between the GAE and HetGNN results and suggesting that the models identify different aspects of anomalousness (the top-100 anomaly sets are disjoint, Jaccard = 0).

Keywords: anomaly detection, electronic health records, heterogeneous graph neural networks.

Вступ

Впровадження електронних медичних записів (ЕМЗ) у систему Національної служби здоров'я України (НСЗУ) відкрило можливості для масштабного аналізу медичних даних, однак водночас поставило задачу автоматизованого виявлення аномалій та помилок кодування. ЕМЗ НСЗУ містять структуровану інформацію із застосуванням трьох міжнародних систем класифікації: ICD-10 — для кодування діагнозів; ICP-2 — для фіксації причин звернення пацієнта впродовж року до встановлення діагнозу; АСНІ — для реєстрації процедур. Методи машинного навчання є ефективним інструментом для виявлення аномалій у медичній статистиці, де традиційний підхід ускладнено дефіцитом розмічених даних [1, 2].

Графові нейронні мережі (GNN) — клас методів глибокого навчання, що безпосередньо опрацьовує дані у вигляді графів. На відміну від класичних архітектур (CNN, RNN), GNN здійснюють ітеративне передавання повідомлень між пов'язаними вузлами: кожен вузол оновлює свій прихований стан, агрегуючи інформацію від сусідів. Це робить GNN особливо придатними для аналізу медичних записів, де коди ICD-10, ICP-2 та АСНІ пов'язані клінічними залежностями: симптом (ICP-2) передуює діагнозу (ICD-10), а діагноз визначає набір процедур (АСНІ). Аномальні записи отримують ізольоване представлення у просторі ембедингів, що слугує сигналом для виявлення. GNN успішно застосовуються для ієрархічного кодування ICD-10 [3], моделювання гетерогенних структур ЕМЗ [5] та виявлення аномальних патернів у задачах медичного аналізу даних [2].

Метою роботи є розроблення та апробація методу виявлення аномалій у медичних записах НСЗУ на основі гетерогенного графа, що відображає зв'язки ICP-2 → ICD-10 → АСНІ, а також комплексне виявлення структурних та семантичних аномалій в єдиному аналітичному фреймворку.

Результати дослідження

Для дослідження використано датасет, що містить 281 167 записів від 192 799 унікальних пацієнтів з 149 медичних закладів. Медичні записи перетворено у гетерогенний граф $G = (V, E)$ із 6 типами вузлів та 8 типами ребер (рис. 1); характеристики наведено у табл. 1.

Таблиця 1. Характеристики гетерогенного графа EM3. (*) — ключовий сигнал аномалії

Тип вузла	Кількість вузлів	Тип ребра	Кількість ребер
Пацієнти	192 799	has_symptom (пацієнт → ICPC-2)	997 126
ICPC-2	1 073	diagnosed_with (пацієнт → ICD-10)	279 767
ICD-10	43	preceded (*) (ICPC-2 → ICD-10)	18 468
АСНІ	3 990	received (пацієнт → АСНІ)	3 119 313
Заклади	149	at (пацієнт→заклад)	202 700
Направлення	1 688	treated_by (ICD-10→АСНІ)	48 284
—	—	generated_completed (АСНІ → напр.)	590 148
—	—	generated_unfulfilled (АСНІ → напр.)	447 667

Інспекція ребер preceded на основі правил виявила 17 980 з 18 468 пар (97,4 %) зі структурною невідповідністю. Аналіз показав, що значна частина таких випадків пов'язана з компонентними кодами ICPC-2, які позначають направлення у вторинну ланку, а не клінічний симптом. Це свідчить не лише про наявність потенційних невідповідностей у даних, а й про необхідність формування формальної таблиці відповідності ICPC-2 → ICD-10 для зменшення кількості хибнопозитивних спрацювань.

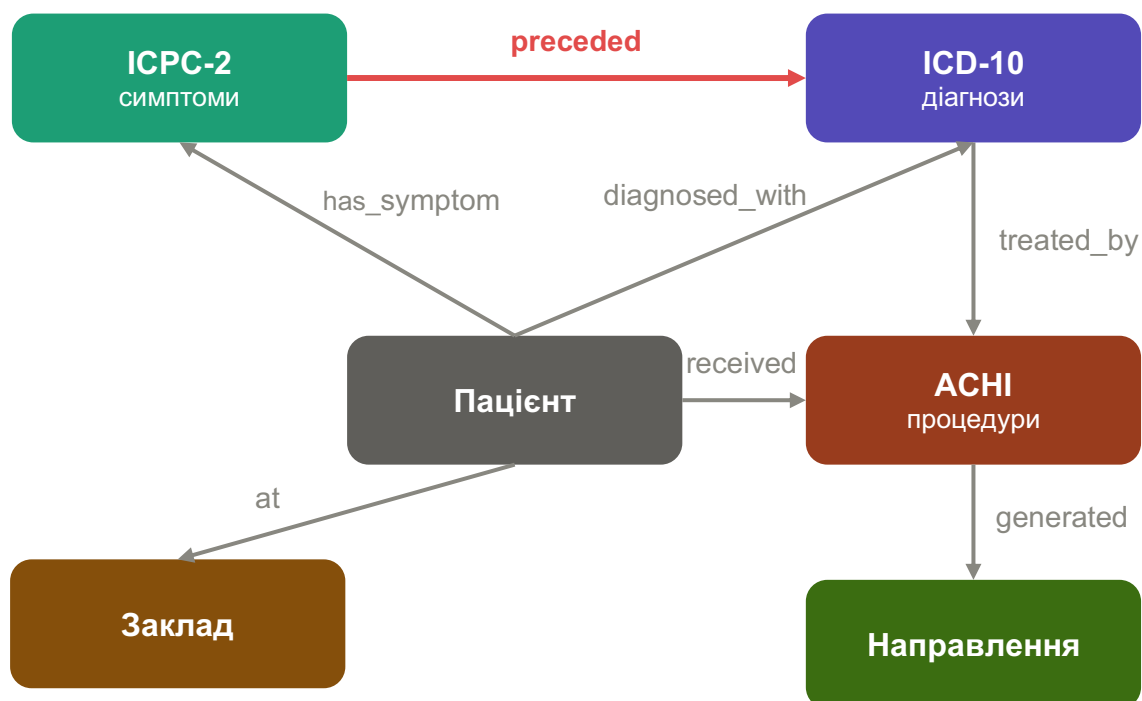


Рисунок 1 – Схема гетерогенного графа EM3: 6 типів вузлів, 8 типів ребер; ребро preceded – ключовий сигнал аномалії кодування

GAE [4] навчається на k-NN графі пацієнтів ($k = 15$, 2 891 985 ребер) з 8-вимірними ознаками (вік, стать, циклічно закодований місяць діагнозу (sin/cos) та чотири показники зв'язності у графі — кількість пов'язаних кодів ICD-10, ICPC-2, АСНІ та закладів), оптимізуючи комбіновану функцію втрат реконструкції суміжності та ознак. Оцінка аномальності — локальна відносна помилка реконструкції ознак:

$$\text{score}(v) = \log \left(1 + \frac{e(v)}{\mu_k(v)} \right), \quad (1)$$

де $e(v) = \|x_v - \hat{x}_v\|^2$ — помилка реконструкції 8-вимірного вектора ознак вузла v декодером із латентного представлення (x_v — вектор ознак вузла: вік, стать, циклічно закодований місяць діагнозу (sin/cos) та чотири показники зв'язності у графі; \hat{x}_v — його реконструкція декодером); $\mu_k(v)$ — середня помилка реконструкції серед $k = 15$ найближчих сусідів вузла v у просторі ознак. Перед використанням значення (1) додатково нормалізується у діапазон $[0, 1]$. Формула виявляє пацієнтів, чия помилка реконструкції суттєво перевищує помилку їхнього найближчого оточення, тобто структурно нетипових у локальному контексті.

HeteroGNN на основі HeteroConv із агрегацією GraphSAGE для кожного типу ребра гетерогенного графа (сім із восьми типів зв'язків — received виключено через велику кількість ребер (~3,1 млн), прихований вимір 64, латентний вимір 32) навчається відновлювати вектор ознак пацієнта; оцінка аномальності — Local Outlier Factor (LOF, $k = 50$) у нормалізованому просторі отриманих ембедингів пацієнтів.

GAE (200 епох, прихований вимір 64, латентний 32) ідентифікував як найбільш аномальні пацієнтів з показником $\text{score} = 1,000$, у яких зафіксовано ICD-10 діагнози з діагностичних глав L (шкіра), N (сечостатева), G (нервова) та M (опорно-рухова). Через масовий збіг максимальних оцінок ($\text{score} = 1,000$) GAE радше виокремлює структурно нетипових пацієнтів за профілем зв'язності у графі, ніж стабільно ранжує окремі випадки. HetGNN виявив аномальні семантичні поєднання: J00–J99 (дихальна система) з O24 (гестаційний діабет) або D50–D53 (анемія); код F10 (розлади внаслідок вживання алкоголю) присутній у 6 з 10 найбільш аномальних пацієнтів за версією HetGNN. Коефіцієнт Спірмена між рейтингами моделей становив $\rho = -0,010$ ($p < 0,001$), що свідчить про низьку узгодженість результатів GAE та HetGNN. Оскільки в задачі виявлення аномалій вирішальною є узгодженість саме на вершині рейтингу, а не глобальна рангова кореляція (яку розмиває велика маса типових пацієнтів), додатково оцінено перетин множин найаномальніших пацієнтів. Він виявився порожнім: $\text{Jaccard}@10 = \text{Jaccard}@50 = \text{Jaccard}@100 = 0,000$, тобто GAE та HetGNN не поділяють жодного спільного випадку серед 100 найаномальніших пацієнтів. Це може вказувати на потенційну взаємодоповнюваність моделей: GAE краще виявляє структурні аномалії документування, тоді як HetGNN фіксує семантичні відхилення у поєднаннях медичних кодів.

Висновки

У роботі запропоновано та апробовано метод виявлення аномалій у медичних записах на основі гетерогенних графових нейронних мереж. У використаному датасеті було виявлено два рівні аномалій: структурний (пацієнти зі структурно нетиповим профілем зв'язності у графі, виявлені GAE) та семантичний (незвичайні поєднання кодів F10, J00–J99 з O24/D50–D53, виявлені HetGNN). Значення $\rho = -0,010$ та порожній перетин top-k аномалій ($\text{Jaccard} = 0$) свідчать про практично відсутню рангову узгодженість і взаємодоповнюваність моделей. У поєднанні з якісним аналізом виявлених випадків це може вказувати на їхню потенційну взаємодоповнюваність.

Результати відкривають додаткові напрями для подальшої роботи: розроблення формальної таблиці відповідності ICPC-2 → ICD-10 на основі SNOMED CT/UMLS для точнішого визначення аномальних пар кодів; розроблення федеративного варіанта методу (FL-GNN) для виявлення міжзакладових аномалій без централізації персональних даних у мережі зі 149 закладів; інтеграція методів пояснюваності, зокрема SHAP та GNN Attribution, як необхідна умова практичного впровадження.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. De Meulemeester H., De Smet F., van Dorst J., Derroitte E., De Moor B. Explainable unsupervised anomaly detection for healthcare insurance data. BMC Medical Informatics and Decision Making. 2025. Vol. 25, No. 1. Article 14. DOI: <https://doi.org/10.1186/s12911-024-02823-6>

2. Бобко Б. В., Жуков С. О. Методи машинного навчання для виявлення аномалій у медичній статистиці України: аналітичний огляд. Вінницький національний технічний університет. URL: <https://ir.lib.vntu.edu.ua/handle/123456789/49988>
3. Xi S., Shi J., Yan J., Lin M., Zhou X., Cheng Y., Ding H., Kang C.C. Breaking barriers in ICD classification with a robust graph neural network for hierarchical coding. Scientific Reports. 2025. Vol. 15. Article 25676. DOI: <https://doi.org/10.1038/s41598-025-10590-1>
4. Kipf T.N., Welling M. Variational Graph Auto-Encoders. arXiv preprint. 2016. arXiv:1611.07308. DOI: <https://doi.org/10.48550/arXiv.1611.07308>
5. Hu Z., Dong Y., Wang K., Sun Y. Heterogeneous Graph Transformer. Proceedings of The Web Conference 2020 (WWW '20). ACM, 2020. P. 2704–2710. DOI: <https://doi.org/10.1145/3366423.3380027>

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