ВІСНИК ПРИАЗОВСЬКОГО ДЕРЖАВНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ
2025р. Серія: Технічні науки Вип. 52
p-ISSN: 2225-6733; e-ISSN: 2519-271X

УДК 004.75+004.932.2:616 DOI: 10.31498/2225-6733.52.2025.350990

МЕТОДИ DATA MINING ТА МАШИННЕ НАВЧАННЯ ЯК ЗАСОБИ АНАЛІЗУ СТАНУ ПАЦІЄНТІВ ТА УСКЛАДНЕНЬ ЗАХВОРЮВАНЬ ЗА КЛІНІЧНИМИ ДАНИМИ

Білявенко Л.В. *аспірант, молодший науковий співробітник, Інститут інформаційних технологій та систем НАН України, м. Київ, ORCID: https://orcid.org/0009-0003-0602-1072, e-mail: leonidtill@gmail.com;*

Коваленко О.С. *д-р мед. наук, професор, ORCID: https://orcid.org/0000-0001-6635-0124*

У статті досліджено можливості застосування методів Data Mining та машинного навчання для аналізу клінічних даних у контексті складних медичних станів, що характеризуються високою гетерогенністю перебігу та підвищеним ризиком ускладнень. Актуальність роботи зумовлена стрімкою цифровізацією системи охорони здоров'я, зростанням обсягів медичної інформації, а також необхідністю впровадження персоналізованих підходів до діагностики, лікування та реабілітації пацієнтів. Особливу увагу приділено аналізу даних у педіатричній практиці та ендокринології, де традиційні статистичні методи не завжди є достатніми для виявлення складних нелінійних залежностей. Об'єктом дослідження стали дві клінічні когорти: діти віком від шести місяців до шістнадцяти років із захворюваннями дихальної системи на тлі посттравматичного стресового розладу та дорослі пацієнти з цукровим діабетом 1-го типу з різними ускладненнями перебігу захворювання. У межах дослідження застосовано комплексний підхід, що поєднує методи дескриптивної статистики з алгоритмами машинного навчання, зокрема ансамблевими моделями, методами оцінки важливості ознак та кластерним аналізом. Для педіатричної вибірки проаналізовано взаємозв'язки між тяжкістю бронхіту, частотою гострих респіраторних вірусних інфекцій, показниками посттравматичного стресу, розладами сну, інтегральним показником якості життя та інтенсивністю бойових дій у регіоні проживання. Для когорти пацієнтів із цукровим діабетом 1-го типу виконано аналіз ускладнень перебігу захворювання із застосуванням методів класифікації та кластеризації, що дало змогу виявити внутрішню стратифікацію ризиків навіть у межах бінарно заданих клінічних ознак. Результати дослідження демонструють, що використання методів машинного навчання дозволяє не лише підвищити точність аналізу клінічних даних, а й ідентифікувати приховані групи пацієнтів із різним ступенем тяжкості стану та потенційним ризиком розвитку ускладнень. Запропонований підхід може бути використаний як інструмент підтримки прийняття клінічних рішень, а також як основа для подальших досліджень у напрямі персоналізованої та превентивної медицини.

Ключові слова: *Data Mining, машинне навчання, аналіз даних, ансамблеві моделі, випадковий ліс, кластерний аналіз, оцінка важливості ознак, стратифікація ризиків, клінічні дані.*

Постановка проблеми

Впровадження та застосування машинного навчання та комплексного підходу Data Mining у системі охорони здоров'я та в клінічній медицині є наразі актуальними. Із стрімким збільшенням медичних даних та впровадженням цифрових технологій діагностики, запису та зберігання постає необхідність в комплексних підходах для підвищення ефективності лікувально-реабілітаційних заходів. Також слід зауважити, що певні інформаційні системи на основі моделей машинного навчання активно починають застосовуватись у системах підтримки прийняття лікарем рішень, в системах персоналізованої медицини.

Аналіз останніх досліджень та публікацій

Стрімкий розвиток цифрових технологій та, зокрема, цифровізація медицини та системи охорони здоров'я приводять до накопичення значних обсягів даних внаслідок демографічних, медико-соціальних, клінічних досліджень та лабораторних обстежень. Значні обсяги накопиченої інформації та можливість швидко нею оперувати створюють підґрунтя для використання не тільки класичних статистичних методів, а також методів та підходів Data Mining, машинного навчання (ML).

Основне завдання використання цих потужних комплексних підходів – це пошук інформації та нових знань шляхом аналізу великих обсягів даних, зокрема медичних [1, 2]. Побудова прогностичних моделей, систем підтримки прийняття рішень, кластерний аналіз, стратифікація пацієнтів – все це є викликами для застосування Data Mining та ML. Серед переваг використання цих засобів зазначено виявлення складних нелінійних зв'язків, підвищення точності діагностування. Але звертають увагу також на складність інтерпретації певних моделей і результатів, на необхідну якість і повноту медичних даних та обов'язковість клінічної валідації отриманих результатів.

Кластеризація – це тип машинного навчання без вчителя, який групує об'єкти за схожістю в просторі ознак. У клінічному контексті вона застосовується для виділення фенотипів пацієнтів, прогнозування відповіді на лікування, стратифікації ризику ускладнень, відбору пацієнтів у дослідження тощо [2].

Є багато алгоритмів кластеризації, але k-means та його варіанти – найпоширеніший алгоритм через простоту, ефективність і низьку обчислювальну складність. Він групує дані навколо центроїдів, але має

обмеження, пов'язані з необхідністю заздалегідь вказати число кластерів і чутливістю до вибору початкових центрів [3, 4].

Окремим підтипом таких алгоритмів є ієрархічна кластеризація, яка спрямована на побудову ієрархії кластерів у вигляді дендрограми, що дає змогу гнучко вибирати рівень поділу на групи [4].

Застосовуючи Fuzzy-кластеризацію, ми надаємо умови, за якими елементи можуть належати до кількох груп одночасно, що корисно у разі розмитих меж між фенотипами [3].

Кластерні методи охоплюють не лише основні алгоритми, а й їх модифікації та варіанти, які підвищують ефективність аналізу нестандартних форм даних, покращують стійкість до викидів чи здатність знаходити складні структури.

Для коректного застосування кластерних алгоритмів до медичних даних необхідно належним чином попередньо опрацювати дані: оброблення пропущених значень і викидів, нормалізація та масштабування змінних, перетворення категоріальних ознак, а також вибір або зменшення розмірності ознак, що є критично важливим в умовах великих наборів даних.

Стандартний алгоритм (k-means) ділить дані на k кластерів, мінімізуючи суму квадратів відстаней між точками і центроїдами.

До його переваг належать простота реалізації, швидка робота та широка практика застосування, зокрема в медицині, сегментації зображень тощо.

Водночас алгоритм має обмеження, серед яких необхідність заздалегідь визначати кількість кластерів, чутливість до початкових центрів і слабка робота з несферичними групами чи з їхнім перекриттям.

Розроблено численні варіанти реалізації k-means, спрямовані на подолання наявних обмежень в оригінального алгоритмі, зокрема методи автоматичного визначення кількості кластерів та адаптаційні метрики для підвищення стійкості алгоритму [3].

Окрім цього, сучасні дослідження вказують на зростання ролі density-based методів кластеризації для аналізу медичних даних, оскільки вони дають змогу ідентифікувати кластери довільної форми та автоматично виділяти шумові спостереження без необхідності попереднього задання кількості груп [5].

Серед таких підходів HDBSCAN демонструє особливу ефективність у разі змінної щільності даних та високого рівня шуму, що є типовим для біомедичних часових рядів і великих клінічних реєстрів. Алгоритм автоматично визначає стабільні кластери та підвищує надійність результатів фенотипування пацієнтів [6].

Поєднання кластеризації з методами зменшення розмірності, такими як PCA або UMAP, дає змогу ефективніше аналізувати високорозмірні медичні дані, виявляти приховані структури та формувати основу для персоналізованих підходів у e-healthcare [6].

Також використовуються ці методи в ендокринології, а саме в аналізі захворювання на діабет, яке характеризується високою ймовірністю ускладнень та гетерогенністю.

Somolinos-Simón з колегами досліджували гетерогеність дорослих пацієнтів, які хворіють на цукровий діабет 1-го типу. Автори використовували кластерний аналіз (метод k-means). Було виявлено 5 кластерів, які були додатково проаналізовані на предмет суттєвих відмінностей, що дає в перспективі змогу приймати зваженіші рішення та формувати стратегії лікування з врахуванням особливостей та ризиків, а також відслідковувати перебіг та динаміку ускладнення [7].

За допомогою кластеризації показано, що ризик розвитку і перебіг захворювання (після діагнозу) є суттєво гетерогенним. Kahkoska аналізували ризики макро- та мікросудинних ускладнень на тлі таких показників, як маса тіла та глікемічний контроль [8]. Lu You використовували outcome-guided clustering для когорти людей без діабету для виявлення груп з принципово різними ймовірностями прогресування поточного стану тяжкості захворювання на діабет 1-го типу [9].

Науковці застосовували методи Data Mining та машинного навчання для аналізу результатів лікування пацієнтів кардіологічного та діабетичного профілів. Зокрема, було використано класифікаційні моделі для оцінювання ефективності лікування та виявлення закономірностей у клінічних даних, що дає змогу здійснювати детальнішу стратифікацію пацієнтів за перебігом захворювання [10], розроблено програмний модуль підтримки прийняття рішень лікарем, який інтегрує алгоритми машинного навчання для вибору тактики лікування з урахуванням індивідуальних особливостей пацієнтів [11-12]. Окремим напрямом є застосування методів глибокого навчання для підтримки клінічних рішень у діагностуванні серцево-судинних, урологічних та інших захворювань, де нейронні мережі використовуються для аналізу складних нелінійних взаємозв'язків між клінічними показниками [13-14].

Мета статті

Метою статті є аналіз можливостей застосування методів Data Mining та машинного навчання для аналізу клінічних даних задля визначення ступеня тяжкості стану пацієнтів і стратифікації ризику ускладнень у двох завданнях дослідження: визначення тяжкості стану пацієнтів із захворюваннями дихальної системи на тлі ПТСР та ускладнень у дорослих пацієнтів із цукровим діабетом 1-го типу. У межах дослідження передбачається: (1) виявити асоціації між тяжкістю бронхіту, частотою епізодів ГРВІ та показниками ПТСР, розладами сну, інтегральним показником якості життя та інтенсивністю бойових дій у регіоні проживання; (2) проаналізувати ускладнення перебігу цукрового діабету 1-го типу за допомогою класифікаційних моделей та кластеризації для виявлення внутрішньої стратифікації ризиків бінарних клінічних ознак.

p-ISSN: 2225-6733; e-ISSN: 2519-271X

Матеріали та методи

Об'єктом нашого дослідження є дві групи пацієнтів: діти віком від шести місяців до 16 років із захворюваннями дихальної системи, посттравматичним стресовим розладом (ПТСР), розладами сну та дорослі хворі на діабет першого типу. Проаналізувати залежності тяжкості захворювання бронхітом від регіону проживання за інтенсивністю бойових дій, кількістю епізодів ПТСР, а також вплив показників «якості життя» (за результатами тестування пацієнтів лікарем). У пацієнтів з діабетом 1-го типу виявлено наявність ускладнень. Детальніша стратифікація пацієнтів на більш ніж бінарні категорії (наявне або відсутнє ускладнення) можуть допомогти у діагностичних заходах, у визначенні ризиків ускладнень, зокрема виявлення кластерної структури ускладнень у разі діабету 1-го типу.

Виклад основного матеріалу

Для аналізу клінічних даних було застосовано комплексний підхід Data Mining, який дає змогу використовувати широкий спектр методів від дескриптивної статистики до машинного навчання (зокрема нейронні мережі). Для кожного завдання, кожної зазначеної досліджуваної когорти (дата сетів) було застосовано різні підходи.

Виявлення впливу психоемоційних стані дітей на тяжкість та частоту гострої респіраторної вірусної інфекції (ГРВІ). Для когорти дітей з ГРВІ було застосовано статистичні методи для аналізу якості життя (інтегральна оцінка шляхом анкетування лікарем пацієнта), тяжкості ПТСР, бронхіту, частотою ГРВІ та місця проживання (за інтенсивністю бойових дій). Досліджено дані про дітей віком від 6ти місяців до 16ти років з різних міст та регіонів України, із захворюванням дихальної системи в часовому проміжку з 2022 по 2023 роки.

Діти із захворюваннями дихальних шляхів на тлі ПТСР та інтенсивності «бойових дій» у місці проживання розділені на три категорії за місцем перебування, кожна категорія відповідає інтенсивності бойових дій, де 0 – найменша інтенсивність.

Визначено три кластера (групи) дітей за ступенем інтенсивності бойових дій у місті проживання пацієнтів. Попередній аналіз виявив часткову невідповідність ступеня тяжкості бронхіту інтенсивності бойових дій у місті проживання пацієнтів. Статистично значуща різниця є між групами з найменшою та найбільшою інтенсивностями. Середня група за інтенсивністю статистично не відрізняється від двох інших (рис. 1).

Таблиця 1

Розподіл міст проживання досліджуваних пацієнтів за категорією інтенсивності «бойових дій»

Місто	Категорія (інтенсивність)
Львів	0
Черкаси	0
Вінниця	0
Кропивницький	0
Житомир	1
Бердичів	1
Київ	1
Біла Церква	1
Харків	2
Одеса	2
Запоріжжя	2
Ромни	2
Кривий Ріг	2
Полтава	2
Дніпро	2
Суми	2

p-ISSN: 2225-6733; e-ISSN: 2519-271X

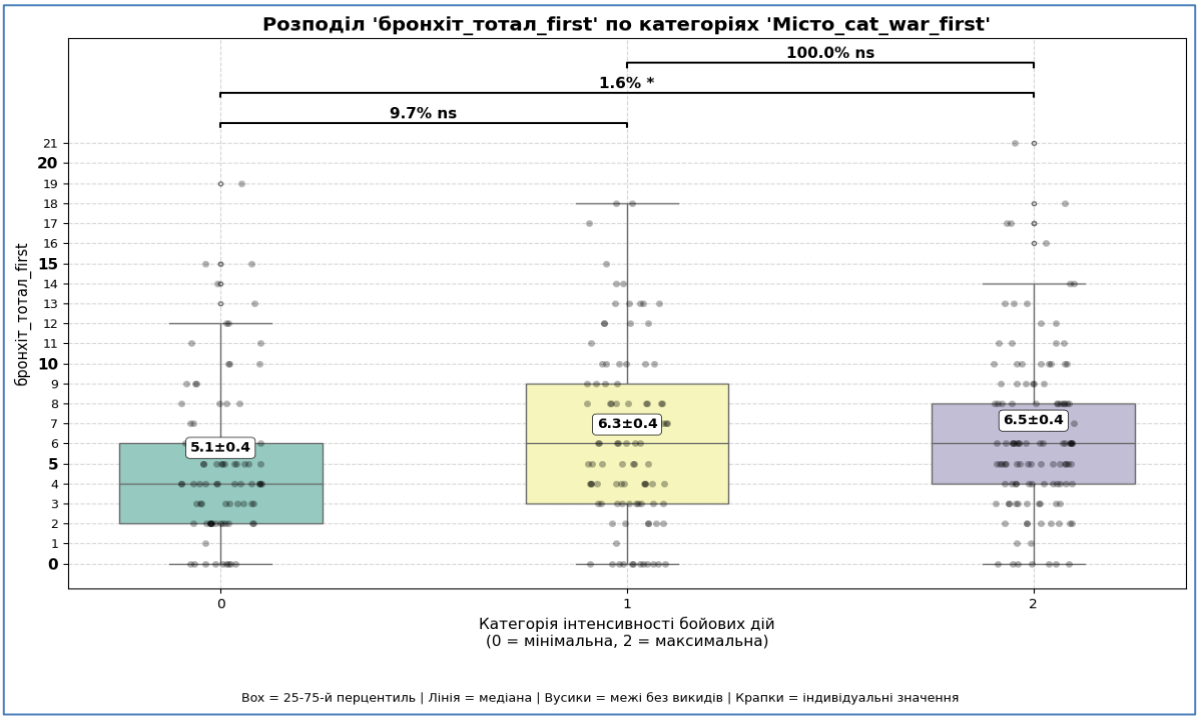

Рис. 1 – Розподіл ступеня тяжкості бронхіту у групах за інтенсивністю бойових дій у місті проживання пацієнтів

Здійснено аналіз змін інтегрального показника тяжкості бронхіту в залежності від частоти епізодів ГРВІ в 2022-2023 роки (рис. 2). Виявлено сім кластерів (груп) пацієнтів за цим показником. Визначено тенденцію до зростання тяжкості тяжкості бронхіту із збільшенням частоти ГРВІ.

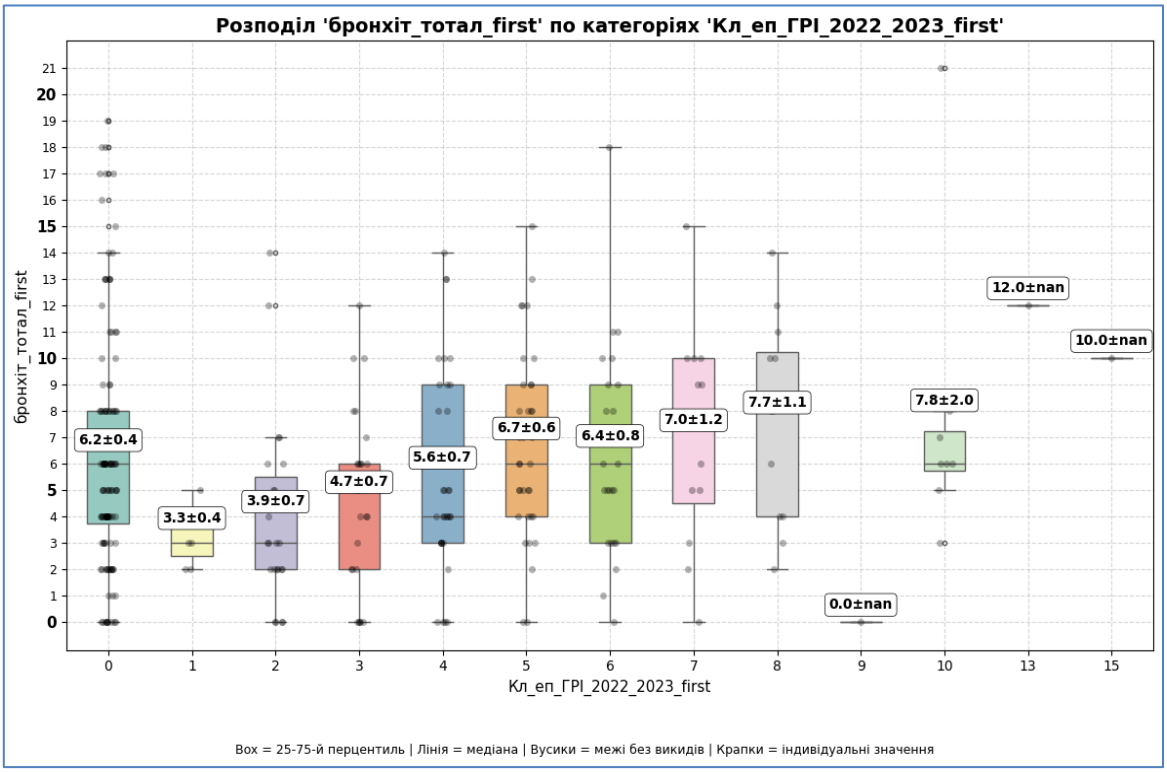

Рис. 2 – Зміни інтегрального показника тяжкості бронхіту в залежності від частоти епізодів ГРВІ в 2022-2023 роки

p-ISSN: 2225-6733; e-ISSN: 2519-271X

Подальший аналіз цільових показників, а саме показника якість життя (зважена сума балів внаслідок проведеного лікарем опитування), виявив статистично достовірну різницю між визначеними двома групами пацієнтів – зі збільшенням кількості епізодів ГРВІ погіршується якість життя (рис. 3). Слід зауважити, що з огляду на специфіку лікарського опитувальника для якості життя більший бал – гірше.

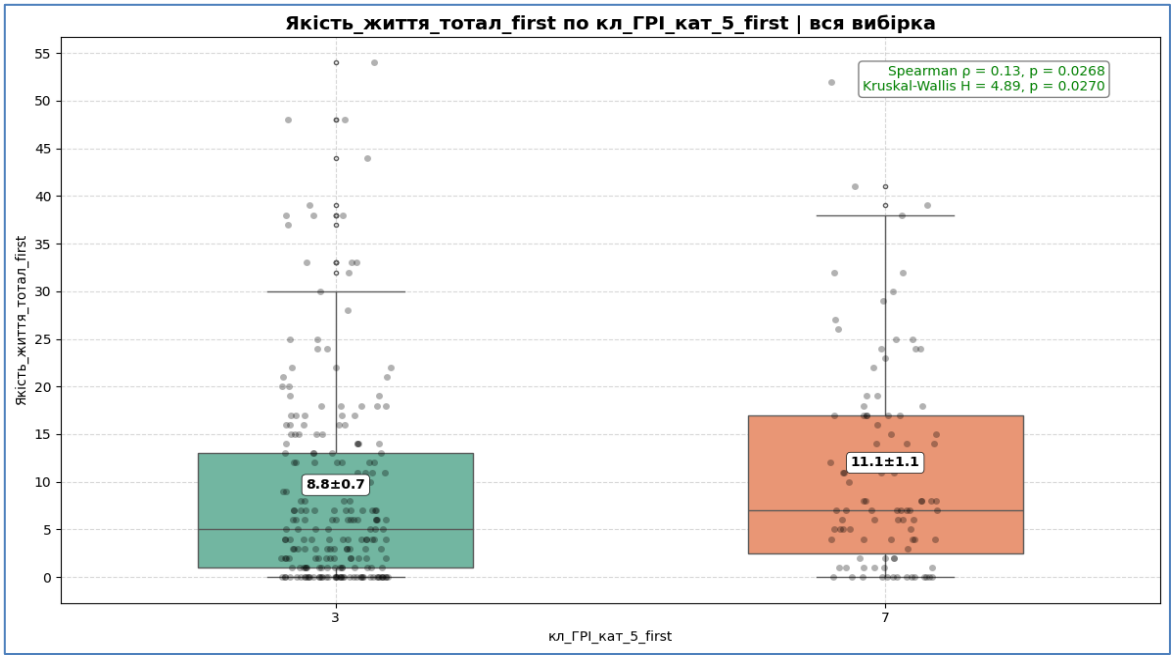

Рис. 3 – Зміни показника «якість життя» від частоти ГРВІ

Для поглибленого аналізу залежності якості життя було використано машинне навчання, а саме випадковий ліс (Random Forest, алгоритм Permutation Importance), який дає змогу оцінити реальну важливість ознак для класифікації (або регресії).

Виявлено залежність якості життя від PCL (показник ПТСР), SDSC (розлади сну), сукупності показників стресових розладів тощо. Визначено, що ПТСР, розлад сну та стресові розлади демонструють статистично достовірну кореляцію із якістю життя (рис. 4).

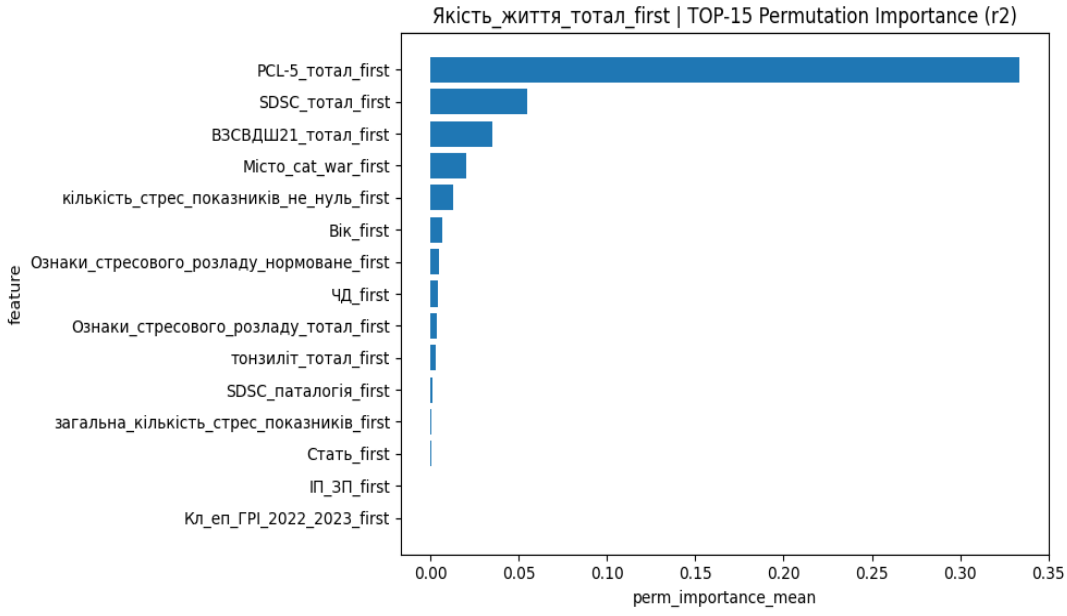

Рис. 4 – Кореляція ПТСР, розлад сну та стресові розлади демонструють статистично достовірну із якістю життя

Використання машинного навчання дає змогу за необхідності оцінити результати дослідження, виявити залежності в першому наближенні. Можливість пришвидшеного попереднього аналізу дає змогу за необхідності зосередитись на детальному аналізі особливостей визначених кластерів (груп) та пошуку асоціацій і кореляцій всередині певних кластерів та групах.

Дослідження тяжкості захворювання на діабет 1-го типу та стратифікацію ускладнення на гепатоз. Для виконання другого завдання, за допомогою методів машинного навчання (моделей класифікації, кластеризації тощо) було проаналізовано ускладнення у разі захворювання на діабет 1-го типу. Для цього за даними пацієнтів було вибрано чотири цільові показники (ускладнення перебігу діабету 1го типу), а саме: діабетична полінейропатія; нефропатія; діабетичний гепатоз; мікроангіопатія ніг. Використання методів Data mining надало можливість проаналізувати важливість та ступінь впливу клінічних показників на ускладнення перебігу діабету (рис. 5).

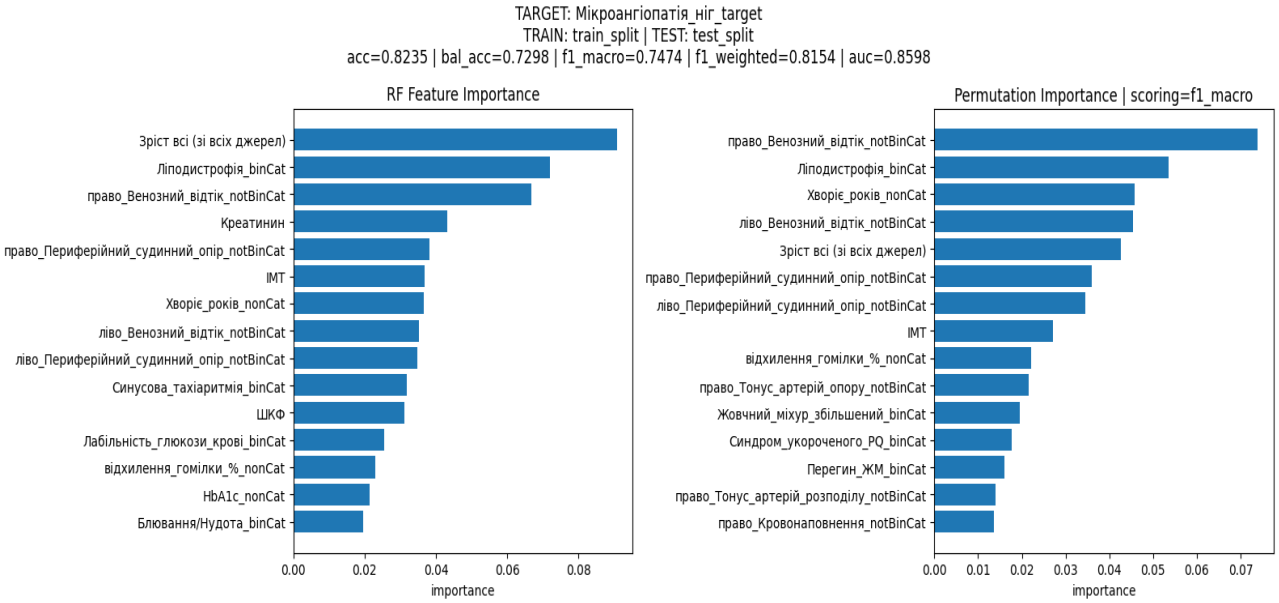

Рис. 5 – Розподіл важливості клінічних показників на наявність діагнозу мікроангіопатія ніг

Також дало змогу проаналізувати та стратифікувати рівні ризиків для певних ускладнень (гепатоз). Оскільки ускладнення перебігу захворювання наявні у бінарному вигляді (0 – без ускладнення, 1 – з ускладненням), серед моделей машинного навчання було вибрано випадковий ліс (Random Forest). Випадковий ліс є одним з найкращих методів, оскільки має нечутливість до статистичних викидів, виду даних (категоріальні, числові тощо), а також високу ефективність на відносно невеликих масивах даних. Застосовані моделі машинного навчання за методом випадкового лісу продемонстрували значення міри точності моделі F-1 на рівні 75-78% для більшості досліджуваних ускладнень та на рівні 66% – для діабетичного гепатозу. Міра F-1 це гармонійне середнє між характеристиками специфічності (*precision)* та чутливості (*recall),* що дає змогу у разі дисбалансу класів адекватно оцінити ефективність моделі.

Показник загальної точності (ACC, тобто відсоток правильних класифікацій) у разі дисбалансу класів не є інформативним, оскільки може бути високим (на рівні 90%) для класів з суттєвою різницею кількості пацієнтів в кожному з класів. Інформативнішим у таких випадках є показник F-1. Один з підходів до покращення вибірки та моделі це балансування даних. Для подолання суттєвої нерівномірності досліджуваних класів застосовано метод балансування (SMOTE) та проведено оптимізацію гіперпараметрів моделі, зокрема глибини кількості та розгалуження дерев рішень, і вибір критерію розгалуження. Для аналізу було вибрано показник діабетичного гепатозу, за яким точність класифікаційної моделі була найменшою. Оптимізацію архітектури та балансування даних привело до покращення показника точності з 66% до 74%. Незважаючи на те, що цільовий параметр має 2 значення (наявність або відсутність ускладнення), кластеризація надала шість кластерів, які мають показники якості найкращі від 2-х до 10-ти включно (рисунок 6).

Кількість розмежованих кластерів може свідчити про різну тяжкість стану пацієнтів. Навіть для однієї групи, наприклад без ускладнень, можуть бути визначено різні ступені ризику ускладнення. Аналізуючи розподіл наявності та відсутності ускладнення у кластерах, можливо дійти висновку, що певні групи мають тяжіння до ускладнення.

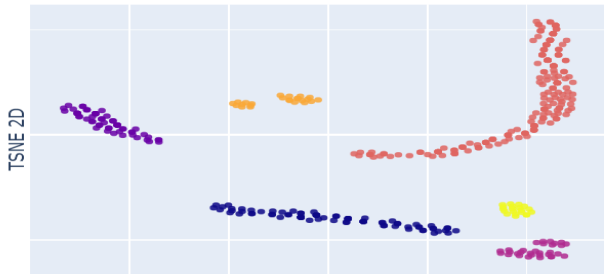

Рис. 6 – Результати кластеризації пацієнтів хворих на діабет за ознакою гепатоз

Алгоритмічну реалізацію методів Data Mining та машинного навчання здійснено в програмному середовищі Python з використанням бібліотек NumPy, Pandas, Scikit-learn, а також методів зниження розмірності та візуалізації багатовимірних даних (t-SNE, PCA) для аналізу результатів кластеризації.

Результати та їх обговорення

За допомогою класичних статистичних методів та комплексних підходів Data Mining отримано результати, які можуть бути корисними в клінічній медицині та в цілому у системі охорони здоров'я. Аналіз тяжкості стану дітей із захворюванням дихальної системи продемонстрував залежність тяжкості бронхіту від частоти епізодів ГРВІ, інтенсивності бойових дій у місті проживання тощо. Виявлено найважливіші предиктори для класифікації характеристики «якість життя». Проведений аналіз за допомогою методів машинного навчання тенденцій ускладнення на гепатоз для хворих на діабет 1го типу показав, що розподіл на більшу кількість кластерів, аніж два, надає додаткове уточнення структури ускладнень. Це уможливлює створення ефективніших прогностичних моделей розвитку певного ускладнення, що забезпечуватиме раннє ідентифікування пацієнтів, які знаходяться в групі ризику.

Висновки

Застосування методів Data Mining та машинного навчання до аналізу клінічних даних дає змогу ефективно виявляти приховані залежності та гетерогенні підгрупи пацієнтів, що не завжди є очевидним за використання виключно класичних статистичних підходів.

Встановлено залежність тяжкості бронхіту у дітей із захворюваннями дихальної системи від частоти епізодів ГРВІ, рівня показників посттравматичного стресу, розладів сну та інтенсивності бойових дій у регіоні проживання, а також визначено ключові предиктори, що впливають на якість життя пацієнтів.

Аналіз даних пацієнтів із цукровим діабетом 1-го типу показав, що розподілення досліджуваного масиву даних на більшу кількість кластерів надає можливість проведення докладнішого аналізу ризиків виникнення та подальшого розвитку ускладнень основного захворювання.

Отримані результати підтверджують доцільність використання ансамблевих моделей машинного навчання, зокрема випадкового лісу, для оброблення медичних даних, що характеризуються невеликими вибірками, зашумленістю та незбалансованістю класів.

Перелік використаних джерел

[1] Досвід та перспективи створення медичних інформаційних систем та інформаційних технологій підтримки надання медичної допомоги / Коваленко О. С., Козак Л. М., Наджифіан Туманджані М., Романюк О. О. *Medical and Biological Cybernetics* 2022. Vol. 1 (207). Pp. 59-73. DOI: https://doi.org/10.15407/kvt207.01.059.

[2] Phenotype clustering in health care: A narrative review for clinicians / T. J. Loftus et al. *Frontiers in Artificial Intelligence*. 2022. Vol. 5. Pp. 1-11. DOI: https://doi.org/10.3389/frai.2022.842306.

[3] K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data / A. M. Ikotun et al. *Information Sciences*. 2023. Vol. 622. Pp. 178-210. DOI: https://doi.org/10.1016/j.ins.2022.11.139.

[4] A Brief Review on Medical Data and Clustering Algorithms for Smart Healthcare System: Challenges and Opportunities / Jandachot S., Simmachan T., Shakya S., Boonkroong P. *The 7th International Conference on Applied Statistics 2024 (ICAS2024)*, Chiang Mai, Thailand, 24-25 October 2024. Pp. 24-40.

[5] Quispe Hilasaca S. M. Review on the Optimization of Unsupervised Clustering Models in Healthcare. Preprint 2109.07771. 2024. DOI: https://doi.org/10.20944/preprints202412.1423.v1.

[6] Scalable Clustering of Complex ECG Health Data: Big Data Clustering Analysis with UMAP and HDBSCAN / V. Kaverinskiy et al. *Computation*. 2025. Vol. 13(6). Article 144. DOI: https://doi.org/10.3390/computation13060144.

[7] Cluster analysis of adult individuals with type 1 diabetes: Treatment pathways and complications over a five-year follow-up period / F. J. Somolinos-Simón et al. *Diabetes Research and Clinical Practice*. 2024. Vol. 215. Article 111803. DOI: https://doi.org/10.1016/j.diabres.2024.111803.

[8] Longitudinal Phenotypes of Type 1 Diabetes in Youth Based on Weight and Glycemia and Their Association With Complications / A. R. Kahkoska et al. *The Journal of Clinical Endocrinology & Metabolism*. 2019. Vol. 104, iss. 12. Pp. 6003-6016. DOI: https://doi.org/10.1210/jc.2019-00734.

[9] Identification of type 1 diabetes risk phenotypes using an outcome-guided clustering analysis / Y. Lu et al. *Diabetologia Clinical, Translational and Experimental Diabetes and Metabolism*. 2024. Vol. 67. Pp. 2507-2517. DOI: https://doi.org/10.1007/s00125-024-06246-w.

[10] Застосування класифікаційних моделей за

методами Data Mining та інформаційної технології для аналізу результатів лікування пацієнтів кардіологічного та діабетичного профілів / О.С. Коваленко та ін. *Medical and Biological Cybernetics*. 2023. Vol. 1(211). Pp. 77-89. DOI: https://doi.org/10.15407/kvt211.01.077.

[11] Програмний модуль підтримки прийняття рішень лікарем при виборі тактики лікування / О. С. Коваленко та ін. *Біомедична інженерія і технологія*. 2025. Вип. 17(1). С. 52-59. DOI: https://doi.org/10.20535/.2025.17.328254.

[12] Застосування методів глибокого навчання у підтримці прийняття клінічних рішень при діагностиці серцево-судинних захворювань / О. С. Коваленко та ін. *Автоматизація та біомедичні і комп'ютерні технології* : збірник праць Всеукраїнської науково-технічної інтернет-конференції, м. Дніпро, 26 березня 2025. С. 182-186.

[13] Clinical Applications of Machine Learning for Urolithiasis and Benign Prostatic Hyperplasia: A Systematic Review / D. Bouhadana et al. *Journal of Endourology*. 2023. Vol. 37(4). Pp. 182-185. DOI: https://doi.org/10.1089/end.2022.0311.

[14] Universal representations in cardiovascular ECG assessment: A self-supervised learning approach / Z.-Y. Liu et al. *International Journal of Medical Informatics*. 2025. Vol. 195. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105742.

DATA MINING METHODS AND MACHINE LEARNING AS TOOLS FOR ANALYZING PATIENT CONDITIONS AND DISEASE COMPLICATIONS BASED ON CLINICAL DATA

Biliavenko L.V. *postgraduate student, junior Researcher, Institute of information technologies and systems of the National academy of sciences of Ukraine, Kyiv, ORCID: https://orcid.org/0009-0003-0602-1072, e-mail: leonidtill@gmail.com;*

Kovalenko O.S. *D.Sc. (Medical Sciences), professor, ORCID: https://orcid.org/0000-0001-6635-0124*

The article investigates the possibilities of applying Data Mining methods and machine learning techniques for the analysis of clinical data in the context of complex medical conditions characterized by high heterogeneity of disease progression and an increased risk of complications. The relevance of the study is determined by the rapid digitalization of the healthcare system, the growth of medical data volumes, and the need to implement personalized approaches to diagnosis, treatment, and rehabilitation of patients. Particular attention is paid to data analysis in pediatric practice and endocrinology, where traditional statistical methods are often insufficient for identifying complex nonlinear relationships. The objects of the study include two clinical cohorts: children aged from six months to sixteen years with respiratory system diseases associated with post-traumatic stress disorder, and adult patients with type 1 diabetes mellitus presenting various complications of disease progression. Within the scope of the research, a comprehensive approach was applied that combines descriptive statistical methods with machine learning algorithms, including ensemble models, feature importance evaluation methods, and clustering analysis. For the pediatric cohort, relationships between bronchitis severity, frequency of acute respiratory viral infections, post-traumatic stress indicators, sleep disorders, an integrated quality-of-life index, and the intensity of hostilities in the region of residence were analyzed. For the cohort of patients with type 1 diabetes mellitus, complications of disease progression were analyzed using classification and clustering methods, which made it possible to identify internal risk stratification even within clinically binary outcome variables. The results demonstrate that the application of machine learning methods not only improves the accuracy of clinical data analysis but also enables the identification of latent patient groups with different degrees of disease severity and potential risk of complication development. The proposed approach can be used as a clinical decision support tool and as a basis for further research in the field of personalized and preventive medicine.

Keywords: *Data Mining, machine learning, data analysis, ensemble models, random forest, clustering analysis, feature importance evaluation, risk stratification, clinical data.*

References

[1] O.S. Kovalenko, L.M. Kozak, M.N. Tumajani, and O.O. Romanyuk, "Dosvid ta perspektyvy stvorennia medychnykh informatsiinykh system ta informatsiinykh tekhnolohii pidtrymky nadannia medychnoi dopomohy" ["Experience and prospects of creating medical information systems and informtion technologies to support medical care"], *Medical and Biological Cybernetics*, vol. 1 (207), pp. 59-73, 2022. **doi: 10.15407/kvt207.01.059**. (Ukr.)

[2] T.J. Loftus et al., "Phenotype clustering in health care: A narrative review for clinicians," *Frontiers in Artificial Intelligence*, vol. 5, pp. 1-11, 2022. **doi: 10.3389/frai.2022.842306**.

[3] A.M. Ikotun, A.E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data," *Information Sciences*, vol. 622, pp. 178-210, 2023. **doi: 10.1016/j.ins.2022.11.139**.

[4] S. Jandachot, T. Simmachan, S. Shakya, and P. Boonkroong, "A Brief Review on Medical Data and Clustering Algorithms for Smart Healthcare System: Challenges and Opportunities," in Proc. of the *7th Int. Conf. on Applied Statistics 2024 (ICAS2024)*, Chiang Mai, Thailand, Oct. 24-25, 2024, pp. 24-40.

[5] S.M. Quispe Hilasaca, "Review on the Optimization of Unsupervised Clustering Models in Healthcare," 2024, *preprint 2109.07771*. **doi: 10.20944/preprints202412.1423.v1**.

[6] V. Kaverinskiy, I. Chaikovsky, A. Mnevets, T. Ryzhenko, M. Bocharov, and K. Malakhov, "Scalable Clustering of Complex ECG Health Data: Big Data Clustering Analysis with UMAP and HDBSCAN," *Computation*, vol. 13(6), article 144, 2025. **doi: 10.3390/computation13060144**.

[7] F.J. Somolinos-Simón, G. García-Sáez, J. Tapia-Galisteo, R. Corcoy, and M.E. Hernando, "Cluster analysis of adult individuals with type 1 diabetes: Treatment pathways and complications over a five-year follow-up period," *Diabetes Research and Clinical Practice*, vol. 215, article 111803, 2024. **doi: 10.1016/j.diabres.2024.111803**.

[8] A.R. Kahkoska et al., "Longitudinal Phenotypes of Type 1 Diabetes in Youth Based on Weight and Glycemia and Their Association With Complications," *The Journal of Clinical Endocrinology & Metabolism*, vol. 104, iss. 12, pp. 6003-6016, 2019. **doi: 10.1210/jc.2019-00734**.

[9] Y. Lu et al., "Identification of type 1 diabetes risk phenotypes using an outcome-guided clustering analysis," *Diabetologia Clinical, Translational and Experimental Diabetes and Metabolism*, vol. 67, pp. 2507-2517, 2024. **doi: 10.1007/s00125-024-06246-w**.

[10] O.S. Kovalenko, L.M. Kozak, O.A. Kryvova, V.V. Bychkov, and L.V. Nenasheva, "Zastosuvannia klasyfikatsiinykh modelei za metodamy Data Mining ta informatsiinoi tekhnolohii dlia analizu rezultativ likuvannia patsiientiv kardiolohichnoho ta diabetychnoho profiliv" ["Application of classification models by data mining and information technology for analyze the results of treatment of cardiac and diabetic patients"], *Medical and Biological Cybernetics*, vol. 1(211), pp. 77-89, 2023. **doi: 10.15407/kvt211.01.077**. (Ukr.)

[11] O. Kovalenko, L. Kozak, O. Averyanova, L. Bilyavenko, and O. Kutsiak, "Prohramnyi modul pidtrymky pryiniattia rishen likarem pry vybori taktyky likuvannia" ["Software module for doctor decision-making support when choosing treatment tactics"], *Biomedychna inzheneriia i tekhnolohiia – Biomedical Engineering and Technology*, vol. 17(1), pp. 52-59, 2025. **doi: 10.20535/.2025.17.328254**. (Ukr.)

[12] O.S. Kovalenko et al., "Zastosuvannia metodiv hlybokoho navchannia u pidtrymtsi pryiniattia klinichnykh rishen pry diahnostytsi sertsevo-sudynnykh zakhvoriuvan" ["Application of deep learning methods in supporting clinical decision-making in the diagnosis of cardiovascular diseases"], in *Proc. of the All-Ukrainian Sci. and Techn. Internet Conf. «Automation and biomedical and computer technologies»*, Dnipro, Ukraine, March 26, 2025, pp. 182-186. (Ukr.)

[13] D. Bouhadana et al., "Clinical Applications of Machine Learning for Urolithiasis and Benign Prostatic Hyperplasia: A Systematic Review," *Journal of Endourology*, vol. 37(4), pp. 182-185, 2023. **doi: 10.1089/end.2022.0311**.

[14] Z.-Y. Liu et al., "Universal representations in cardiovascular ECG assessment: A self-supervised learning approach," *International Journal of Medical Informatics*, vol. 195, 2025. **doi: 10.1016/j.ijmedinf.2024.105742**.

Стаття надійшла 18.09.2025
Стаття прийнята 01.11.2025
Стаття опублікована 29.12.2025

Цитуйте цю статтю як: Білявенко Л. В., Коваленко О. С. Методи Data Mining та машинне навчання як засоби аналізу стану пацієнтів та ускладнень захворювань за клінічними даними. *Вісник Приазовського державного технічного університету. Серія: Технічні науки*. 2025. Вип. 52. С. 55-63. DOI: https://doi.org/10.31498/2225-6733.52.2025.350990.