Biomedical Engineering and Technology
Issue 12(4), 20223 ISSN (online) 2707-8434

УДК 004.8 + 616.12 + 519.254.3

ПОРІВНЯЛЬНИЙ АНАЛІЗ АНСАМБЛЕВИХ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ У ПРОГНОЗУВАННІ НАЯВНОСТІ ЗАХВОРЮВАНЬ СЕРЦЯ

Беспалов Ярослав Володимирович
y.bespalov-fbmi24@lll.kpi.ua
Настенко Євген Арнольдович
nastenko.e@gmail.com
Бабенко Віталій Олегович
vbabenko2191@gmail.com
кафедра біомедичної кібернетики
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
м. Київ, Україна

Реферат *– Серцево-судинні захворювання (ССЗ) продовжують бути провідною причиною летальних випадків та інвалідизації на глобальному рівні, становлячи загрозу для здоров'я мільйонів осіб. Незважаючи на значні досягнення в області медичних технологій, існують виклики, пов'язані з ранньою діагностикою та точним прогнозуванням ССЗ, що ускладнюється різноманітністю клінічних даних та складністю патологій. Дане дослідження має на меті оцінити ефективність застосування ансамблевих алгоритмів машинного навчання для прогнозування ССЗ, аналізуючи їх точність, надійність та інтегрованість з клінічними даними. Особлива увага приділяється потенціалу цих алгоритмів у вдосконаленні клінічного прогнозування та терапевтичних підходів до лікування ССЗ. Науковий проект фокусується на реалізації алгоритмів машинного навчання, зокрема ансамблевих методів, які застосовуються для створення моделей бінарної класифікації. Використовуються такі методи ансамблевого навчання, як Random Forest, XGBoost та LightGBM. Основна увага зосереджена на оптимальному розподілі даних для забезпечення точної оцінки, з використанням 10% даних для екзамену, 80% для тренування та 20% для тестування. Параметри моделей оптимізуються за допомогою 5-fold перехресної валідації. Модель Random Forest продемонструвала високу точність під час тренування, однак показала меншу точність під час тестування і екзамену, що може свідчити про перенавчання. У контрасті, моделі LightGBM та XGBoost показали більш стабільні результати на всіх етапах, зокрема LightGBM виявилася більш ефективною з точки зору швидкості навчання. Висновки дослідження підтверджують, що ансамблеві алгоритми машинного навчання, особливо LightGBM, є ефективними у прогнозуванні ССЗ. Результати також акцентують увагу на тому, що вік, систолічний кров'яний тиск та індекс маси тіла є ключовими індикаторами для оцінки ризику ССЗ.*

Ключові слова: *серцево-судинні захворювання, аналіз медичних даних, алгоритми прогнозування, машинне навчання, ансамблеве навчання*

I. ВСТУП

Серцево-судинні захворювання (ССЗ) залишаються однією з основних причин смерті населення світу, незважаючи на значні досягнення медичної науки та охорони здоров'я [1]. Розробка ефективних методів прогнозування ССЗ може відіграти вирішальну роль у профілактиці та лікуванні цих станів, що в кінцевому підсумку може значно покращити якість життя пацієнтів та зменшити витрати на охорону здоров'я [2].

Останні наукові розробки в галузі застосування методів машинного навчання для прогнозування ССЗ свідчать про значний прогрес у цій сфері. Мета-аналіз 55 досліджень [3] демонструє високу ефективність алгоритмів машинного навчання в прогнозуванні розвитку ішемічної хвороби серця. Було проаналізовано різноманітні алгоритми, включаючи згорткові нейронні мережі, машини опорних векторів, алгоритми бустінгу (англ. boosting), налаштовані алгоритми та методи випадкових лісів.

У роботі [4] автори використовували ансамблеві методи та глибокі нейронні мережі для точного виявлення серцевої

недостатності. Результати цієї роботи свідчать про значний потенціал глибокого навчання в галузі медичної діагностики, зокрема, у точному визначенні станів, пов'язаних із серцево-судинною системою. Це дослідження є прикладом інтеграції передових технологій машинного навчання в практичну медицину, що може допомогти підвищити точність методів діагностики і, як наслідок, поліпшити результати лікування пацієнтів із захворюваннями серця.

У дослідженні [5] автори розробили інноваційну модель стекінгу (англ. stacking) для аналізу впливу забруднення повітря та метеорологічних умов на частоту госпіталізації пацієнтів із ССЗ. Для реалізації цього підходу в дослідженні було використано комбінований набір різних класифікаторів, що дозволило побудувати більш комплексну і точну модель для прогнозування. Отримані результати підкреслюють важливість міждисциплінарного підходу до медичних досліджень і демонструють, як взаємодія між навколишнім середовищем і здоров'ям людини може бути більш ефективно оцінена за допомогою складних обчислювальних моделей.

У контексті використання машинного навчання для прогнозування ССЗ важливо визначити невирішені науково-технічні питання:

1. Хоча машинне навчання досягло значного прогресу в прогнозуванні ССЗ, є випадки, коли точність і надійність все ще недостатні. Необхідні подальші дослідження для визначення оптимальних параметрів та алгоритмів для покращення цих показників [3].

2. Важливим аспектом є інтеграція клінічних даних в алгоритми машинного навчання. Багато існуючих систем не враховують всю необхідну клінічну інформацію, яка є критично важливою для забезпечення точних прогнозів [4].

3. Інтерпретація результатів моделей машинного навчання часто буває складною. Необхідно розробити методи, які допоможуть медичним працівникам більш ефективно інтерпретувати ці результати, щоб підвищити якість обслуговування пацієнтів.

Дана робота є поглибленим дослідженням вищезазначених аспектів, зосереджуючись на ключових невирішених питаннях у використанні машинного навчання для прогнозування ССЗ.

II. МЕТА РОБОТИ

Визначення та аналіз можливості підвищення точності та надійності прогнозування серцево-судинних захворювань (ССЗ) за допомогою алгоритмів машинного навчання.

III. АНАЛІЗ ВХІДНИХ ДАНИХ

У дослідженні використано набір даних медичних записів пацієнтів, отриманих з загальнодоступного джерела *Kaggle* [6]. Він містить 63130 унікальних записів, з яких 31868 стосуються здорових осіб, а 31262 – пацієнтів із серцево-судинними захворюваннями. Дані включають наступні 11 параметрів:

1. Вік пацієнта, виміряний у днях, із середнім віком 53 роки і діапазоном від 39 до 65 років.

2. Гендер вибірки, включаючи 41060 чоловіків і 22070 жінок.

3. Зріст учасників дослідження, виміряний у сантиметрах, із середнім значенням 165 см і діапазоном від 141 до 188 см.

4. Вага, виміряна в кілограмах, із середнім значенням 73 кг і діапазоном від 37 до 110 кг.

5. Систолічний артеріальний тиск, виміряний у міліметрах ртутного стовпчика, із середнім значенням 126 мм рт.ст. і діапазоном від 90 до 174 мм рт.ст.

6. Діастолічний артеріальний тиск, також вимірюється в міліметрах ртутного стовпчика, із середнім значенням 82 мм рт.ст. і діапазоном від 63 до 107 мм рт.ст.

7. Рівень холестерину, з 47580 пацієнтами з нормальним рівнем, 8334 з підвищеним рівнем і 7216 з рівнем, значно вищим за норму.

8. Рівень глюкози: 53871 особа з нормальним рівнем, 4481 особа з підвищеним рівнем та 4778 осіб зі значно підвищеним рівнем.

9. Мітка куріння: 57653 пацієнти не курять і 5477 курять.

10. Мітка вживання алкоголю: 59818 не вживають алкоголь і 3312 вживають.

11. Мітка фізичної активності: 12369 пацієнтів, які не займаються фізичними вправами, і 50761 пацієнт, які займаються.

Ці показники, як відомо, є основними факторами ризику розвитку серцево-судинних захворювань [7]. Однак для досягнення більшої точності аналізу до початкового набору даних були внесені зміни:

• пункти 9-11 були вилучені через їх суб'єктивний характер та низьку кореляцію [8] з цільовою змінною, що ставить під сумнів достовірність даних самозвітів;

• зріст і вага пацієнтів були замінені на індекс маси тіла (IMT), що є практичним кроком для фокусування на ключовому показнику здоров'я;

• додано нові показники, такі як пульсовий тиск, середній артеріальний тиск та співвідношення систолічного до діастолічного тиску, що можуть надати додаткову інформацію про загальний стан артеріального тиску пацієнта.

У контексті аналізу даних лінійні моделі кореляції між цільовою змінною та вхідними даними часто є недостатніми для повної оцінки впливу останніх [8]. Вивчення нелінійних взаємозв'язків вимагає більш складних методологічних підходів. Одним з ефективних методів є використання локально оціненого згладжування діаграм розсіювання (англ. LOESS), яке дозволяє візуалізувати складні взаємозв'язки в даних. На рис. 1 показано діаграми розсіювання, згенеровані за допомогою цього методу.

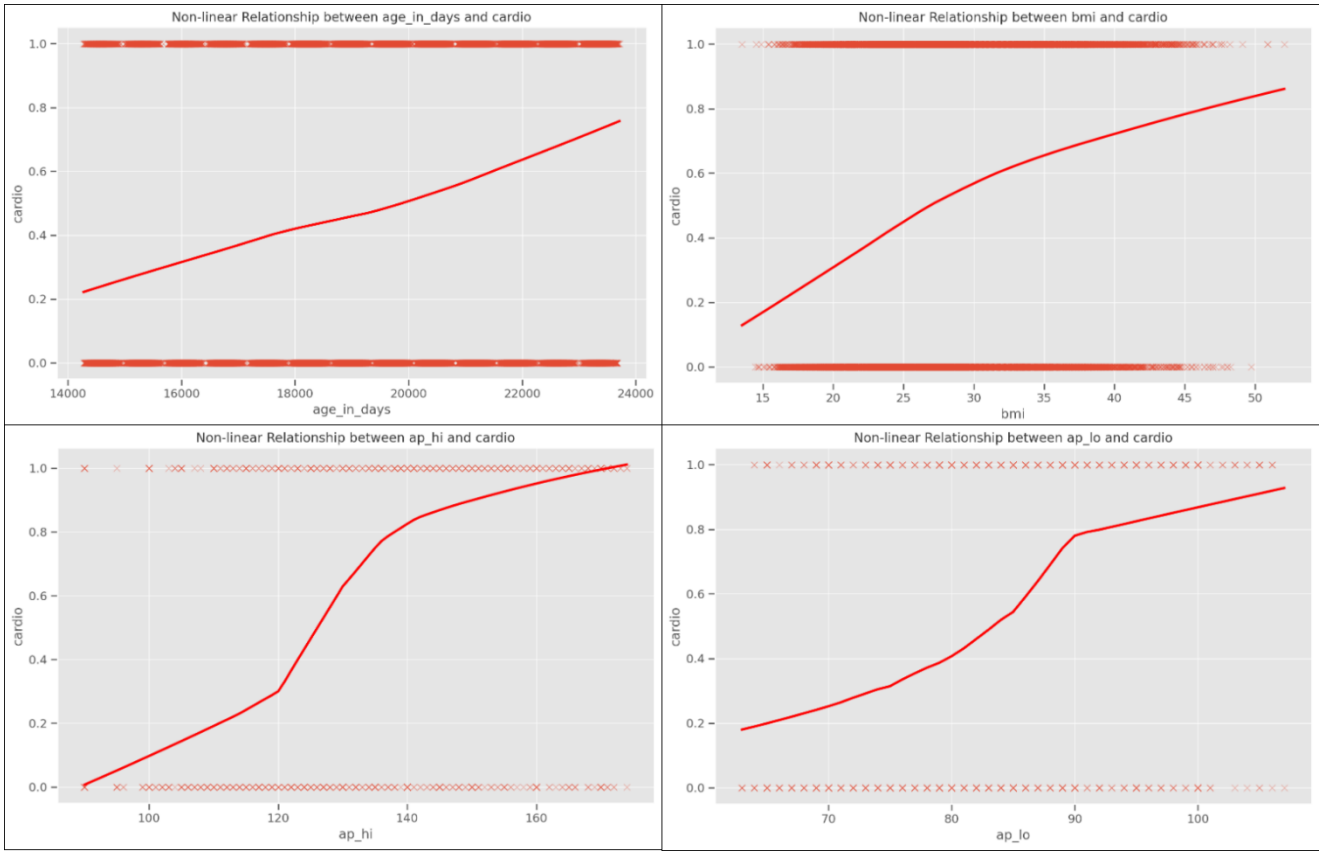

Рис. 1. Діаграми розсіювання основних показників пацієнтів

На графіках показані зв'язки між віком, IMT, систолічним та діастолічним тиском і серцево-судинними захворюваннями. Лінія LOESS показує, як змінюється ймовірність патології серця зі зміною кожного з цих параметрів, забезпечуючи візуальне представлення можливих нелінійних взаємозв'язків. Аналіз показує, що ймовірність зростає з віком та IMT і стає більш вираженою при високому рівні артеріального тиску, що

підкреслює необхідність диференційованого підходу до оцінки цих ризиків.

IV. МЕТОДИКА РЕАЛІЗАЦІЇ

Для розробки системи, яка прогнозує ризик серцево-судинних захворювань на основі клінічних даних, необхідно створити бінарні класифікаційні моделі. Оптимальним рішенням для цієї задачі є застосування методів машинного навчання, які, однак, можуть зіткнутися з труднощами при моделюванні нелінійних зв'язків у даних. У цьому контексті значний інтерес становлять ансамблеві методи машинного навчання, які здатні краще справлятися з такими завданнями [9].

Випадковий ліс (англ. Random Forest) [10], XGBoost [11] та LightGBM [12] визнані провідними ансамблевими алгоритмами завдяки їхній здатності ефективно працювати з великими наборами даних, забезпечувати високу точність та зменшувати перенавчання. Random Forest використовує множину дерев рішень для зменшення варіабельності, тоді як XGBoost і LightGBM оптимізують градієнтне підсилення, що значно підвищує швидкість навчання і продуктивність, особливо на великих і розріджених наборах даних, що робить їх еталоном у машинному навчанні. Крім того, нещодавні дослідження [13-14] підтверджують, що застосування модифікацій до подібних алгоритмів може призвести до покращення результатів.

У процесі моделювання важливим фактором, що впливає на її ефективність, є спосіб поділу даних на вибірки [15]. Вважається, що найкращою практикою є поділ всього набору даних на дві частини:

- навчання, де будується модель;
- тестування, де відбувається незалежна перевірка якості моделі.

Під час розробки моделі існує значний ризик перенавчання, що вимагає наявності валідаційної вибірки, яка використовується для оптимального налаштування параметрів моделі [15]. Валідаційний набір даних часто генерується в процесі *k*-fold перехресної валідації, що дозволяє оцінити модель та уникнути перенавчання, використовуючи різні підмножини даних для валідації і тренування [15]. Тактика полягає в тому, щоб розділити дані на *k* підмножин. Кожна підмножина використовується як тестова (валідаційна), а решта *k*-1 підмножин об'єднуються для формування тренувальної. Процес повторюється *k* разів, при цьому кожна з *k* підмножин використовується рівно один раз як тестова. Перевага цього методу полягає в тому, що він використовує всі доступні дані як для навчання, так і для тестування, гарантуючи, що кожне спостереження використовується лише один раз для тестування.

Прийнято вважати, що для забезпечення об'єктивності оцінки моделі вибірка для тестування повинна визначатися випадковим чином, що запобігає зміщенню і витоку даних [16]. Витік даних відбувається, коли використовується інформація, що не призначена для навчання. Це може призвести до переоцінки продуктивності моделі та зниження ефективності на нових даних.

В оглядовій статті [16] було виявлено 329 досліджень, які постраждали від витоку даних, що призвело до надмірно позитивних висновків. Автори виділили три основні категорії витоку даних:

1. У часовому просторі: відбувається, коли модель навчається з майбутнього моменту часу для прогнозування результатів у більш ранній момент часу. Це може призвести до нереалістичних результатів, оскільки модель має доступ до інформації, якої вона не мала б у реальному світі.

2. Подібність даних: виникає, коли в навчанні та тестуванні використовуються схожі або повторювані набори даних. Продуктивність моделі видається штучно високою, оскільки вона вже бачила подібні дані під час навчання.

3. Валідація: відбувається, коли інформація з тестування використовується в процесі розробки моделі, включаючи відбір ознак (англ. Feature Selection), налаштування параметрів моделі, нормалізацію даних тощо.

Таким чином було прийнято рішення реалізувати наступну стратегію моделювання

- виділити екзаменаційну вибірку, що складає 10% від загального набору даних, яка буде використана для додаткової (фінальної)

перевірки побудованих моделей (оскільки даних більш ніж достатньо, цей маневр дозволений);

• решту даних розділити на навчання (80%) і тестування (20%);

• навчання моделей Random Forest, XGBoost та LightGBM, використовуючи 5-fold перехресну валідацію для оптимізації параметрів налаштування (перевага надається тим параметрам, при яких модель показує найкращі результати в середньому на всіх 5 підмножинах валідації);

• оцінити отримані оптимізовані моделі на тестовій вибірці, використовуючи такі метрики, як точність, чутливість, специфічність та коефіцієнт кореляції Метьюса (ККМ);

• провести фінальне оцінювання на екзаменаційні вибірці для закріплення результату.

.

V. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Відповідно до визначеної стратегії моделювання були отримані наступні результати, що показані в табл. 1

Таблиця 1

Результати моделювання наявності серцево-судинного захворювання у пацієнтів

Ансамбль	Точність	Чутливість	Специфічність	ККМ
Навчання				
Random Forest	0.999	0.999	0.999	0.999
XGBoost	0.736	0.676	0.794	0.474
LightGBM	0.742	0.690	0.793	0.486
Тестування				
Random Forest	0.689	0.673	0.705	0.379
XGBoost	0.729	0.675	0.782	0.460
LightGBM	0.729	0.681	0.775	0.459
Екзамен				
Random Forest	0.694	0.674	0.715	0.389
XGBoost	0.724	0.666	0.782	0.451
LightGBM	0.726	0.680	0.772	0.454

Згідно з результатами, представленими в табл. 1:

1. Модель Random Forest показала високу точність прогнозування наявності серцево-судинного захворювання (99.9%) на етапі навчання, але результати були значно гіршими на тестуванні (68.9%) та екзамені (69.4%), що свідчить про можливе перенавчання побудованої моделі.

2. XGBoost показав більш збалансовані результати з точністю в 73.6% на тренуванні. Стабільність результатів прогнозування зберіглась на тестуванні (72.9%) та екзамені (72.4%).

3. LightGBM показав дещо гірші результати прогнозування на тестуванні (72.9% з вищою чутливістю та нижчою специфічністю) та кращі результати на екзамені (72.6%) порівняно з XGBoost.

Отже, було виявлено, що найефективнішими моделі забезпечують ансамблеві алгоритми машинного навчання LightGBM та XGBoost, які показали приблизно однакову точність у прогнозуванні наявності серцево-судинних захворювань на всіх трьох вибірках. Особливо слід відзначити LightGBM, яка показала вищу чутливість у прогнозуванні патологічних випадків на всіх етапах – навчання (чутливість 69%), тестування (68.1%) та екзаменаційної вибірки (68%).

Крім того, якщо порівняти швидкість навчання всіх трьох алгоритмів (рис. 2), то LightGBM також показує найкращий результат. В середньому: Random Forest навчається за 924,23 секунди (близько 16 хвилин), XGBoost за 298,77 секунди (близько 5 хвилин), а LightGBM за 72,38 секунди (близько 2 хвилин). Це означає, що LightGBM

у 2.5-3 рази випереджає XGBoost за швидкістю навчання моделі. Цей факт підтвердили самі автори алгоритму в [12], що свідчить про високу цінність його використання в моделюванні медичних даних, де, окрім точності прогнозування, ключовим фактором є швидкодія подібних моделей.

Таким чином, для подальшого дослідження основний акцент буде ставитись на алгоритмі LightGBM

.

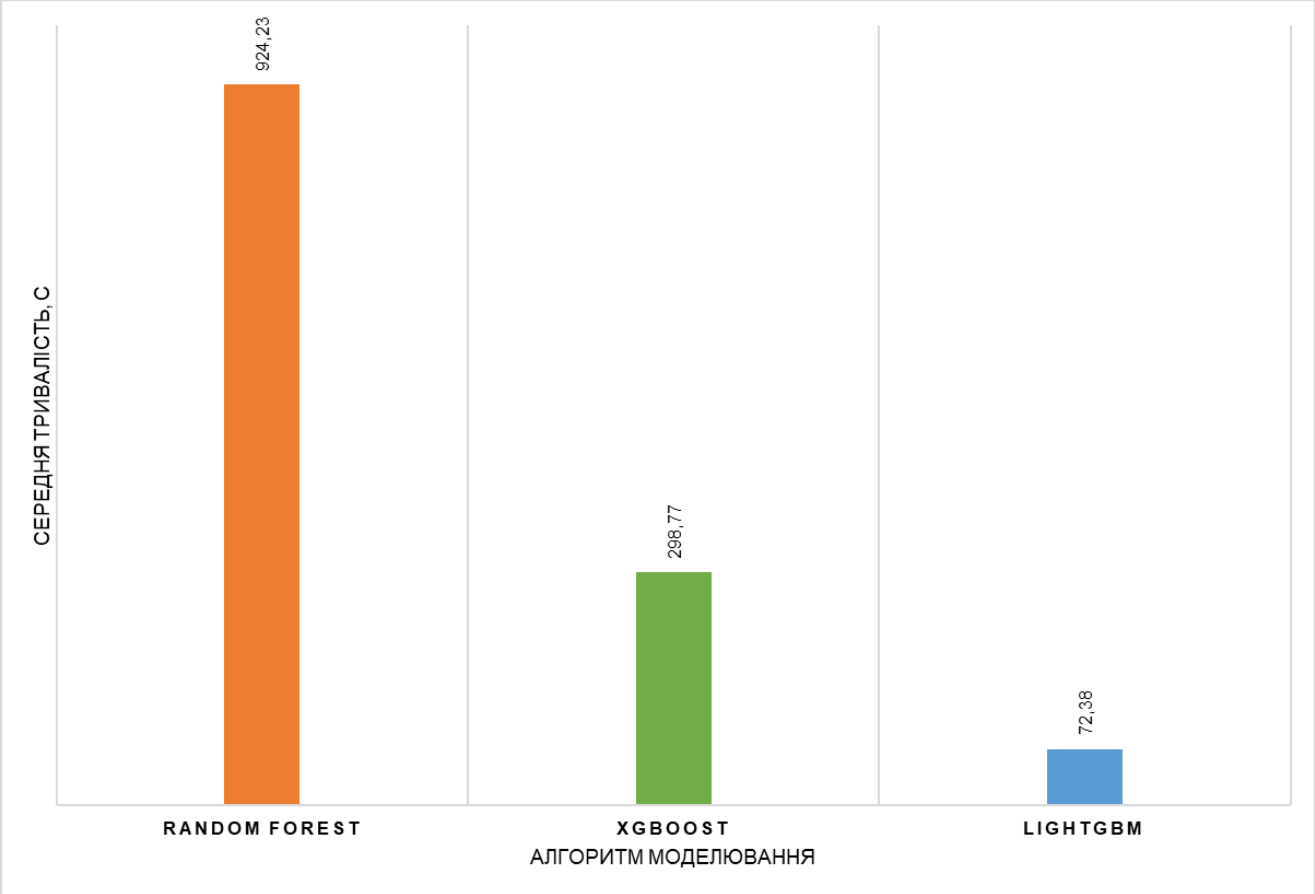

Рис. 2. Порівняння швидкості навчання алгоритмів ансамблевого машинного навчання

Модель LightGBM здатна оцінити важливість кожної характеристики, що використовується в моделюванні. Найбільшу цінність мають такі показники пацієнта:

• вік (чим старша людина, тим вищий ризик серцево-судинних захворювань);

• систолічний артеріальний тиск (знову ж таки, чим вище значення цього показника, тим вищий ризик);

• IMT (пацієнти зі значеннями нижче або вище норми знаходяться в групі ризику).

VI. ВИСНОВКИ

Відповідно до поставленої мети були використані методи машинного навчання, а саме ансамблеві алгоритми, для точного прогнозування наявності серцево-судинних захворювань у пацієнтів на основі їхніх медичних даних.

Було порівняно три основні алгоритми ансамблевого навчання: Random Forest, XGBoost та LightGBM. Серед трьох алгоритмів Random Forest продемонстрував найнижчий рівень точності прогнозування, досягнувши 69.4% на екзаменаційній (фінальній) вибірці.

Алгоритм LightGBM згенерував найкращу модель з точністю 72.6% на іспиті. Крім того, виявилось, що даний алгоритм навчається швидше, ніж інші моделі, що є додатковою перевагою.

Представлені LightGBM результати показують, що вік, систолічний артеріальний тиск та індекс маси тіла є найбільш

значущими показниками для оцінки ризику серцево-судинних захворювань, що корелює з існуючою клінічною літературою.

Що стосується обмежень цього дослідження, то точність прогнозування моделі LightGBM була посередньою через недостатню клінічну інформацію про пацієнтів.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

ORCID ID та внесок авторів.

1. Yaroslav Bespalov (A, B, C) 0009-0009-9167-592X
2. Ievgen Nastenko (F) – 0000-0002-1076-9337
3. Vitalii Babenko (D, E) – 0000-0002-8433-3878

A – Концепція роботи та дизайн, B – Проєктування та реалізація моделювання наявності серцево-судинних захворювань, C – Написання статті, D – Валідація побудованих моделей, E – Критичний огляд, F – Остаточне схвалення статті.

ПЕРЕЛІК ПОСИЛАНЬ

[1] Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216-9. doi: 10.1056/NEJMp1606181. PMID: 27682033; PMCID: PMC5070532.

[2] Kim JOR, Jeong YS, Kim JH, Lee JW, Park D, Kim HS. Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. Diagnostics (Basel). 2021 May 25;11(6):943. doi: 10.3390/diagnostics11060943. PMID: 34070504; PMCID: PMC8229422.

[3] Krittanawong, C., Virk, H.U.H., Bangalore, S. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 10, 16057 (2020). https://doi.org/10.1038/s41598-020-72685-1

[4] Srinivasan, S., Gunasekaran, S., Mathivanan, S.K. et al. An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. Sci Rep 13, 13588 (2023). https://doi.org/10.1038/s41598-023-40717-1

[5] Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K and Rohini K (2023) Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front. Med. 10:1150933. doi: 10.3389/fmed.2023.1150933

[6] Kaggle. Cardiovascular Disease dataset. 2018. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[7] Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10;14(10):e30119. doi: 10.7759/cureus.30119. PMID: 36381818; PMCID: PMC9644238.

[8] E. Szmidt and J. Kacprzyk, "The Spearman rank correlation coefficient between intuitionistic fuzzy sets," 2010 5th IEEE International Conference Intelligent Systems, London, UK, 2010, pp. 276-280, doi: 10.1109/IS.2010.5548399.

[9] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[10] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

[13] Babenko, V., Nastenko, I., Pavlov, V. et al. Classification of Pathologies on Medical Images Using the Algorithm of Random Forest of Optimal-Complexity Trees. Cybern Syst Anal 59, 346–358 (2023). https://doi.org/10.1007/s10559-023-00569-z

[14] Y. Hladkyi, O. Radchenko, V. Pavlov, O. Matviichuk and O. Horodetska, "A classifier of the Random Forest type based on GMDH, logistic transformation and positional voting," 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2023, pp. 1-4, doi: 10.1109/CSIT61576.2023.10324054.

[15] Оптимізація результатів моделювання шляхом розбиття вибірок за критерієм подібності відстані Махаланобіса / М. Гупало, В. Павлов, Є. Настенко, Г. Корнієнко. // Біомедична інженерія і технологія. – 2023. – №11. – С. 21–30.

[16] Leakage and the Reproducibility Crisis in ML-based Science, Sayash Kapoor and Arvind Narayanan, 2022, 2207.07048, arXiv

REFERENCES

[1] Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216-9. doi: 10.1056/NEJMp1606181. PMID: 27682033; PMCID: PMC5070532.

[2] Kim JOR, Jeong YS, Kim JH, Lee JW, Park D, Kim HS. Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. Diagnostics (Basel). 2021 May 25;11(6):943. doi: 10.3390/diagnostics11060943. PMID: 34070504; PMCID: PMC8229422.

[3] Krittanawong, C., Virk, H.U.H., Bangalore, S. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 10, 16057 (2020). https://doi.org/10.1038/s41598-020-72685-1

[4] Srinivasan, S., Gunasekaran, S., Mathivanan, S.K. et al. An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. Sci Rep 13, 13588 (2023). https://doi.org/10.1038/s41598-023-40717-1

[5] Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian

K, Anbarasu K and Rohini K (2023) Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front. Med. 10:1150933. doi: 10.3389/fmed.2023.1150933
[6] Kaggle. Cardiovascular Disease dataset. 2018. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
[7] Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10;14(10):e30119. doi: 10.7759/cureus.30119. PMID: 36381818; PMCID: PMC9644238.
[8] E. Szmidt and J. Kacprzyk, "The Spearman rank correlation coefficient between intuitionistic fuzzy sets," 2010 5th IEEE International Conference Intelligent Systems, London, UK, 2010, pp. 276-280, doi: 10.1109/IS.2010.5548399.
[9] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
[10] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157.
[13] Babenko, V., Nastenko, I., Pavlov, V. et al. Classification of Pathologies on Medical Images Using the Algorithm of Random Forest of Optimal-Complexity Trees. Cybern Syst Anal 59, 346–358 (2023). https://doi.org/10.1007/s10559-023-00569-z
[14] Y. Hladkyi, O. Radchenko, V. Pavlov, O. Matviichuk and O. Horodetska, "A classifier of the Random Forest type based on GMDH, logistic transformation and positional voting," 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2023, pp. 1-4, doi: 10.1109/CSIT61576.2023.10324054.
[15] Modeling results optimization based on data splitting by Mahalanobis distance similarity criterion / M. Hupalo, V. Pavlov, Ie. Nastenko, G. Kornienko. // Biomedichna injeneriya i technologiya. – 2023. – No. 11. – pp. 21–30.
[16] Leakage and the Reproducibility Crisis in ML-based Science, Sayash Kapoor and Arvind Narayanan, 2022, 2207.07048, arXiv

Biomedical Engineering and Technology
Issue 12(4), 20223 ISSN (online) 2707-8434

UDC 004.8 + 616.12 + 519.254.3

A COMPARATIVE ANALYSIS OF ENSEMBLE MACHINE LEARNING ALGORITHMS FOR PREDICTING THE PRESENCE OF CARDIOVASCULAR DISEASE

Yaroslav Bespalov
y.bespalov-fbmi24@lll.kpi.ua
Ievgen Nastenko
nastenko.e@gmail.com
Vitalii Babenko
vbabenko2191@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine

Abstract – *Cardiovascular disease (CVD) continues to be the leading cause of death and disability globally, threatening the health of millions of people. Despite significant advances in medical technology, there are challenges associated with early diagnosis and accurate prognosis of CVD, which is complicated by the diversity of clinical data and the complexity of pathologies. This study aims to evaluate the effectiveness of ensemble machine learning algorithms for CVD prediction by analyzing their accuracy, reliability, and integration with clinical data. Particular attention is paid to the potential of these algorithms to improve clinical prognosis and therapeutic approaches to CVD treatment. The research project focuses on the implementation of machine learning algorithms, in particular ensemble methods used to create binary classification models. The ensemble learning methods used are Random Forest, XGBoost and LightGBM. The focus is on optimal data distribution to ensure accurate scores, using 10% of the data for the exam, 80% for training, and 20% for testing. Model parameters are optimized using 5-fold cross-validation. The Random Forest model demonstrated high accuracy during training, but showed lower accuracy during testing and the exam, which may indicate overfitting. In contrast, the LightGBM and XGBoost models showed more stable results at all stages, with LightGBM proving to be more efficient in terms of learning speed. The findings of the study confirm that ensemble machine learning algorithms, especially LightGBM, are effective in predicting CVD. The results also emphasize that age, systolic blood pressure, and body mass index are key indicators for assessing CVD risk.*

Keywords *– Cardiovascular Diseases, Medical Data Analysis, Prediction Algorithms, Machine Learning, Ensemble Learning*