УДК 004.89:311.2:616-036.22

Б. В. Бобко[1]
С. О. Жуков[1]

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

[1]Вінницький 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, ICPC-2, ACHI, заклад та направлення. На датасеті (281 167 записів, 192 799 пацієнтів, 149 закладів) побудовано граф з понад 5 млн ребер та виявлено три рівні аномалій: системний (43,1 % невиконаних направлень, 51 код ACHI без виконань), структурний (пацієнти без ICPC-2-симптоматики за наявних ICD-10 діагнозів, виявлені GAE) та семантичний (незвичайні поєднання кодів, виявлені HetGNN). Коефіцієнт Спірмена між моделями $\rho = -0{,}012$, що підтверджує їхню взаємодоповнюваність.

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 is constructed with six node types (patients, ICD-10, ICPC-2, ACHI, facilities, referrals) and eight edge types. Applied to a dataset (281,167 records, 192,799 patients, 149 facilities), the graph contains over 5 million edges. Three anomaly levels were identified: system-level (43.1% unfulfilled referrals; 51 ACHI codes with exclusively unfulfilled referrals), structural (patients with ICD-10 diagnoses but no ICPC-2 symptom history, detected by GAE), and semantic (unusual code co-occurrences, detected by HetGNN). Spearman $\rho = -0.012$ between the two models confirms their complementarity.

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; ICPC-2 — для фіксації причин звернення пацієнта впродовж року до встановлення діагнозу; ACHI — 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, ICPC-2 та ACHI пов'язані клінічними залежностями: симптом (ICPC-2) передує діагнозу (ICD-10), а діагноз визначає набір процедур (ACHI). 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 ICPC-2 → ICD-10 → ACHI, а також комплексне виявлення структурних, семантичних і системних 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 ЕМЗ. (*) — ключовий сигнал аномалії; (†) — частка 43,1 % від сумарних направлень

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
ACHI	3 990	received (пацієнт → ACHI)	3 119 313
Заклади	149	at (пацієнт→заклад)	202 700
Направлення	1 688	treated_by (ICD-10→ACHI)	48 284
—	—	generated_completed (ACHI → напр.)	590 148
—	—	generated_unfulfilled (†) (ACHI → 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 (символ «*»), 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.

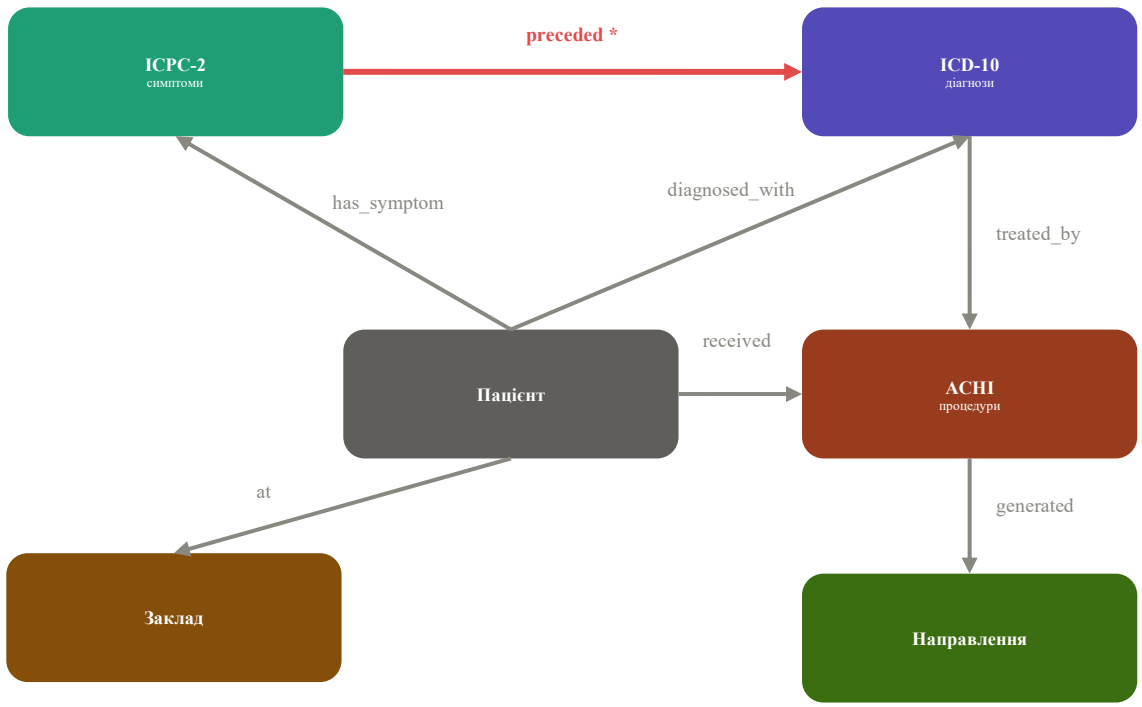

Рисунок 1 – Схема гетерогенного графа ЕМЗ: 6 типів вузлів, 8 типів ребер; ребро preceded (*) – ключовий сигнал аномалії кодування

GAE/VGAE [4] навчається відновлювати матрицю суміжності графа пацієнтів (2 891 985 ребер); оцінка аномальності:

$$score(v) = \left\|A_v - \hat{A}_v\right\|_2 + \lambda \cdot \|x_v - \hat{x}_v\|_2, \quad (1)$$

де A_v — стовпець реальної матриці суміжності для вузла v (з ким пацієнт пов'язаний у графі); $\hat{A}_v$ — стовпець відновленої матриці (очікуване сусідство за нормальним патерном); x_v — реальні ознаки вузла (вік, стать, місяць діагнозу); $\hat{x}_v$ — ознаки, відновлені декодером із латентного вектора; λ — ваговий коефіцієнт балансу між складовими ($\lambda = 1$ у поточній реалізації). Перший доданок вимірює структурну аномалію зв'язків пацієнта у графі, другий — відхилення його ознак від норми.

HetGNN на основі HGT [5] визначає окрему матрицю W_r для кожного типу ребра та агрегує повідомлення від усіх відношень; оцінка — відстань до центроїда кластера K-means.

GAE (200 епох, прихований вимір 64, латентний 32) ідентифікував як найбільш аномальні пацієнтів з показником score = 1,000, у яких зафіксовано ICD-10 діагнози (шкіра L, сечостатева N, нервова G, опорно-рухова M), проте відсутня будь-яка ICPC-2-симптоматика за попередній. Такий результат можна інтерпретувати як структурну прогалину в первинній медичній документації. 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 5 з 10 найбільш аномальних пацієнтів ансамблю. Коефіцієнт Спірмена між рейтингами моделей становив $\rho = -0{,}012$ ($p < 0{,}001$), що свідчить про низьку узгодженість результатів GAE та HetGNN. 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 три рівні аномалій: системний (43,1 % невиконаних направлень; 51 код ACHI без виконань), структурний (пацієнти без ICPC-2 при наявному ICD-10, виявлені GAE) та семантичний (незвичайні поєднання кодів F10, J00–J99 з O24/D50–D53, виявлені HetGNN). Значення $\rho = -0{,}012$ підтверджує взаємодоповнюваність моделей та цінність ансамблевого підходу.

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; клінічна верифікація виявлених аномалій; 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