# Кібернетика та комп'ютерні технології

УДК 004.8:616.1
DOI:10.34229/2707-451X.25.4.4

М.О. ПРАЗДНІКОВА

## МОДЕЛІ ПРОГНОЗУВАННЯ РИЗИКУ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ

*Запропоновано підхід до покращення прогнозування ризику серцево-судинних захворювань. Створена нова модель на основі Multinomial Naive Bayes для визначення ураження судин головного мозку на основі аналізу лікарських висновків.*

***Ключові слова:*** *інфаркт міокарда, інсульт, машинне навчання, прогнозування ризику, Multinomial Naive Bayes, медичні тексти, аналіз даних.*



**Вступ.** Неінфекційні захворювання дуже поширені серед населення. Особливе місце серед неінфекційних патологій посідають серцево-судинні захворювання, які є однією з провідних причин смертності у світі. Найнебезпечнішими проявами таких захворювань є інфаркт міокарда та інсульт. Інфаркт міокарда – це ураження серцевого м'яза, викликане гострим порушенням його кровопостачання, що призводить до некрозу ураженої частини. Інсульт – характеризується гострим порушенням мозкового кровообігу, зумовлюючи неврологічні симптоми або навіть смерть [1]. Для ефективної профілактики і лікування неінфекційних хвороб надзвичайно важливе – це раннє виявлення пацієнтів із підвищеним ризиком розвитку цієї патології. Так як неінфекційні захворювання характеризуються стабільністю і повільною зміною показників захворюваності, то це створює сприятливі умови для аналізу великих обсягів медичних даних і знаходження рівня ризику їх появи у пацієнтів. Сучасні методи прогнозування та оцінки ризику базуються на аналізі медичної інформації, зокрема, на сукупності текстів лікарських записів, що містять цінні клінічні ознаки та історію хвороби пацієнта. Використання методів машинного навчання та обробки текстових записів лікарів дозволяє автоматизовано і ефективно аналізувати великі масиви текстових даних для ідентифікації ключових факторів ризику, покращуючи тим самим точність прогнозування загострень і персоналізацію підходів до лікування.

У попередній роботі [2] було розроблено та випробувано модель для прогнозування інфаркту міокарда на основі аналізу текстових лікарських висновків. Наразі ведеться робота з удосконалення цієї моделі, а також розглядається можливість її адаптації для визначення ризику інших неінфекційних захворювань.

### 1. Огляд відомих підходів

У сучасній медицині для прогнозування та оцінки ризику інфаркту міокарда активно впроваджуються різноманітні алгоритми машинного навчання. Серед них найбільш поширені – гауссовий наївний Байєс, випадковий ліс, метод опорних векторів (SVM), логістична регресія та XGBoost. Наприклад, у Великобританії розроблена система NIHR CALIBER, яка базується

на загальнодержавній електронній системі медичних записів із п'яти лікарень та містить дані понад 80 тисяч пацієнтів і 586 клінічних параметрів [3]. Ця система успішно використовується для прогнозування інфаркту міокарда й дозволяє аналізувати вплив численних факторів, включаючи супутні серцево-судинні захворювання, ожиріння, патології нирок і респіраторної системи.

Детальний систематичний огляд методів визначення ризику захворювання на інфекційні хвороби представлено у статті [4]. Проаналізовано понад 100 досліджень, у яких для прогнозування серцево-судинних подій застосовувалися різні моделі, серед яких виділяються наївний байєсівський класифікатор, логістична регресія, випадковий ліс, XGBoost і штучні нейронні мережі. Особлива увага приділена роботі з неструктурованими даними, зокрема текстами лікарських висновків. Підкреслюється, що попередня обробка тексту, виділення релевантних клінічних ознак та застосування методів обробки природної мови (NLP) є критично важливими для підвищення точності прогнозування. Висновки огляду підтверджують ефективність використання машинного навчання для аналізу медичних текстів з метою раннього виявлення пацієнтів із підвищеним ризиком інфаркту міокарда, що цілком відповідає обраному в нашій роботі підходу.

У роботі [5] проаналізовано дослідження 2021–2024 рр., де порівнювали класичні методи оцінки ризику серцево-судинних подій (Framingham, QRISK) із сучасними підходами машинного навчання, зокрема, SVM, Random Forest, а також глибокими нейронними мережами (CNN, RNN). Моделі CNN і RNN використовувалися переважно для обробки часових рядів, наприклад, сигналів фотоплетизмографії (PPG), тоді як SVM і Random Forest – для аналізу електронних медичних записів (EHR) із фокусом на прогнозування серцевих подій, таких як інфаркт чи інсульт. Автори зазначають, що методи ML та глибокого навчання здатні враховувати ширший спектр предикторів і забезпечувати точніші та більш персоналізовані прогнози, часто перевершуючи традиційні шкали ризику.

Розроблена модель прогнозування ішемічної хвороби серця, яка використовує сучасні методи глибокого навчання [6]. У моделі враховано ключові проблеми медичних даних, зокрема, незбалансованість класів і вибір релевантних ознак. Дані включають понад 37,000 лікарських записів, у яких ідентифіковано 51 атрибут; до найбільш інформативних належать вік, рівень креатиніну, глюкоза, наявність діабету тощо. Для оптимізації ваг мережі використано гібридну архітектуру на основі Particle Swarm Optimization (PSO) та штучної нейронної мережі (ANN), що забезпечило точність класифікації на рівні 91%. Додатково застосовано метод SMOTE для балансування класів, що покращило здатність моделі виявляти позитивні випадки CHD. Порівняльний аналіз з традиційними алгоритмами виявив перевагу запропонованої моделі за точністю та стабільністю.

Сучасні дослідження демонструють можливість ефективного використання як традиційних, так і глибоких алгоритмів машинного навчання для аналізу ризиків не лише неінфекційної, але й інфекційної патології. Зокрема, у роботі [7] моделі машинного навчання були застосовані для прогнозування ризиків поширення заразних хвороб на різних етапах – від пандемій і епідемій до фаз ендемії та елімінації, що свідчить про універсальність цих підходів для завдань медичної аналітики.

**2. Попередні результати**

У ході роботи було використано деперсоналізований фрагмент бази даних «Ескулап», який містить інформацію про 22 тисячі пацієнтів та їх 773 тисячі візитів за період 2013–2023 рр. Серед цих пацієнтів 893 особи перенесли інфаркт міокарда. Аналіз здійснювався на основі сукупності текстів лікарських висновків, які містять відомості про анамнез, діагнози та призначене лікування.

Зібрані дані включають прийоми 22 тисяч пацієнтів, зокрема 63 тисячі візитів до кардіолога до розвитку інфаркту у відповідної когорти та майже 130 тисяч візитів після події. Вихідні дані було підготовлено та анонімізовано ДНУ ЦІТОЗ ДУС і передано ІК НАН України на основі договору про співробітництво.

Для вирішення поставленої задачі була розроблена модель прогнозування інфаркту міокарда на основі класичного байєсівського підходу, яка працює з векторами слів-ознак, створених за допомогою алгоритму Левенштейна для ототожнення словоформ. Для навчання моделі здійснювався спеціальний відбір значущих медичних термінів.

У 2024 році було реалізовано та випробувано модель прогнозування інфаркту міокарда на основі аналізу текстів лікарських висновків і результати наведені в табл. 1. Наведена матриця помилок і бачимо, що правильно прогнозований інфаркт – 80 %. Тестування моделі на збалансованій вибірці (893 пацієнти з ІМ і 900 без ІМ) показало високу точність прогнозу – 80,0 % для визначення високого ризику та 81,6 % для визначення низького ризику. Площа під ROC-кривою становила 0,898, що свідчить про значну надійність моделі.

ТАБЛИЦЯ 1. Матриця помилок для визначення інфаркту

| **Точність = 80%** | Прогноз | |
|---|---|---|
| Факт | **Інфаркт** | **Ні** |
| **Інфаркт** | TP = 80.0 % | FN = 20.0 % |
| **Ні** | FP = 18.4 % | TN = 81.6 % |

**3. Розширення наявного методу**

Також розширили метод, який застосовували. На його базі було побудовано ще одну модель для прогнозування ураження судин головного мозку. В фрагментах деперсоналізованної бази даних «Ескулап» було знайдено 17006 пацієнтів із кодами I60–I69 (цереброваскулярні хвороби) та 4646 пацієнтів які мали код I але не мали ураження судин головного мозку. Саме для обробки і побудови моделі було сформовано тренувальний і тестовий набори з обох груп: до тренувального датасету потрапили 3600 пацієнтів із кодами I60–I69 та 1200 – до тестового набору. Без коду виділили 3600 для тренувального датасету і 1145 пацієнтів для тестового. Для моделювання було виділено 77821 унікальне слово-ознаку, з яких інформативними (ті, що мали ненульові ваги в моделі) виявилися 16787.

Для побудови моделі була використана бібліотека scikit-learn – це безкоштовна бібліотека для мови Python, що надає функціонал для створення і тренування різних алгоритмів машинного навчання. Як у попередній статті [2] автори обрали метод Naive Bayes, а саме MultinomialNB, який був у пакеті. MultinomialNB – це один із класичних алгоритмів наївного Байєса. Multinomial Naive Bayes це спеціалізована версія наївного байєсівського класифікатора, яка оптимально підходить для роботи з дискретними ознаками, зокрема, з текстовими даними. Різниця між методом, який я вже застосовувала і MultinomialNB – це те, що новий метод використовує саме кількість появи слів, а не частотність. Класичне припущення моделі полягає у незалежності ознак (слів) у межах одного документа для кожного класу [8]. Автори продовжили працювати з методом Байєса через те, що в нас велика кількість незалежних атрибутів і цей метод добре з ними працює.

Ймовірність того, що текст належить до певного класу $C_k$, обчислюється за формулою:

$$P(C_k|x) = P(C_k)\prod_{i=1}^{n} P(w_i|C_k)^{x_i}, \quad (1)$$

де

- $P(C_k)$ – апріорна ймовірність класу $C_k$;

- $x_i$ – кількість появ слова $w_i$ у документі;
- $P(w_i|C_k)$ – ймовірність появи слова $w_i$ у класі $C_k$;
- $n$ – розмір словника (кількість унікальних ознак).

Ймовірність появи ознаки $w_i$ для класу $C_k$ оцінюється за допомогою згладженого максимуму правдоподібності:

$$P(w_i|C_k) = \frac{N_{k,i} + \alpha}{N_k + \alpha n}, \tag{2}$$

де

- $N_{k,i}$ – кількість появи слів$i$ у всіх документах класу $C_k$;
- $N_k$ – загальна кількість усіх слів у класі $C_k$;
- $\alpha$ – параметр згладжування;
- $n$ – кількість ознак.

Згладжування α дозволяє враховувати слова, які не з'являлися у навчальних документах певного класу, та уникати нульових ймовірностей у розрахунках.

Також цей підхід дозволив виявити ключові лексичні маркери, пов'язані з високим ризиком захворювання (рис. 1). Важливість кожної ознаки оцінювалася шляхом обчислення різниці логарифмічних ймовірностей зустрічі цієї ознаки у класах "є ураження" та "без ураження" на основі параметрів навченого наївного байєсівського класифікатора. Ознаки з найбільшою позитивною різницею вважаються найхарактернішими для класу "є ураження", тоді як від'ємні значення вказують на більшу типовість для іншого класу.

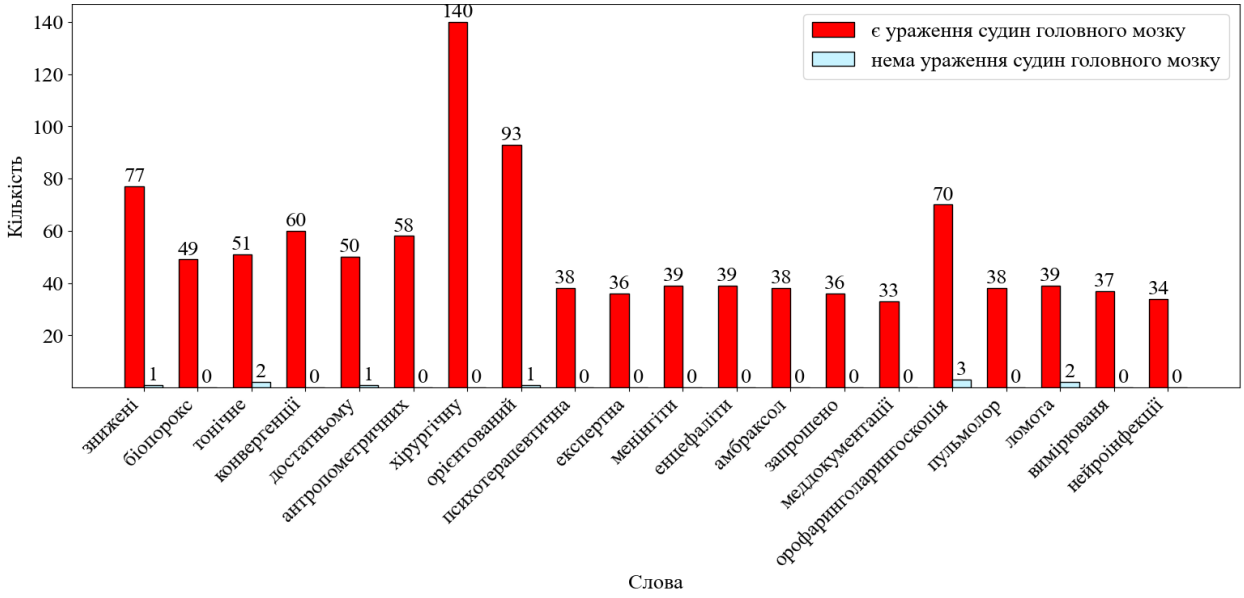


РИС. 1. Кількість 20 найбільш впливових слів визначено sklearn

**4. Удосконалення моделі**

Покращення існуючої моделі було спрямовано на усунення дублювання ознак, що виникає внаслідок використання різних граматичних форм одного й того самого слова. Це дозволяє підвищити точність і однозначність аналізу текстів. Для вирішення цієї проблеми розглядався метод Левенштейна [9], який дозволяє зіставляти слова за схожістю написання, однак цей підхід не дає змогу прибрати всі дублювання. Тому було вирішено використати інші методи, застосувати бібліотеку

spaCy [10], яка виконує лематизацію, тобто приводить слова до їхньої нормальної (лематизованої) форми. Це забезпечує уніфікацію всіх слів до однієї граматичної форми, що суттєво покращує якість аналізу текстових даних. Результати виділення нових слів покращили читабельність і виділили нові групи слів, які характеризують ураження головного мозку (рис. 2).

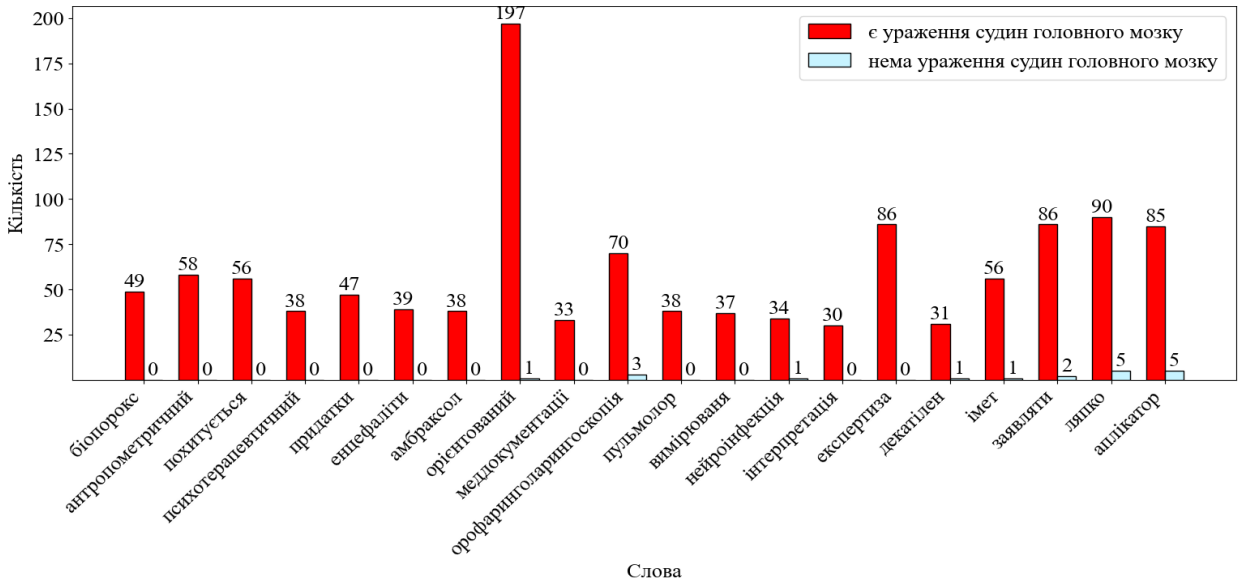


РИС. 2. Кількість 20 найбільш впливових слів визначено після лемматизації

**5. Експериментальні результати**

Використання MultinomialNB дає хороші результати для прогнозування ураження судин головного мозку. Оцінка якості моделі Multinomial Naive Bayes для задачі прогнозування продемонструвала високі показники ефективності. Чутливість для класу з кодом I60–I69 становила 76,2 %, тобто модель правильно ідентифікує понад три чверті пацієнтів із відповідним діагнозом, тоді як частка хибнонегативних результатів дорівнювала 23,8 %. Для контрольної групи без коду специфічність досягла 96,2 %, а частка хибнопозитивних результатів склала лише 3,8 %. Матриця помилок та відповідні метричні показники моделі (табл. 2).

ТАБЛИЦЯ 2. Матриця помилок і основні метрики класифікації

| **Точність = 86%** | Прогноз | |
|---|---|---|
| Факт | **P (I60-I69)** | **N (ні)** |
| **P (I60-I69)** | TP = 76.2 % | FN = 23.8 % |
| **N (ні)** | FP = 3.8 % | TN = 96.2 % |

Крива операційних характеристик (ROC-крива) для побудованої моделі (рис. 3). Площа під ROC-кривою (AUC) становила 0,92, що підтверджує високу здатність моделі розмежовувати пацієнтів із наявністю та відсутністю ураження головного мозку. F1-метрика моделі становить 86 %, що свідчить про збалансовану роботу алгоритму як за точністю (precision), так і за повнотою (recall).

Такий рівень F1-score означає, що модель досить впевнено відокремлює пацієнтів із ризиком інфаркту від інших, мінімізуючи як хибнопозитивні, так і хибнонегативні результати.

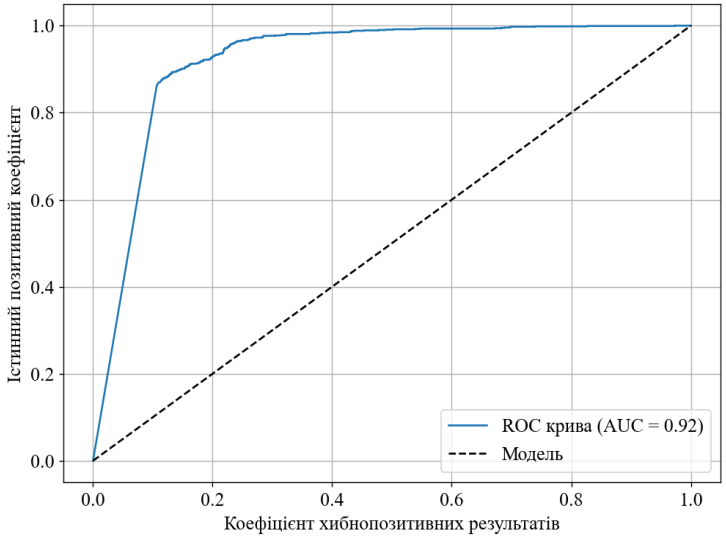


РИС. 3. ROC-крива для моделі (AUC = 0,92)

T-SNE [11] проєкція (рис. 4) візуально показує розподіл медичних записів з різними ризиками. На графіку видно, що пацієнти з високим ризиком утворюють щільне скупчення в одній частині простору, а щільність низького рівня ризику більш зосереджена в іншій зоні і менше перетинається з високим ризиком. Такий розподіл свідчить про те, що текстові характеристики записів (наприклад, ключові слова чи фрази) відрізняються між групами, і ці ознаки можуть бути успішно використані моделлю для класифікації ризику. Щільність відображено за допомогою кольорової шкали: чим темніший відтінок, тим вища концентрація записів певної категорії у цій області простору.

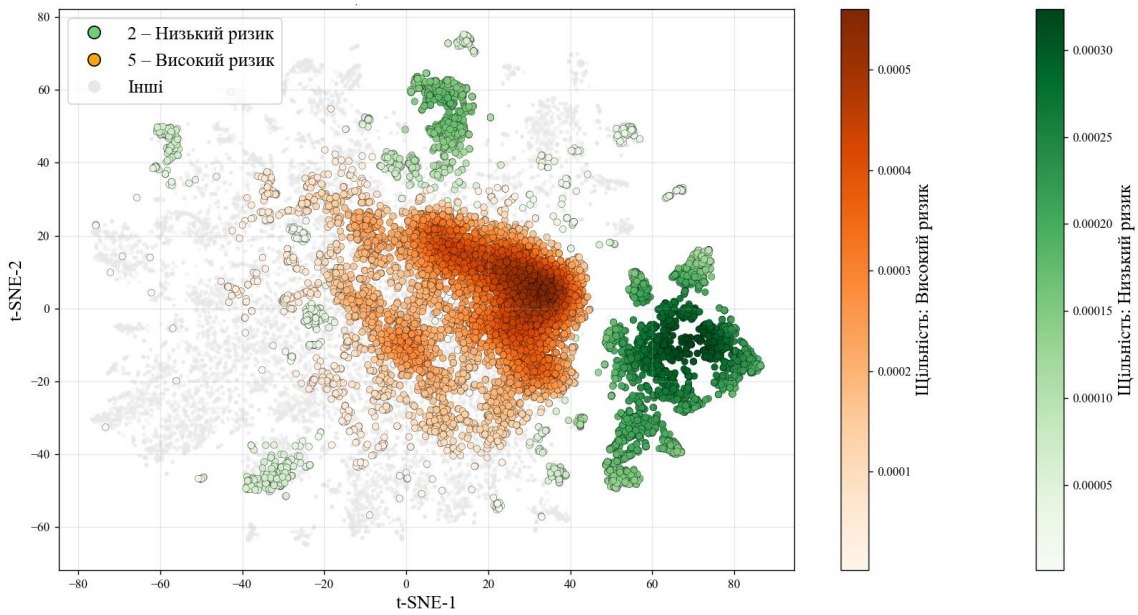


РИС. 4. T-SNE розподіл по записам пацієнтів за рівнем ризику

На основі отриманих результатів можна обчислити індивідуальний ризик ураження головного мозку для кожного пацієнта (рис. 5) та визначити порогові значення для п’яти груп ризику.

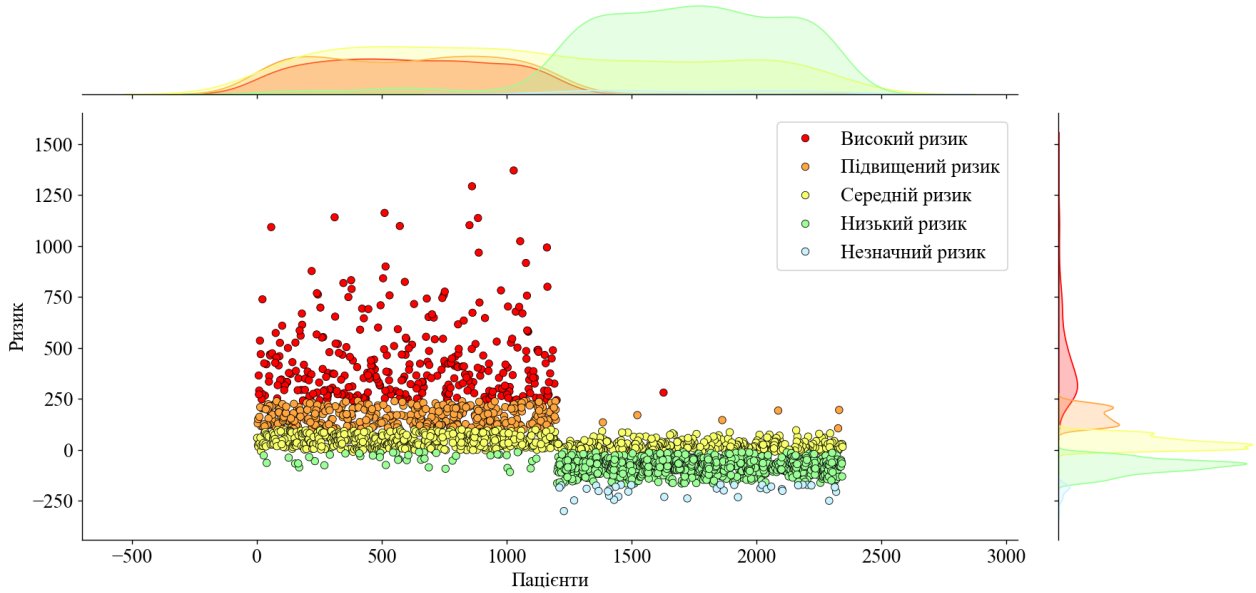


РИС. 5. Рівень ризику ураження судин головного мозку по пацієнтам

Таким чином, отримані результати свідчать про практичну придатність моделі для автоматизованої оцінки ризику ураження головного мозку на основі текстових медичних записів (табл. 3).

ТАБЛИЦЯ 3. Шкала ризику

| Ризик | Пороги |
|---|---|
| Високий | > 240 |
| Підвищений | 100 – 240 |
| Середній | –10 – 100 |
| Низький | –170 – –10 |
| Незначний | < –170 |

**5. Практичне застосування**

Розроблена система випробувана на реальних даних та інтегрована у ДНУ «ЦІТОЗ» ДУС, де створено захищене середовище, в якому на окремому комп’ютері, підключеному до реальної бази даних «Ескулап» та ізольованому від мережі Інтернет, було встановлено першу версію програми для прогнозування ризику інфаркту міокарда, показано на рис. 6. Розроблена модель успішно інтегрована у цей захищений контур для тестового використання на реальних медичних даних.

Програмне забезпечення підтримує два режими роботи: прогнозування ризику для групи осіб за ідентифікаторами пацієнтів, перелічених у вхідному файлі (input.csv), із виведенням результатів у відповідну папку (output), а також прогнозування ризику для окремого пацієнта, дані якого можуть

бути введені вручну через форму із виведенням результатів у вікно програми та збереженням у файл. У перспективі передбачається збереження результатів безпосередньо у базі даних як окремого типу медичного дослідження.

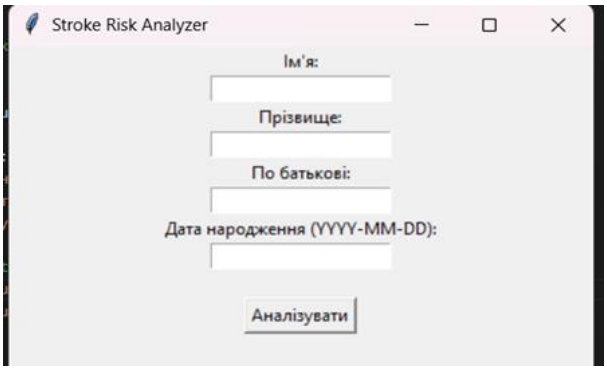


РИС. 6. Графічний інтерфейс користувача

**Висновки.** У роботі представлено підходи до автоматизованого прогнозування ризику неінфекційних захворювань на основі аналізу текстів лікарських висновків із використанням сучасних інструментів машинного навчання. Проведено удосконалення попередньої моделі за рахунок перекладу текстів, лематизації, а також використання цілеспрямованого відбору медичних термінів із залученням мовної моделі GPT-3.5-turbo. Розширення підходу на прогнозування цереброваскулярних захворювань та попередня обробка тексту дозволили виявити сильні сторони запропонованих рішень. Дослідження показало високу ефективність моделі MultinomialNB (F1 = 86 %, AUC = 0,92) для задачі оцінки ризику.

Запропонований підхід до прогнозування ризику неінфекційних захворювань на основі аналізу текстів лікарських висновків має низку важливих переваг.

Модель дозволяє здійснювати як середньострокові, так і довгострокові прогнози щодо загострень найбільш небезпечних серцево-судинних захворювань, зокрема, інфаркту міокарда та інсульту. У перспективі підхід може бути адаптований для прогнозування ризику розвитку й інших неінфекційних хвороб.

Система забезпечує персоналізовану оцінку ризику для кожного пацієнта. Оцінювання ризику ґрунтується на комплексному аналізі медичної інформації: анамнезу, супутніх діагнозів, даних щодо призначеного лікування іншими спеціалістами, а також, за подальшого розвитку системи, – результатів ЕКГ, медичних зображень тощо.

Також реалізована можливість персоналізованого управління лікуванням на основі аналізу даних.

Таким чином, використання розробленого підходу сприяє підвищенню якості медичної допомоги за рахунок автоматизованої, всебічної та персоналізованої підтримки прийняття клінічних рішень.

Запропонована модель прогнозування ризику неінфекційних захворювань має широкий спектр можливих напрямків практичного застосування, що охоплює різні групи користувачів.

Лікувальні установи можуть використовувати результати прогнозу як додаткове джерело інформації для індивідуального планування лікування, моніторингу стану пацієнтів та визначення пріоритетності клінічних втручань.

Страхові компанії можуть застосовувати модель для контролю та оптимізації видатків, розрахунку страхових тарифів і покращення механізмів управління ризиками на популяційному рівні.

Пацієнти можуть отримати можливість підвищити якість життя завдяки інформації про власне здоров'я, активній участі у прийнятті рішень щодо лікування та профілактики, а також більш свідомому ставленню до рекомендацій медичних фахівців.

Державні та муніципальні служби можуть використовувати результати прогнозування для оптимізації фінансування та більш адресного розподілу ресурсів у сфері охорони здоров'я, впровадження програм профілактики та управління епідеміологічною ситуацією.

Навчальні центри можуть інтегрувати модель у програми підвищення кваліфікації лікарів та медичного персоналу, сприяючи впровадженню сучасних інструментів аналітики у практичну медицину.

Наукові установи отримують додаткові можливості для дослідження нових зв'язків між факторами ризику, перевірки гіпотез та розвитку доказової медицини шляхом роботи з великими масивами структурованих і неструктурованих даних.

Розроблена система випробувана на реальних даних та інтегрована у захищене середовище, що підтверджує її практичну придатність для використання у клінічній, страховій, дослідницькій та освітній діяльності. Модель має потенціал для подальшого розвитку – зокрема, за рахунок додавання нових джерел медичних даних, оптимізації структури ознак і впровадження ансамблевих методів. Визначено, що для підвищення якості прогнозування доцільно комбінувати статистичні методи відбору ознак із фаховою експертизою та враховувати як медичні, так і немедичні терміни.

Отримані результати демонструють перспективність використання машинного навчання для аналізу неструктурованих медичних текстів і підтримки прийняття рішень у системі охорони здоров'я.

**Наявність даних**. ДНУ ЦІТОЗ ДУС є власником оригінальної бази даних «Ескулап» та її деперсоналізованого фрагмента, що був використаний в даній роботі. Договір про співробітництво, в рамках якого був отриманий доступ до даних не дає нам право розповсюджувати їх, чи викладати у відкритий доступ.

**Фінансування**. Автор не отримував додаткового фінансування для проведення досліджень та написання статті.

**Список літератури**

1. Інсульт – не завжди крововилив: які є різновиди небезпечного захворювання. 2023**.** https://phc.org.ua/news/insult-ne-zavzhdi-krovoviliv-yaki-e-riznovidi-nebezpechnogo-zakhvoryuvannya (звернення 15.05.2025)
2. Празднікова М.О. Прогнозування і оцінка ризику інфаркту міокарду за сукупністю текстів лікарських висновків. *Кібернетика та комп`ютерні технології*. 2024. 3. С. 71–80. https://doi.org/10.34229/2707-451X
3. Making a difference - using 'big data' to shape patient care. https://www.uclhospitals.brc.nihr.ac.uk/making-difference-using-big-data-shape-patient-care (звернення 15.05.2025)
4. Singh M., Kumar A, etc. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review. 2024. **73** (3). https://doi.org/10.1016/j.eclinm.2024.102660
5. Shishehbori F., Awan Z. Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models. arXivLabs. 2024. https://doi.org/10.48550/arXiv.2401.17328
6. Rehman M.U., Naseem S., Butt A. et al. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Sci Rep*, 15, 13361. 2025. https://doi.org/10.1038/s41598-025-96437-1
7. Liu M., Liu Y., Liu J. Machine Learning for Infectious Disease Risk Prediction: A Survey. *ACM Computing Surveys*. 2024. **57** (8). P. 1–39. https://doi.org/10.1145/3719663
8. Bayes N. https://scikit-learn.org/stable/modules/naive_bayes.html (звернення 15.05.2025)
9. Navarro G. A guided tour to approximate string matching. ACM Computing Surveys (CSUR). 2001. **33** (1). P 31–88. https://doi.org/10.1145/375360.375365
10. spaCy: Промисловий рівень NLP для реальних застосувань. https://products.documentprocessing.com/uk/parser/python/spacy/#google_vignette (звернення 15.05.2025)
11. Becht E., Dutertre C.-A., Kwok I.W.H., Ng L.G., Ginhoux F., Newell E.W. Evaluation of UMAP as an alternative to t-SNE for single-cell data. The preprint server of biology. April 10, 2018. https://doi.org/10.1101/298430

Одержано 27.06.2025

**Празднікова Маргарита Олександрівна,**
аспірантка Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
prazdnikovamargarita@gmail.com

MSC 68T50, 68T20

**Margaryta Prazdnikova**

## Cardiovascular Disease Risk Prediction Models

*V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv*
*Correspondence: prazdnikovamargarita@gmail.com*

**Introduction.** Non-communicable diseases, especially cardiovascular pathologies, remain the leading cause of mortality worldwide, creating a significant burden on society, the economy, and healthcare systems. Heart attacks and strokes are particularly dangerous because they often develop suddenly and without symptoms, which complicates timely diagnosis and prevention. Identification of patients at increased risk can improve disease prevention and clinical outcomes, enhance the quality of medical care. In recent years, growing attention has been directed toward the use of artificial intelligence, machine learning, and big data processing techniques – particularly the analysis of unstructured medical texts – to improve the accuracy of medical predictions. The analysis of medical reports, patient histories, and other textual information can reveal hidden patterns that are inaccessible to traditional manual review and can greatly contribute to personalized treatment strategies.

**The aim** of the study is to improve the model for predicting the risk of myocardial infarction by introducing new methods of preprocessing medical reports and feature selection. In addition, the study aims to develop a new model for determining the risk level of cerebral vascular damage. The work focuses on integrating these models into modern information systems used in medical institutions and testing them on real clinical datasets.

**Results.** The study proposed and evaluated several approaches for improving myocardial infarction risk prediction, including text translation, lemmatization, and automated extraction of medical terms. Building on an extended version of the existing methodology, a new model was developed to predict cerebral vascular lesions. The analysis was conducted using the depersonalized “Eskulap” database, which contains records of more than 22,000 patients. The improved models demonstrated strong performance, achieving 80% accuracy (AUC = 0.898) for myocardial infarction and 86% accuracy (AUC = 0.92) for cerebral vascular lesions. The new model has already been successfully implemented in a medical center.

**Conclusions.** The proposed methods for improving the analysis of medical texts, including preprocessing, automated selection of relevant features, lemmatization, and adaptation to language-specific characteristics – enhanced the quality of risk prediction for cardiovascular and cerebrovascular diseases. The development of the new model for predicting cerebral vascular lesions further confirmed the effectiveness of this approach, and its implementation demonstrates the feasibility of integrating such solutions into clinical, insurance, and scientific practice. The model supports personalized prevention and treatment, facilitates the identification of high-risk groups, optimizes resource allocation, and improves clinical decision-making. It may also be used for calculating insurance rates or guiding targeted funding by governmental and municipal institutions.

The model also has strong potential for further development through the integration of additional data sources (such as laboratory indicators, instrumental examination results, and medical images), the adoption of more advanced ensemble algorithms, and deeper incorporation of expert assessments. Taken together, these results reinforce the conclusion that machine learning is a promising tool for analyzing unstructured medical texts, supporting clinical decision-making, and improving overall healthcare efficiency.

**Keywords:** non-communicable diseases, myocardial infarction, stroke, machine learning, risk prediction, Multinomial Naive Bayes, medical texts, data analysis.