ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА КОМП'ЮТЕРНИХ НАУК

Кваліфікаційна робота

на тему: "**ДОСЛІДЖЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ ЗАХВОРЮВАНЬ НА ОСНОВІ ТЕХНОЛОГІЇ BIG DATA**"

на здобуття освітнього ступеня магістра
зі спеціальності __122 Комп'ютерні науки_____
(код, найменування спеціальності)
освітньо-професійної програми __Комп'ютерні науки _____
(назва)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ ___Гоменюк Андрій_____
(підпис) *Ім'я, ПРІЗВИЩЕ здобувача*

Виконав: здобувач вищої освіти
гр.КНДМ-62_____Андрій ГОМЕНЮК_____
(Ім'я, ПРІЗВИЩЕ)

Керівник: __к.т.н., доцент__Андрій ЧИЧУР_____
Науковий ступінь, вчене звання *(Ім'я, ПРІЗВИЩЕ)*

Рецензент: _____
Науковий ступінь, вчене звання *(Ім'я, ПРІЗВИЩЕ)*

Київ – 2024

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально–науковий інститут інформаційних технологій

Кафедра Комп'ютерних наук

Ступінь вищої освіти − «Магістр»

Спеціальність підготовки − «122 Комп'ютерні науки»

Освітньо-професійна програма − «Комп'ютерні науки»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Комп'ютерних наук
_____Віктор Вишнівський
"____" _____2024 року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Гоменюк Андрій Вікторович

(прізвище, ім'я, по батькові)

1.Тема кваліфікаційної роботи: «Дослідження системи прогнозування захворювань на основі технології Big Data»

Керівник кваліфікаційної роботи ____Андрій Чичур, к.т.н., доцент кафедри комп'ютерних наук

(ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «15» жовтня 2024р. № 320.

2. Строк подання студентом роботи 15.12.2024 р.

3. Вихідні дані до роботи: Серцево-Судинні Захворювання, Ліпопротеїн, Машинне Навчання, Прогнозування, Big Data, LightGBM, HPE OnDemand.

4. Зміст розрахунково-пояснювальної записки
1. Дослідження наявних технологічних інструментів прогнозування ССЗ.
2. Розробка системи прогнозування рівня ліпопротеїну(а) та потенційних ССЗ за допомогою технологій Big Data.

5. Перелік ілюстративного матеріалу: *презентація*
1. Мета роботи, об`єкт та предмет дослідження
2. Традиційні та технологічні інструменти діагностики ССЗ
3. Основні технології Big Data
4. Основні технології ML
5. Дослідження безпеки системи прогнозування захворювань
6. Огляд набір даних для дослідження
7. Визначення типу моделі для прогнозування
8. Характеристики навченої моделі
9. Метрики ефективності моделі
10. Додатки прогнозування на основі HPE OnDemand
11. Висновки

6. Дата видачі завдання 05.09.2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Затвердження теми дипломної роботи	15.10.2024	вик.
2	Підбір науково-технічної літератури	25.10.2024	вик.
3	Проведення дослідження за темою дипломної роботи	01.11.2024	вик.
4	Завершення роботи над першим варіантом частини дипломної роботи	15.11.2024	вик.
5	Проведення роботи над експериментальною частиною дипломної роботи	01.12.2024	вик.
6	Оформлення дипломної роботи та аналіз отриманих результатів	12.12.2024	вик.
7	Розробка демонстраційних матеріалів	13.12.2024	вик.

Здобувач вищої освіти _____ Андрій Гоменюк
(підпис) (Ім'я, ПРІЗВИЩЕ)

Керівник кваліфікаційної роботи _____ Андрій Чичур
(підпис) (Ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 70 стор., 3 рис., 1 табл., 1 додаток, 22 джерела посилання.

Мета дослідження – узагальнення наявних теоретичних підходів використання технологій Big Data для прогнозування ССЗ.

Об'єкт дослідження – процес прогнозування серцево-судинних захворювань.

Предмет дослідження – системи прогнозування ССЗ за допомогою технологій Big Data.

Короткий зміст роботи:

Проаналізовано наявні технологічні рішення в сфері прогнозування ССЗ. Визначено необхідні технології Big Data та ML для побудови системи прогнозування. Створено нову систему прогнозування ССЗ, що визначає рівень ліпопротеїну в крові без біолабораторних досліджень та прогнозує ймовірність розвитку ССЗ.

Було навчено і протестовано модель для передбачення рівня ліпопротеїну (а) в крові; було проведено аналіз метрик моделі і запропоновано методи імплементації дослідженої системи через API сервіс від HPE OnDemand.

Отриманий продукт дозволяє прогнозувати ймовірність розвитку ССЗ у пацієнтів, здійснювати корекцію догляду за власним здоров'ям, зменшувати фінансове навантаження на систему охорони здоров'я.

КЛЮЧОВІ СЛОВА: СЕРЦЕВО-СУДИННІ ЗАХВОРЮВАННЯ, ЛІПОПРОТЕЇН, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ, BIG DATA, LIGHTGBM, HPE ONDEMAND, ДАНІ.

ABSTRACT

Text part of the master's qualification work: 70 pages, 3 pictures, 1 table, 1 appendix, 22 sources.

The purpose of the work – summarize existing theoretical approaches to the use of Big Data technologies for cardiovascular disease prediction (CVD).

Object of research: the process of CVD predicting.

Subject of research: CVD prediction systems using Big Data technologies.

Summary of the work:

The existing technological solutions in the field of CVD prediction were analysed. The necessary Big Data and Machine Learning technologies for building a prediction system were identified. A novel CVD prediction system was developed, capable of determining blood lipoprotein levels without laboratory tests and predicting the likelihood of CVD development.

A model for predicting lipoprotein(a) levels in the blood was trained and tested; the model's metrics were analysed, and implementation methods were proposed via the HPE OnDemand API service.

The resulting product enables the prediction of CVD development risks for patients, supports personal health management adjustments, and reduces the financial burden on the healthcare system.

KEYWORDS: CARDIOVASCULAR DISEASES, LIPOPROTEIN, MACHINE LEARNING, PREDICTION, BIG DATA, LIGHTGBM, HPE ONDEMAND, DATA.

ЗМІСТ

Стор

ВСТУП

В атестаційній кваліфікаційної роботі магістра розглядається актуальна проблема обґрунтування доцільності застосування технологій Big Data для побудови системи прогнозування серцево-судинні захворювання (ССЗ).

Обґрунтування вибору теми та її актуальність. Серед головних причин природної смертності у світі визначено серцево-судинні захворювання (ССЗ). Основними видами ССЗ є ішемічна хвороба серця, інсульти, серцева недостатність, гіпертонія та її ускладнення, ревматичні хвороби серця тощо. Щороку від таких хвороб помирає близько 18 мільйонів людей за даними Всесвітньої організації охорони здоров'я. І ця цифра постійно зростає. Україна має близько 67% усіх смертей, що обумовлені ССЗ, це один із найвищих рівнів смертності у Європі [1]. Для подолання такого стану сучасна медицина в світі переходить від реактивного підходу (тобто від медичного втручання переважно після появи симптомів захворювань) до проактивного підходу (тобто застосування методів раннього виявлення та моніторингу захворювань, проведення статевих та вікових профілактичних оглядів для використання інтелектуальних технологій Big Data), з метою прогнозування ризиків захворювань. Одним з шляхів прогнозування ризиків рівня є визначення рівня ліпопротеїну (Lp(a)) в крові. Це є важливим кроком у сучасній профілактиці ССЗ. Дослідження системи прогнозування захворювань на основі технологій Big Data є актуальним.

Ступінь вивчення проблеми. Ступінь вивченості проблеми обґрунтування доцільності застосування технологій Big Data для побудови системи прогнозування ССЗ розглядається в багатьох роботах [1-7, 10-15], але є деякі прогалини в літературі та практиці.

Специфіка джерельної бази. Оцінка джерельної бази відзначає, що джерела інформації неоднорідні та її треба перебрати за певними критеріями щодо їх

різновиду, а саме: до важливості, якості, унікальності та ін., щоб відмітити проблеми та недоліки.

Мета дослідження – узагальнення наявних теоретичних підходів використання технологій Big Data для прогнозування ССЗ.

Метод дослідження. Метод дослідження заснований на відомих методах системного підходу, а саме: аналітично-розрахунковий, теоретичного аналізу з використанням моделювання, реконструювання і типологізації на основі аналізу емпіричних даних.

Завдання дослідження: проаналізувати наявні технологічні інструменти прогнозування ССЗ; виявити релевантні технології Big Data для побудови системи прогнозування ССЗ; розробити систему прогнозування ССЗ за допомогою технологій Big Data.

Об’єкт дослідження – Процес прогнозування серцево-судинних захворювань.

Предмет дослідження – Системи прогнозування ССЗ за допомогою технологій Big Data.

Методи дослідження – методи інтелектуального аналізу даних (Data Mining), опрацювання даних в реальному часі, машинне навчання за допомогою глибокого аналізу.

Джерела дослідження – деанонімізовані медичні дані Медико-дослідницького центру Family Heart Foundation (США).

Завдання роботи:

1. Провести аналіз передумов щодо обгрунтування доцільності застосування технологій BIG DATA для побудови системи прогнозування ССЗ.

2. Обґрунтувати доцільності застосування технологій Big Data для побудови системи прогнозування ССЗ.

3. Розробити пропозиції щодо побудови системи прогнозування ССЗ на основі технології BIG DATA.

Результати дослідження. Обґрунтоване необхідність застосування технологій Big Data для створення систему прогнозування ССЗ на основі

технологій Big Data, що визначає рівень ліпопротеїну в крові без біолабораторних досліджень та прогнозує ймовірність ССЗ.

Наукова новизна. Наукова новизна цього дослідження полягає в розробці пропозицій щодо побудови системи прогнозування ССЗ на основі технології BIG DATA.

Практичне застосування результатів. Отримані результати дозволяють прогнозувати ймовірність розвитку ССЗ у пацієнтів, здійснювати корекцію догляду за власним здоров'ям, зменшувати фінансове навантаження на систему охорони здоров'я.

Апробація результатів надається в 1 статті, 1 тезисах в рецензованих наукових журналах і виданнях.

Апробація результатів: Матеріали були опубліковані в тезисах:

Гоменюк А.В., Вишнівський В.В. ДОСЛІДЖЕННЯ БЕЗПЕКИ СИСТЕМИ ПРОГНОЗУВАННЯ ЗАХВОРЮВАНЬ НА ОСНОВІ ТЕХНОЛОГІЇ BIG DATA (5.11.2024) // науково-практична конференція «Актуальні проблеми безпеки інформаційно-телекомунікаційних систем» Збірник тез. – К.: ДУІКТ, 2024. 05 листопада 2024, С-38-43. https://duikt.edu.ua/uploads/p_2661_93669247.pdf

Гоменюк А.В., Вишнівський В.В. Сучасні досягнення компанії Hewlett Packard Enterprise в галузі ІТ та нові можливості їх вивчення і застосування (14.12.2024). // науково-практична конференція «Сучасні досягнення компанії Hewlett Packard Enterprise в галузі іт та нові можливості їх вивчення і застосування» - подане до збірнику тез. – К.: ДУІКТ, 2024. 14 грудня 2024, С. https://duikt.edu.ua/ua/2661-konferencii-2024-roku-konferencii-ta-seminari

1 АНАЛІЗ ПЕРЕДУМОВ ЩОДО ОБҐРУНТУВАННЯ ДОЦІЛЬНОСТІ ЗАСТОСУВАННЯ ТЕХНОЛОГІЙ BIG DATA ДЛЯ ПОБУДОВИ СИСТЕМИ ПРОГНОЗУВАННЯ ССЗ

1.1 Аналіз проблеми обґрунтування доцільності застосування технологій Big Data для побудови системи прогнозування ССЗ

Серцево-судинні захворювання становлять найбільшу загрозу здоров'ю населення у світі, забираючи близько 31% усіх смертей щороку. Згідно зі статистикою Всесвітньої організації охорони здоров'я (ВООЗ), більшість випадків смерті від ССЗ припадає на інфаркти та інсульти [1]. Смертність і інвалідність від цих захворювань викликають значне соціальне та економічне навантаження. Вплив захворювань ще більше посилюється через старіння населення та поширення факторів ризику, таких як ожиріння, малорухливий спосіб життя та стрес.

В Україні ситуація є особливо гострою. За даними Центру громадського здоров'я МОЗ України, серцево-судинні захворювання спричиняють понад 65% усіх смертей, ставлячи країну серед світових лідерів за рівнем смертності від ССЗ [2]. Серед основних причин виділяються ішемічна хвороба серця (ІХС), цереброваскулярні захворювання та гіпертонічна хвороба.

ССЗ призводять до значних економічних витрат, які включають прямі витрати на лікування (госпіталізації, медичні процедури, ліки) та непрямі втрати, пов'язані зі зниженням продуктивності через інвалідність чи передчасну смерть. Наприклад, у країнах із низьким і середнім рівнем доходу, де ресурси системи охорони здоров'я є обмеженими, це навантаження стає критичним, адже ресурси спрямовуються на лікування, а не на профілактику.

Основними факторами ризику розвитку ССЗ є нездоровий спосіб життя (незбалансоване харчування, багате насиченими жирами, цукром та сіллю, недостатня фізична активність), шкідливі звички (куріння та надмірне вживання алкоголю), хронічні захворювання (діабет, артеріальна гіпертензія, дисліпідемія)

та соціальні фактори (стрес, соціальна ізоляція, бідність та недостатній доступ до якісної медичної допомоги). Наприклад, регулярне куріння не лише безпосередньо пошкоджує стінки судин, але й сприяє утворенню атеросклеротичних бляшок. Брак фізичної активності веде до зниження серцево-судинної витривалості, що посилює ризик серцевої недостатності. Крім того, стрес, особливо хронічний, підвищує рівень гормонів, таких як кортизол, що може впливати на серцевий ритм і кров'яний тиск. Недостатній доступ до медичної допомоги особливо критичний у регіонах із низьким рівнем економічного розвитку, де рання діагностика захворювань часто недоступна.

Значну частину випадків ССЗ можна запобігти через зміни способу життя, а також за допомогою ранньої діагностики та втручання. Зокрема, здорове харчування з низьким вмістом насичених жирів та цукру, а також збагачення раціону овочами, фруктами та цільнозерновими продуктами значно знижує ризик серцевих захворювань. Регулярна фізична активність, навіть у помірному обсязі, зміцнює серце та покращує кровообіг. Рання діагностика, наприклад, через регулярні медичні огляди, дозволяє виявляти хвороби на ранніх стадіях, коли лікування є більш ефективним. Крім того, впровадження технологій, таких як носимі пристрої, що відслідковують показники здоров’я, може мотивувати людей до більш активного способу життя. Профілактика повинна включати освітні програми щодо здорового харчування та фізичної активності.

Такі програми мають охоплювати всі вікові групи, починаючи зі шкільного віку, формуючи культуру здорового способу життя з ранніх років. Освітні кампанії також повинні включати інформацію про шкідливий вплив куріння та вживання алкоголю на серцево-судинну систему. Важливо забезпечити доступність цих програм через медіа, школи та громадські організації, особливо в регіонах із низьким рівнем обізнаності населення. Інтерактивні інструменти, наприклад, мобільні додатки, можуть стати ефективним засобом підвищення мотивації до змін у поведінці.

Масовий скринінг і використання інноваційних інструментів прогнозування ризиків. Регулярні скринінгові обстеження дозволяють виявляти передумови

розвитку серцево-судинних захворювань ще до появи симптомів. Використання алгоритмів прогнозування на основі великих даних дозволяє оцінити ризики для окремих пацієнтів з урахуванням їхньої історії захворювань та способу життя. Інструменти штучного інтелекту вже використовуються для аналізу даних електрокардіограм і виявлення аномалій, які можуть свідчити про загрозу серцевого нападу. Крім того, сучасні платформи надають можливість інтегрувати результати скринінгу з персоналізованими рекомендаціями для пацієнтів.

Застосування машинного навчання для підтримки ухвалення клінічних рішень. Алгоритми машинного навчання допомагають лікарям аналізувати складні дані, такі як зображення медичних сканувань або результати лабораторних досліджень. Також ці алгоритми можуть передбачати реакцію пацієнтів на певні методи лікування, оптимізуючи підхід до терапії. Використання таких інструментів особливо ефективне у випадках, коли є велика кількість даних, яку неможливо швидко опрацювати вручну. Розробка моделей для прогнозування ускладнень дозволяє знижувати ризики і підвищувати ефективність лікування.

Генетичні передумови відіграють ключову роль у розвитку ССЗ. Певні генетичні мутації можуть змінювати функціонування білків, що впливають на кровообіг та стан судин. Також дослідження показують, що у деяких людей з певними генетичними особливостями ефективність ліків може значно відрізнятися. Знання про ці варіації дозволяють розробляти персоналізовані підходи до лікування, які враховують генетичний профіль пацієнта. Генетичні дослідження також допомагають ідентифікувати нові біомаркери, що можуть використовуватися для ранньої діагностики ССЗ [7, 8].

Наявність випадків ССЗ у найближчих родичів є важливим маркером генетичної схильності. Це вимагає від людей із сімейною історією захворювань більш уважного ставлення до свого здоров’я та регулярних обстежень. Генетичне тестування допомагає визначити конкретні мутації, які збільшують ризик, і вжити відповідних профілактичних заходів. Важливо, щоб лікарі враховували ці фактори під час оцінки ризиків та призначення лікування. Сімейна історія також стимулює

проведення досліджень, спрямованих на розробку ліків, що ефективніше впливають на пацієнтів із подібними генетичними особливостями.

Поведінкові аспекти значно впливають на розвиток ССЗ, оскільки багато з них є модифікованими, тобто такими, на які можна впливати. Наприклад, відмова від шкідливих звичок, таких як куріння, уже через кілька місяців знижує ризик серцевих нападів. Покращення харчування та збільшення фізичної активності можуть значно зменшити ризик розвитку гіпертонії та атеросклерозу. Також важливо враховувати психологічний стан: управління стресом через медитацію або психологічну підтримку сприяє зниженню рівня кортизолу та покращенню роботи серцево-судинної системи. Підвищення обізнаності населення про ці фактори є важливим компонентом профілактичних програм.

Варто зазначити, що взаємодія генетичних і поведінкових факторів може підсилювати ризик ССЗ. Наприклад, люди з генетичною схильністю до гіпертонії, які також мають шкідливі звички, можуть бути більш схильними до розвитку ускладнень [6]. Цей факт підкреслює необхідність поєднання генетичного аналізу з модифікацією поведінки як частини комплексного підходу до профілактики. Розуміння генетичних та поведінкових факторів дозволяє сформувати чітке уявлення про шляхи розвитку серцево-судинних захворювань. Впровадження генетичного тестування в клінічну практику разом із модифікацією поведінкових факторів може значно знизити захворюваність і смертність від ССЗ.

Актуальність проблеми серцево-судинних захворювань обумовлена не лише їх високою поширеністю, але й складністю діагностики на ранніх стадіях. Ускладнення діагностики часто виникають через відсутність очевидних симптомів на ранніх етапах захворювання. Інноваційні технології, такі як штучний інтелект, дозволяють виявляти тонкі зміни в даних, які можуть свідчити про розвиток патологій. Також використання мобільних пристроїв та телемедицини сприяє більш оперативному виявленню потенційних проблем. Це забезпечує можливість раннього втручання, що є ключовим для запобігання серйозним ускладненням. Впровадження машинного навчання дозволяє змінити парадигму боротьби із ССЗ, забезпечуючи більш точні прогнози, персоналізований підхід і підвищення

ефективності профілактики. Це відкриває нові перспективи для зниження захворюваності та смертності від серцево-судинних патологій.

1.2 Аналіз методів діагностики ССЗ

Традиційні методи діагностики серцево-судинних захворювань є основою для оцінки та ідентифікації різноманітних патологічних станів серцево-судинної системи. Ці методи охоплюють широкий спектр технік, що включають збір медичної історії пацієнта, фізичні обстеження та лабораторні дослідження.

Одним із основних етапів у процесі діагностики ССЗ є отримання детальної медичної історії пацієнта. Цей етап включає збір даних щодо особистої та сімейної медичної історії, факторів способу життя, а також симптомів, що спостерігаються у пацієнта. Через ретельне опитування лікар може виявити потенційні фактори ризику, такі як гіпертонія, діабет або наявність ССЗ у сімейному анамнезі. Оцінка медичної історії є важливим інструментом для ідентифікації пацієнтів, які потребують додаткових діагностичних досліджень.

Фізичні обстеження є необхідною складовою процесу діагностики серцево-судинних захворювань, оскільки вони забезпечують цінну інформацію про загальний стан здоров’я пацієнта та функціонування серцево-судинної системи. У ході фізичного огляду медичні фахівці оцінюють важливі показники, такі як артеріальний тиск, пульс та частоту дихання, що дає змогу виявити відхилення від норми, що можуть свідчити про наявність патології. Окрім того, лікар проводить оцінку фізичних ознак, які можуть допомогти в уточненні діагнозу та подальшому лікуванні. Загалом варто ознайомитися з традиційними методами діагностики, щоб усвідомити їх вади для повного обстеженням пацієнтів з СС.

Аускультація серцевих звуків є одним із основних методів діагностики, що включає прослуховування звуків, які видає серце, за допомогою стетоскопа. Ця техніка дозволяє медичним фахівцям виявляти аномальні серцеві звуки, відомі як серцеві шуми, які можуть бути ознакою різних порушень, таких як стеноз або

недостатність клапанів серця, а також структурних дефектів. Зокрема, серцеві шуми можуть вказувати на порушення нормальної роботи серцевих клапанів, що потребує додаткової діагностичної оцінки. Аускультація є важливим інструментом для виявлення таких відхилень та своєчасного призначення додаткових досліджень, таких як ехокардіографія, для точного визначення діагнозу.

Пальпація периферичних пульсів є важливою складовою фізичного обстеження, яка надає лікарям цінну інформацію про якість і регулярність кровообігу. За допомогою пальпації пульсів, що вимірюються в різних артеріях, таких як променева артерія на зап’ясті, сонна артерія на шиї, пахова артерія та артерії на дорсальній поверхні стопи, медичний спеціаліст може оцінити силу, ритм та симетрію пульсу. Це допомагає виявити потенційні артеріальні перешкоди, а також серцево-судинні розлади, зокрема периферичну артеріальну хворобу. Пальпація пульсу є важливим методом для первинної оцінки стану серцево-судинної системи та може спрямовувати пацієнта на подальші обстеження для підтвердження діагнозу.

Виявлення ознак затримки рідини або периферичного набряку. Під час фізичного обстеження лікарі також здійснюють візуальний огляд пацієнта для виявлення ознак затримки рідини або периферичного набряку. Це включає огляд кінцівок на наявність набряків або припухлостей, що можуть бути сигналами серцевої недостатності, порушення функції нирок або венозної недостатності. Виявлення периферичного набряку є важливим для визначення основної причини серцево-судинних симптомів і для подальшої діагностики. Це дозволяє лікарям зібрати додаткову інформацію, необхідну для точного визначення патологічного стану та планування лікування. За допомогою фізичного огляду медичні фахівці можуть виявити можливі серцево-судинні відхилення та визначити необхідність проведення додаткових діагностичних тестів.

Лабораторні аналізи крові є основними інструментами для діагностики серцево-судинних захворювань, оскільки вони дозволяють виявити важливі біохімічні маркери, що відображають стан серцево-судинної системи. Одним з основних досліджень є аналіз ліпідного профілю, який включає оцінку рівнів

загального холестеролу, ліпопротеїнів низької щільності (LDL), ліпопротеїнів високої щільності (HDL) та тригліцеридів. Підвищені рівні LDL і тригліцеридів у поєднанні зі зниженими рівнями HDL є значущими показниками, що свідчать про підвищений ризик розвитку серцево-судинних захворювань, зокрема атеросклерозу.

Крім того, лабораторні аналізи дозволяють виявити супутні захворювання, такі як діабет або порушення функції нирок, що також є важливими факторами ризику для розвитку ССЗ. Регулярне моніторингування цих показників є важливим для своєчасного виявлення змін у стані здоров’я пацієнта та корекції лікування.

Електрокардіографія (ЕКГ) є неінвазивною методикою, що дозволяє записувати електричну активність серця та виявляти різноманітні серцеві відхилення. ЕКГ є важливим інструментом для діагностики аритмій, порушень провідності, а також ознак міокардіальної ішемії. Процедура включає прикріплення електродів до грудної клітки та кінцівок пацієнта, що дозволяє вимірювати електричні сигнали, генеровані серцем. Інтерпретація результатів ЕКГ дозволяє медичному фахівцю встановити точний діагноз, а також визначити подальшу тактику лікування.

Ехокардіографія є однією з основних неінвазивних методик для оцінки структури та функції серця, яка використовує ультразвукові хвилі для створення зображень. Цей метод дозволяє детально оцінити розмір, форму і рух серцевих камер, а також функцію серцевих клапанів. Окрім того, ехокардіографія дає можливість виміряти фракцію викиду серця, що є важливим показником його функціонального стану. У разі виявлення структурних дефектів, таких як вроджені вади серця або патології клапанів, ехокардіографія є незамінним інструментом для діагностики та планування подальшого лікування.

Стрес-тести, також відомі як тести з фізичним навантаженням, використовуються для оцінки реакції серця на підвищену фізичну активність. Під час тесту пацієнт виконує фізичні вправи, наприклад на біговій доріжці або велотренажері, при цьому моніториться серцева активність, артеріальний тиск та

виникнення симптомів. Стрес-тести є ефективними для виявлення коронарної артеріальної хвороби, оскільки вони дозволяють оцінити здатність серця постачати достатню кількість кисню до міокарду під час навантаження. Ненормальні зміни в серцевому ритмі, артеріальному тиску або виникнення болю в грудях під час виконання тесту можуть свідчити про порушення кровопостачання серцевого м'яза.

В останні роки досягнуто значних проривів у розвитку діагностичних технологій, що сприяло впровадженню передових методів діагностики серцево-судинних захворювань. Інноваційні техніки використовують сучасні технології, включаючи засоби зображення, біомаркери та обчислювальні підходи, для забезпечення більш точної та всебічної оцінки серцево-судинного здоров'я. Метою цього підрозділу є аналіз новітніх діагностичних методів серцево-судинних захворювань, а також виявлення їхніх переваг, обмежень і потенційного впливу на клінічну практику.

Сучасні методи серцевого зображення значно перевершують традиційні підходи до діагностики серцево-судинних захворювань, надаючи можливість детально візуалізувати структуру, функцію та кровотік серця. Це дозволяє медичним фахівцям здійснювати точнішу діагностику та оцінку різноманітних серцевих патологій. До основних технік серцевого зображення, що використовуються сьогодні, відносяться магнітно-резонансна томографія серця (МРТ) (методика дозволяє отримувати високоякісні зображення серця, що дозволяють оцінити як його структуру, так і функцію); комп'ютерна томографічна ангіографія (КТА) (використовується для детального вивчення судин серця, включаючи коронарні артерії, що дає змогу виявити атеросклероз або інші обструкції); позитронно-емісійна томографія (ПЕТ) (ця техніка надає цінну кількісну інформацію, що сприяє вивченню метаболічних процесів у серці, таких як кровопостачання міокарда).

Ці інструменти дозволяють медичним фахівцям отримувати зображення з високою роздільною здатністю, що сприяє більш точному виявленню та оцінці

серцево-судинних захворювань. Вони є надзвичайно важливими для комплексної діагностики та планування лікування пацієнтів із серцевими захворюваннями.

Магнітно-резонансна томографія серця є високоточним діагностичним методом, що дозволяє отримувати високоякісні тривимірні зображення серця та його структур. Ця методика базується на використанні потужних магнітних полів і радіочастотних хвиль для отримання деталізованих зображень серцевих камер, клапанів, судин і м’язів серця. МРТ є незамінним інструментом для оцінки серцевої функції, визначення розміру порожнин серця, обсягу крові, а також виявлення структурних аномалій. Завдяки своїй високій роздільній здатності, магнітно-резонансна томографія застосовується для діагностики таких захворювань, як аномалії розвитку серця, міокардит, інфільтративна кардіоміопатія та ішемічна хвороба серця.

Комп’ютерна томографія є не менш важливим методом, що використовує рентгенівські промені та комп’ютерну обробку для створення високоякісних зображень серцевих структур. Цей метод дозволяє отримати тривимірні зображення, що включають коронарні артерії, аорту та інші важливі судини серця. КТ-ангіографія є ефективним способом для виявлення таких патологій, як коронарна артеріальна хвороба, артеріальні аневризми, легеневі емболії та інші судинні захворювання. Окрім того, цей метод дозволяє оцінити прохідність коронарних стентів, що є важливим для контролю ефективності лікування [5].

Ядерна медицина, зокрема методи однофотонної емісійної комп'ютерної томографії (SPECT) та позитронно-емісійної томографії (PET), використовує радіоактивні речовини для отримання зображень серцево-судинної системи. Метод SPECT базується на використанні радіоактивних фармакологічних агентів, які приєднуються до специфічних молекул у серці, що дозволяє візуалізувати кровотік і оцінити метаболічну активність. У свою чергу, PET використовує радіоактивні ізотопи, які вводяться в організм і захоплюються активними клітинами серця. Це дозволяє здійснити молекулярну візуалізацію процесів в серцевому м’язі та оцінити функцію серця, що є важливим для діагностики ішемії та інших серцево-судинних розладів [6].

Генетичні тести стають важливим інструментом для діагностики серцево-судинних захворювань, особливо у виявленні успадкованих форм патологій серцево-судинної системи. Аналіз генетичного матеріалу дозволяє виявляти специфічні мутації або варіанти генів, що корелюють із розвитком серцево-судинних захворювань [9]. Генетичні тести надають цінну інформацію щодо спадкових чинників, які можуть підвищувати ризик виникнення серцевих захворювань, що особливо важливо для осіб з позитивним сімейним анамнезом.

Такі тести дають можливість не тільки оцінити ризик розвитку серцево-судинних захворювань у родичів, але й допомагають визначити індивідуальні стратегії лікування та профілактики, що враховують специфіку генетичної схильності кожного пацієнта [10]. Це дозволяє більш точно прогнозувати можливі кардіологічні проблеми та забезпечити персоналізований підхід до лікування, що значно підвищує ефективність терапевтичних заходів.

Застосування обчислювальних підходів, зокрема машинного навчання та штучного інтелекту, є перспективним напрямом для покращення точності та швидкості діагностики серцево-судинних захворювань. Ці методи аналізують великі обсяги даних за допомогою алгоритмів і моделей, що дозволяє виявляти приховані закономірності та прогнозувати ризики розвитку серцевих захворювань. Машинне навчання може обробляти клінічні дані, медичні зображення, біомаркери та інші важливі параметри для створення моделей, які допомагають виявляти ранні ознаки захворювань серцево-судинної системи, прогнозувати ризик розвитку ускладнень і підтримувати процес прийняття рішень щодо лікування [10].

Завдяки інтеграції таких обчислювальних методів, як штучний інтелект, в клінічну практику, стало можливим значно підвищити ефективність діагностики. Вони дозволяють не лише визначати потенційні захворювання на ранніх стадіях, але й персоналізувати підхід до лікування, що є критично важливим для оптимізації терапевтичних результатів.

Використання передових діагностичних методів, таких як техніки серцевого зображення, біомаркери, генетичні тести та обчислювальні підходи, є значним

кроком вперед у діагностиці серцево-судинних захворювань. Ці інновації дозволяють отримати більш точну та інформативну оцінку серцево-судинного здоров'я, виявляти ранні ознаки патологій, прогнозувати ризики та формулювати індивідуалізовані підходи до лікування пацієнтів.

Поєднання традиційних та новітніх методів діагностики серцево-судинних захворювань має великий потенціал для підвищення ефективності та точності діагностики. Інтеграція цих методик дозволяє медичним фахівцям отримати комплексну і детальну інформацію, що сприяє більш глибокому розумінню стану здоров’я пацієнта. Використання як традиційних методів (наприклад, клінічний огляд, аускультація, фізичні обстеження), так і новітніх технологій, таких як магнітно-резонансна томографія, комп’ютерна томографія та обчислювальні підходи (штучний інтелект і машинне навчання), дозволяє виявляти захворювання на ранніх стадіях. Це не лише сприяє своєчасному лікуванню, але й підвищує ймовірність повного відновлення пацієнтів [7].

Однак при впровадженні таких технологій варто враховувати ряд чинників: ефективність витрат, доступність сучасних методів та необхідну кваліфікацію медичних працівників для роботи з новими діагностичними інструментами. Використання новітніх методів діагностики вимагає також належної технічної підтримки та регулярного оновлення навичок медичного персоналу.

Продовжуючи розгляд сучасних діагностичних підходів, наступним етапом є детальніше вивчення ролі машинного навчання у діагностиці серцево-судинних захворювань, яке має потенціал значно змінити підходи до медичної діагностики [11].

Машинне навчання (ML) є сучасною науковою галуззю, яка активно використовується в різних сферах, включаючи медицину. В діагностиці серцево-судинних захворювань машинне навчання вже знайшло широке застосування і демонструє обіцяні результати. Розглянемо основні напрямки використання ML в цій сфері.

Машинне навчання вже успішно застосовується в багатьох аспектах діагностики серцево-судинних захворювань, зокрема для прогнозування ризику

розвитку серцевих захворювань на основі клінічних даних та факторів ризику аналізувати великі обсяги даних і виявляти складні взаємозв'язки між факторами ризику та розвитком серцево-судинних захворювань. Це дозволяє створювати індивідуалізовані прогнози ризику і впроваджувати превентивні заходи для запобігання хворобам [15].

Крім того, ML широко використовується для аналізу медичних зображень, отриманих за допомогою таких технік, як ехокардіографія, МРТ та КТ. Використання нейго навчання дозволяє автоматично виявляти аномалії, оцінювати структуру та функцію серця, а також здійснювати точну діагностику. Наприклад, моделі ML можуть виявляти патологічні зміни на зображеннях ЕКГ, що значно сприяє ранньому виявленню порушень серцевого ритму та інших серцевих захворювань.

Дослідження показують, що застосування машинного навчання в діагностиці серцево-судинних захворювань демонструє високу ефективність. Моделі ML не тільки виявляють підвищення ризику виникнення серцевих захворювань, але й можуть класифікувати різні типи хвороб, а також підтримувати лікарів у прийнятті рішень. Вони покращують точність діагностики та допомагають виявляти складні взаємозв'язки між численними факторами [12].

Використання машинного навчання також здатне значно пришвидшити діагностичний процес і зменшити навантаження на медичний персонал. Автоматизація аналізу даних і розпізнавання зображень дозволяє швидше і точніше ставити діагнози, знижуючи ймовірність помилок і підвищуючи ефективність лікування [3].

Машинне навчання вже відіграє важливу роль у діагностиці серцево-судинних захворювань. Його застосування показало високу ефективність у прогнозуванні ризику розвитку захворювань, аналізі медичних зображень, класифікації патологій і підтримці прийняття рішень медичними фахівцями. Зокрема, машинне навчання значно прискорює діагностичний процес, знижує ймовірність помилок і покращує точність діагностичних висновків [14].

Однак важливо враховувати певні обмеження та труднощі, пов’язані з використанням машинного навчання в медицині. До таких викликів відносяться проблеми зі збором і обробкою великих обсягів медичних даних, етичні питання, пов'язані з використанням особистої медичної інформації, а також складність в інтерпретації результатів, зокрема в умовах обмеженого доступу до висококваліфікованих медичних спеціалістів. Тому інтеграція машинного навчання в клінічну практику потребує ретельного підходу до навчання моделей, забезпечення високої якості даних та врахування етичних норм [13].

Подальший розвиток технологій машинного навчання та їхня інтеграція в клінічну практику відкривають нові можливості для поліпшення точності діагностики та ефективності лікування серцево-судинних захворювань, що може значно покращити якість медичної допомоги для пацієнтів.

У цьому розділі було проведено комплексне дослідження предметної області, пов’язаної з ССЗ. Ми ретельно вивчили наявну наукову літературу та дослідження, що охоплюють різні аспекти ССЗ, зокрема їхню епідеміологію, фактори ризику та методи діагностики. Особлива увага була приділена основним показникам здоров’я серцево-судинної системи, таким як артеріальний тиск, рівень холестерину, куріння, спадковість, вік та інші чинники, які мають вагоме значення для оцінки ризику розвитку ССЗ.

Ми також провели аналіз класифікації та типів серцево-судинних захворювань, їхніх симптомів, ускладнень, а також методів лікування та профілактики. Дослідження цієї предметної області дозволило нам глибше зрозуміти проблему ССЗ, їхній вплив на громадське здоров’я та необхідність розробки ефективних стратегій для діагностики й попередження цих захворювань.

У рамках цього розділу ми також розглянули основні методи діагностики ССЗ, зокрема клінічні огляди, лабораторні дослідження та інструментальні методи, такі як ЕКГ, ультразвукове дослідження, магнітно-резонансна томографія та КТ. Крім того, було вивчено сучасні методи, що базуються на машинному навчанні та аналізі медичних даних, що забезпечує нові можливості для прогнозування і діагностики серцево-судинних захворювань.

Загалом, проведене дослідження предметної області надало нам необхідні знання та контекст для подальшого розроблення алгоритмів машинного навчання, які використовуються для прогнозування ризику ССЗ. Це дозволяє створити більш точні та індивідуалізовані підходи до діагностики та лікування пацієнтів.

1.3 Аналіз застосування технологій побудови системи прогнозування ССЗ

Для побудови системи прогнозування ССЗ використовуються різноманітні технології, які забезпечують збір, обробку, аналіз даних і формування прогнозів. Нижче наведено ключові технології:

1. Збір та зберігання даних:

Інтернет речей (IoT): Носимі пристрої (фітнес-трекери, смарт-годинники) для моніторингу фізіологічних параметрів, таких як частота серцевих скорочень, артеріальний тиск, рівень активності. Медичне обладнання (кардіомонітори, апарати ЕКГ).

Хмарні технології: AWS, Google Cloud, Microsoft Azure для зберігання та обробки великих обсягів даних. Інтеграція із системами електронних медичних записів (EMR).

2. Обробка даних

Big Data: Інструменти, такі як Apache Hadoop, Apache Spark, забезпечують паралельну обробку великих обсягів медичних даних. NoSQL бази даних (MongoDB, Cassandra) для зберігання неструктурованих даних.

ETL-процеси: Інструменти для екстракції, трансформації та завантаження даних (Talend, Apache NiFi).

3. Алгоритми машинного навчання (ML):

Методи класифікації та регресії: Логістична регресія, дерева рішень, Random Forest, SVM.

Глибоке навчання (Deep Learning): Нейронні мережі (CNN, RNN, LSTM) для аналізу складних даних, таких як зображення серця (УЗД, МРТ) або динаміка ЕКГ.

Алгоритми кластеризації: k-means, DBSCAN для визначення груп пацієнтів із подібними характеристиками.

Градієнтний бустинг: XGBoost, LightGBM для точних прогнозів із врахуванням великої кількості факторів.

4. Інтелектуальний аналіз даних

Data Mining: Інструменти, як-от Weka, RapidMiner, для пошуку закономірностей у великих наборах медичних даних.

Аналіз часових рядів: Інструменти на основі LSTM або ARIMA для обробки динамічних даних про стан здоров'я.

5. Інтерпретація та візуалізація

Інструменти BI (Business Intelligence): Tableau, Power BI, Qlik для створення інтуїтивно зрозумілих звітів і візуалізації прогнозів.

Методи інтерпретованості моделей: LIME, SHAP для пояснення рішень, прийнятих моделями машинного навчання.

6. Інтеграція та взаємодія з користувачем

API та мікросервіси: Використання REST або GraphQL для інтеграції з медичними платформами.

Мобільні додатки: Розробка інтерфейсів для пацієнтів і лікарів, що надають доступ до персоналізованих рекомендацій.

7. Безпека та конфіденційність

Кібербезпека: Протоколи захисту даних, такі як HTTPS, шифрування AES-256. Системи аутентифікації (OAuth, SAML).

Регуляторні стандарти: Відповідність нормам HIPAA, GDPR для захисту персональних даних.

8.Приклади реальних технологій:

TensorFlow та PyTorch – для створення нейронних мереж.

Scikit-learn – для традиційних ML-алгоритмів.

Keras – для прототипування моделей глибокого навчання.

Apache Kafka – для потокової обробки даних у реальному часі.

Docker та Kubernetes — для масштабування системи прогнозування.

Таким чином ці технології спільно створюють ефективну, масштабовану та інтегровану систему для прогнозування ССЗ.

1.4 Аналіз можливості застосування технологій BIG DATA для побудови системи прогнозування ССЗ

Застосування технологій Big Data для побудови систем прогнозування ССЗ є одним із перспективних напрямів у сучасній медицині. Завдяки обробці великих обсягів даних стає можливим створення точних моделей, які допомагають прогнозувати ризики, виявляти схильності до захворювань і приймати обґрунтовані рішення щодо лікування.

Основні аспекти застосування Big Data в прогнозуванні ССЗ наступні:

1. *Збір даних:* Електронні медичні записи (EMR); дані з фітнес-трекерів, смарт-годинників та інших носимих пристроїв; результати аналізів та обстежень (ЕКГ, КТ, МРТ); генетична інформація (геноміка); дані про спосіб життя пацієнтів (харчування, рівень фізичної активності, стрес).

2. *Обробка даних:* Використання хмарних платформ для зберігання та обробки великих обсягів інформації; методи очищення даних для усунення помилок та дублікатів; інтеграція даних з різних джерел у єдину систему.

3. *Аналіз даних:* Машинне навчання (ML): створення моделей прогнозування, які враховують безліч факторів; аналіз часових рядів: відстеження змін стану здоров'я пацієнта; аналіз факторів ризику: виявлення взаємозв’язків між генетикою, способом життя і захворюваннями.

4. *Прогнозування та інтерпретація:* Розробка систем раннього попередження про ризик виникнення ССЗ; надання рекомендацій лікарям щодо профілактичних заходів або індивідуальних програм лікування.

5. *Реалізація та інтеграція:* Впровадження у лікарняні інформаційні системи; використання мобільних додатків для інформування пацієнтів; інтеграція з телемедичними платформами.

6. *Переваги використання Big Data:* точність (врахування багатьох факторів дозволяє знизити кількість помилкових прогнозів), персоналізація (створення індивідуальних планів лікування та профілактики), економія ресурсів (зниження витрат на лікування завдяки ранньому виявленню захворювань) та масштабованість (можливість обробки даних для мільйонів пацієнтів).

7. *Виклики:* Конфіденційність: забезпечення безпеки персональних даних. Якість даних: необхідність уніфікації форматів і структур даних. Складність інтерпретації: необхідність пояснення лікарям результатів алгоритмів.

8. Реальні приклади:

- Framingham Heart Study – використання історичних медичних даних для прогнозування серцево-судинних ризиків.
- IBM Watson Health – аналіз медичних даних для виявлення схильності до ССЗ.
- Google Fit і Apple Health – обробка даних із носимих пристроїв для моніторингу стану здоров'я.

Таким чином технології Big Data вже роблять значний внесок у медицину, і їхнє вдосконалення може суттєво покращити прогнозування та лікування ССЗ.

1.5 Аналіз можливості застосування технологій машинного навчання для побудови системи прогнозування ССЗ

Застосування технологій машинного навчання (ML) для побудови систем прогнозування ССЗ є важливим напрямом у сучасній медицині. ML дозволяє створювати моделі, що аналізують великі обсяги медичних даних і визначають взаємозв’язки між численними факторами ризику.

Основні етапи використання ML у прогнозуванні ССЗ наступні:

1. Збір та підготовка даних:

- Джерела даних: Електронні медичні записи (EMR). Дані носимих пристроїв (фітнес-трекери, смарт-годинники). Результати обстежень (аналізи

крові, ЕКГ, УЗД). Дані про спосіб життя (харчування, активність, стрес). Генетична інформація.

- Попередня обробка: Усунення пропущених або аномальних значень. Балансування набору даних (наприклад, у разі нерівномірного розподілу здорових і хворих пацієнтів). Нормалізація або стандартизація даних.

2. Розробка моделей машинного навчання:

- Вибір моделі: Логістична регресія: базовий метод для визначення ймовірності захворювання. Дерева рішень та Random Forest: визначають важливість різних факторів ризику. Градієнтний бустинг (XGBoost, LightGBM): висока точність у багатофакторному аналізі. Нейронні мережі: для складних прогнозів і аналізу великих наборів даних. Кластеризація (k-means): визначення груп пацієнтів із подібними ризиками. Обробка часових рядів: LSTM або GRU для аналізу динаміки показників здоров'я.
- Особливості навчання: Використання метрик оцінки (AUC-ROC, F1-score) для об'єктивної оцінки моделі. Перехресна валідація для запобігання перенавчанню.

3. Інтерпретація результатів:

- Аналіз важливості ознак: Визначення основних факторів ризику (вік, тиск, рівень холестерину тощо). Виявлення несподіваних кореляцій.
- Інтерпретованість моделі: Використання методів, таких як LIME або SHAP, для пояснення рішень моделі.

4. Інтеграція в медичну практику:

- Програми та платформи: Мобільні додатки для пацієнтів, що надають персоналізовані рекомендації. Інтеграція з лікарняними інформаційними системами. Системи моніторингу в режимі реального часу для пацієнтів із груп ризику.
- Прогностичні системи: Виявлення пацієнтів із високим ризиком інфаркту або інсульту. Прогнозування ускладнень після операцій.

5. Переваги ML у прогнозуванні ССЗ:

- Висока точність: моделі враховують складні нелінійні взаємозв’язки між факторами ризику.
- Персоналізація лікування: створення індивідуальних рекомендацій для кожного пацієнта.
- Автоматизація: мінімізація людських помилок.
- Економія ресурсів: зниження витрат завдяки ранньому виявленню захворювань.

6. Виклики та обмеження:

- Якість даних: потреба у великій кількості чистих і стандартизованих даних.
- Конфіденційність: забезпечення захисту медичних даних пацієнтів.
- Інтерпретованість: складність пояснення рішень складних моделей (особливо нейронних мереж).
- Етичні аспекти: правильне використання прогнозів для ухвалення рішень.

7. Приклади використання ML у прогнозуванні ССЗ:

- Cleveland Heart Disease Dataset – базовий набір даних для створення моделей прогнозування серцево-судинних ризиків.
- DeepHeart від Cardiogram – нейронна мережа, яка використовує дані смарт-годинників для прогнозування серцевих захворювань.
- IBM Watson Health – інструмент для аналізу медичних даних і прогнозування ризиків.

Таким чином впровадження машинного навчання у прогнозування ССЗ здатне революціонізувати підходи до профілактики та лікування цих захворювань, значно покращуючи якість медичної допомоги.

2 ОБҐРУНТУВАННЯ ДОЦІЛЬНОСТІ ЗАСТОСУВАННЯ ТЕХНОЛОГІЙ BIG DATA ТА МАШИННОГО НАВЧАННЯ ДЛЯ ПОБУДОВИ СИСТЕМ ПРОГНОЗУВАННЯ ССЗ

2.1 Огляд основних понять і термінологій Big Data

Big Data (Великі дані) – це термін, що використовується для опису величезних обсягів структурованих, напівструктурованих та неструктурованих даних, які надзвичайно складні для обробки традиційними методами обробки даних та інструментами. Вони значно перевищують обсяги даних, з якими можуть ефективно працювати звичайні бази даних чи системи для аналітики. Відмінність між Big Data та традиційними даними полягає не тільки в їхньому обсязі, але й у швидкості їх створення та необхідності використовувати нові технології для їх зберігання, обробки та аналізу. У традиційних даних зазвичай використовуються реляційні бази даних і стандартні аналітичні інструменти, що не здатні ефективно працювати з такими великими та різноманітними обсягами.

У Big Data існує 3 головні *V*, які визначають головні особливості цієї сфери.

Volume (Обсяг). Це характеристика, що відображає величезний обсяг даних, який генерується щодня. Наприклад, лише в соціальних мережах кожну хвилину завантажується тисячі годин відео та мільйони постів. Обсяг даних може досягати багатих петабайтів чи ексабайтів, що вимагає особливих інструментів для їх зберігання та обробки. Великі обсяги даних вимагають масштабованих архітектур, таких як розподілені системи і хмарні сервіси, щоб уникнути проблем із продуктивністю. Зростання обсягу даних ставить питання щодо оптимізації індексації, пошуку та управління даними.

Variety (Різноманітність). Великий обсяг даних може бути представленим у різних форматах: від структурованих (наприклад, таблиці баз даних) до напівструктурованих (JSON, XML) та неструктурованих даних (тексти, відео, зображення). Кожен з цих типів вимагає різних підходів до зберігання, обробки і аналізу. Різноманітність даних створює виклики для інтеграції та сумісності

різних форматів, що ускладнює процеси аналітики. Для ефективної роботи з різними типами даних необхідні спеціалізовані інструменти, які здатні обробляти як текстові, так і мультимедійні файли.

Velocity (Швидкість). Швидкість, з якою генеруються, передаються та обробляються дані, є ключовим аспектом Big Data. Данні потоки, як, наприклад, потоки даних від сенсорів Інтернету речей (IoT), необхідно обробляти в режимі реального часу або з мінімальними затримками для забезпечення своєчасної реакції. Це особливо важливо в застосунках, де необхідна миттєва реакція, таких як системи фінансових операцій або моніторинг безпеки. Використання інструментів обробки потокових даних, таких як Apache Kafka та Apache Flink, дозволяє реалізувати обробку великих потоків даних без затримок.

Також в технологіях Big Data виділяють наступні додаткові вимірювання:

Veracity (Правдивість). Це відображає якість даних. Дані можуть бути не точними, неповними або суперечливими, що ускладнює їхнє використання для прийняття рішень. Забезпечення правдивості і чистоти даних є важливим аспектом їх аналізу. Для цього часто застосовують методи перевірки достовірності, фільтрації шуму та виявлення аномалій. Недостатня якість даних може призвести до значних помилок у прогнозуванні або аналітиці, що ставить під загрозу прийняття стратегічних рішень. Тому важливо використовувати інструменти, які забезпечують валідацію та очищення даних перед їх подальшим використанням.

Value (Цінність). Хоча обсяг і швидкість створення даних є важливими, їх цінність для прийняття рішень є найбільш важливою характеристикою. Визначення, які дані є корисними, дозволяє зосередитись на тих, що реально можуть принести вигоду в бізнесі чи науці. Наприклад, аналіз поведінки клієнтів може допомогти розробляти персоналізовані пропозиції, збільшуючи лояльність і продажі. Однак низькоцінні дані можуть створювати зайве навантаження на системи зберігання та обробки, тому їх слід мінімізувати. Здатність витягувати реальну цінність із великих даних залежить від правильного поєднання технологій збору, обробки та аналітики.

Однак, без прикладних технологій сфера Big Data не змогла б існувати. Серед найпоширеніших технологій можна виділити наступні:

Hadoop – це одна з основних технологій для обробки великих даних, яка дозволяє обробляти величезні обсяги інформації за допомогою розподіленої обчислювальної архітектури. Основні компоненти Hadoop включають:

Hadoop Distributed File System (HDFS). Це файлова система, що дозволяє зберігати великі обсяги даних, розподілених по багатьох комп'ютерах (вузлах). HDFS забезпечує високий рівень надійності та стійкості даних за рахунок реплікації. Дані автоматично дублюються на декілька вузлів, що мінімізує ризик втрати інформації через вихід з ладу окремого комп'ютера. Система також оптимізована для обробки великих файлів, що робить її ідеальною для аналізу великих обсягів інформації. HDFS працює за принципом "один раз записати, багато разів читати", що підвищує ефективність зберігання даних. Крім того, вона підтримує горизонтальне масштабування, дозволяючи легко додавати нові вузли для збільшення обсягу зберігання.

MapReduce. Це модель програмування, що дозволяє обробляти великі дані паралельно на розподілених комп'ютерах. Вона складається з двох етапів: Map (розподіл даних) і Reduce (агрегація результатів). На етапі Map дані розбиваються на частини та обробляються незалежно, що значно прискорює виконання завдань. На етапі Reduce результати зводяться в єдине ціле для отримання остаточного результату. Ця модель добре підходить для завдань, що потребують обробки великих масивів неструктурованих даних. MapReduce дозволяє автоматизувати розподіл завдань і зменшує навантаження на окремі вузли системи. Використання MapReduce спрощує реалізацію складних алгоритмів аналізу даних завдяки чітко визначеним етапам обробки.

YARN (Yet Another Resource Negotiator). Це система управління ресурсами в Hadoop, що відповідає за координацію та керування розподіленими обчислювальними ресурсами. YARN дозволяє запускати кілька різних типів обчислювальних завдань на одній Hadoop-кластері, ефективно розподіляючи ресурси між ними. Вона відокремлює управління ресурсами від виконання

завдань, що забезпечує більшу гнучкість системи. Завдяки YARN можна одночасно запускати різні аналітичні процеси, наприклад, MapReduce і Spark, у межах одного середовища. Вона включає модуль ResourceManager, який координує доступ до ресурсів, і NodeManager, що відповідає за обслуговування вузлів. YARN підтримує динамічний розподіл ресурсів, дозволяючи масштабувати обчислення залежно від навантаження. Ця система є важливою частиною сучасних Hadoop-кластерів, забезпечуючи їхню адаптивність і ефективність.

Hadoop є дуже масштабованим і використовує концепцію “розподіленої обробки”, що дозволяє ефективно працювати з величезними обсягами даних. Його використовують великі компанії для зберігання і обробки даних у таких сферах, як фінанси, охорона здоров’я та телекомунікації. Завдяки розподіленій архітектурі Hadoop дозволяє зберігати та аналізувати великі набори медичних даних, включаючи електронні медичні записи, результати тестів та демографічну інформацію пацієнтів. Ця платформа може використовуватися для виявлення закономірностей і факторів ризику ССЗ за допомогою машинного навчання та аналізу даних. Hadoop підтримує інтеграцію з інструментами Big Data, такими як Apache Spark, для прискорення аналізу складних медичних даних і створення прогностичних моделей. У дослідженнях ССЗ Hadoop забезпечує ефективність у виконанні задач, пов’язаних із раннім виявленням захворювань, прогнозуванням ризиків і покращенням стратегій профілактики.

Apache Spark – це ще одна потужна платформа для обробки великих даних, яка значно перевершує Hadoop в аспекті швидкості обробки завдяки обробці даних в пам’яті (in-memory processing). Spark підтримує обробку даних у реальному часі, що робить його ідеальним для застосувань, які потребують швидкої реакції. Spark дозволяє обробляти потоки даних у реальному часі за допомогою модулю Spark Streaming. Spark дозволяє значно швидше обробляти дані в порівнянні з Hadoop завдяки використанню пам’яті для обробки даних. Spark використовує більш простий синтаксис для програмування і підтримує багатоплатформеність (може працювати на різних типах файлових систем).

Spark дозволяє аналізувати великі обсяги медичних даних, таких як записи пацієнтів, дані про ліки, лабораторні тести та результати моніторингу. Він підтримує інтеграцію з мовами програмування, такими як Python і R, що полегшує розробку машинного навчання для прогнозування ризиків ССЗ. Spark також може обробляти потокові дані в реальному часі, що є корисним для моніторингу стану пацієнтів і раннього виявлення аномалій. Завдяки можливості паралельної обробки даних Apache Spark допомагає розробляти ефективні прогностичні моделі та підтримувати рішення в сфері охорони здоров'я [16].

NoSQL – це категорія баз даних, які відрізняються від традиційних реляційних баз даних за моделлю зберігання та обробки даних. Вони розроблені для обробки великих обсягів неструктурованих даних, зокрема даних соціальних мереж, журналів та IoT даних.

Ця категорія баз даних має 4 основних різновиди: key-value stores; document stores; column-family stores; graph databases.

Key-Value Stores. Це найпростіший тип NoSQL баз, де кожен запис складається з пари "ключ-значення". Ідеально підходять для сценаріїв, де потрібен швидкий доступ до значень за унікальним ключем, наприклад, кешування, управління сесіями користувачів або збереження налаштувань додатків. Серед переваг визначається швидкість операцій читання та запису, простота моделі зберігання. Однак, є і недолік – обмежена функціональність у порівнянні з іншими типами NoSQL баз, наприклад, відсутність складних запитів. Приклади: Redis, Riak, Amazon DynamoDB.

Document Stores. Зберігають дані у вигляді документів, зазвичай у форматі JSON, BSON, або XML. Можуть зберігати складні структури даних, що робить їх корисними для додатків, які працюють із неструктурованими або напівструктурованими даними, наприклад, систем керування контентом (CMS) або інтернет-магазинів. Забезпечує гнучкість схеми (schema-less) та підтримку складних запитів і агрегатів. Однак, менш ефективні для транзакційних систем, які потребують високої консистентності даних. Приклади: MongoDB, Couchbase, CouchDB.

Column-Family Stores. Бази даних, які зберігають дані в стовпцях замість рядків, що робить їх ефективними для аналітики та великих обчислень. Дані групуються у “сімейства стовпців”, що дозволяє легко отримувати доступ до певних наборів даних. Застосовуються для зберігання журналів подій, історичних даних або аналітики через масштабованість та ефективність при роботі з великими наборами даних. Є стійкими до збоїв в записі даних. Разом з тим є складність у налаштуванні та управлінні, потреба в розумінні структури даних для оптимальної роботи. Приклади: Apache Cassandra, HBase.

Graph Databases. Орієнтовані на зберігання та обробку графових структур, таких як вузли (nodes) і зв’язки (edges). Використовуються в додатках, які працюють із складними взаємозв'язками між об'єктами, наприклад, соціальні мережі, системи рекомендацій або аналіз мережевих структур. Серед переваг висока ефективність для операцій, які вимагають багаторівневих зв’язків і складних запитів до графів, але присутня Менша продуктивність при роботі з великими наборами даних без явних зв’язків. Приклади: Neo4j, ArangoDB, Amazon Neptune.

NoSQL бази даних зазвичай обирають для додатків, де необхідно працювати з великими обсягами даних, високою швидкістю доступу або складними структурами, які важко вписуються у традиційні реляційні моделі.

NoSQL бази даних використовуються для масштабованих додатків, де необхідна швидка обробка великих обсягів неструктурованих даних або даних із високою швидкістю оновлення. Вони ідеально підходять для великих вебсайтів, мобільних додатків, IoT та соціальних мереж. Загалом такі бази даних можуть зберігати великі обсяги медичних записів, генетичну інформацію, результати аналізів та інші дані пацієнтів, що є важливими для виявлення ризиків ССЗ. Завдяки високій швидкості запису і зчитування, NoSQL сприяє аналізу потокових даних у реальному часі, наприклад, для моніторингу стану серця пацієнтів. Гнучка модель даних дозволяє інтегрувати різноманітні джерела інформації, такі як демографічні дані, спосіб життя та екологічні фактори, для розробки більш точних моделей прогнозування. NoSQL підтримує масштабованість і стійкість, що є

критично важливим для медичних систем, які обробляють величезні масиви даних у дослідженнях і профілактиці ССЗ.

MapReduce – це програмна модель і технологія для паралельної обробки великих обсягів даних. Вона складається з двох основних етапів: Map (розподіл роботи на маленькі підзадачі, де кожна підзадача обробляє частину даних) та Reduce (об’єднання результатів обробки в один кінцевий результат).

MapReduce активно використовувався в Hadoop для паралельної обробки даних на великих обчислювальних кластерах. Моделі, які виконуються за допомогою MapReduce, дозволяють розподіляти дані по різних вузлах і виконувати обчислення незалежно, що значно підвищує продуктивність обробки великих обсягів інформації. Ця модель стала основою для багатьох сучасних систем обробки великих даних, таких як Apache Spark. Технологія MapReduce дозволяє обробляти великі обсяги медичних даних, отриманих із електронних медичних карток, моніторів життєвих показників та баз даних пацієнтів, що значно прискорює аналіз. Вона також може використовуватися для ідентифікації закономірностей у даних, таких як зв'язок між способом життя, генетичними факторами та ризиком серцево-судинних захворювань. Крім того, MapReduce сприяє швидкому та ефективному аналізу медичних зображень, таких як кардіограми чи МРТ серця, для виявлення патологій. Технологія активно застосовується для створення моделей машинного навчання, які прогнозують ризик серцево-судинних захворювань на основі багатовимірних даних. Завдяки паралельній обробці даних, MapReduce дозволяє значно скоротити час аналізу та покращити точність діагностичних і прогностичних моделей у кардіології.

Обробка даних у реальному часі (streaming data) є важливим аспектом сучасної аналітики великих даних, оскільки дозволяє отримувати, обробляти та аналізувати інформацію без затримок, що є необхідним для багатьох бізнесових і наукових застосувань. Обробка даних у реальному часі включає захоплення, обробку та аналіз потокових даних, що надходять від різноманітних джерел, таких як сенсори, соціальні мережі, фінансові ринки, мобільні пристрої, тощо. Завдяки цій технології, компанії можуть миттєво реагувати на події, що відбуваються в

реальному часі, замість того, щоб чекати на аналіз великих зібраних даних. Для обробки потокових даних часто використовуються такі платформи, як Apache Kafka, Apache Flink та Apache Storm, які дозволяють ефективно працювати з даними в реальному часі, здійснювати їх зберігання та аналіз у потоці [16].

Обробка потокових даних має критичне значення для ряду галузей, таких як фінанси, телекомунікації та e-commerce. Наприклад, у фінансовому секторі компанії використовують реальний моніторинг ринкових даних, щоб відстежувати зміни цін на акції в реальному часі і здійснювати автоматизовану торгівлю. Це дозволяє реагувати на зміни на фінансових ринках миттєво, приймаючи рішення на основі актуальної інформації.

Для роздрібної торгівлі аналітика потокових даних допомагає персоналізувати маркетингові кампанії. Amazon та інші компанії використовують ці технології для збору даних про поведінку користувачів, що дозволяє здійснювати персоналізовані рекомендації в реальному часі, збільшуючи таким чином продажі та залучення клієнтів.

Також важливе застосування для управління інфраструктурою та обслуговуванням: для предиктивного обслуговування (наприклад, у транспортних мережах або на виробництві) збираються дані з сенсорів машин і обладнання, що дозволяє передбачити можливі поломки до їхнього виникнення і здійснити необхідні заходи.

Вчені також використовують обробку в реальному часі для аналізу великих обсягів даних у наукових дослідженнях. Наприклад, у медичних дослідженнях використання потокових даних допомагає здійснювати моніторинг стану пацієнтів у реальному часі, що дає можливість оперативно реагувати на зміни стану хворого. Технології візуалізації і аналізу геномних даних дозволяють науковцям здійснювати глибокий аналіз біологічних процесів, таких як аналіз генетичних даних, обробляючи великі потоки біомедичних даних у реальному часі.

У кліматичних науках обробка даних у реальному часі використовуються для моніторингу та прогнозування погодних умов, зміни клімату, відстеження природних катастроф (повені, землетруси, урагани), що дозволяє здійснювати

швидку оцінку ситуації та вжити необхідних заходів з попередження та ліквідації наслідків.

Обробка потокових даних і аналітика великих даних у реальному часі є важливими інструментами, що використовуються в бізнесі та науці для отримання швидких і точних результатів на основі великого обсягу даних. Ці технології допомагають бізнесу підвищувати ефективність операцій, персоналізувати обслуговування клієнтів і оптимізувати виробничі процеси. У науці вони дозволяють здійснювати своєчасні дослідження та приймати інформовані рішення для покращення стану здоров'я, екології та розуміння складних природних процесів.

2.2 Огляд основних понять і термінологій машинного навчання

Машинне навчання (ML) є інноваційною технологією в сфері інформаційних технологій, яка радикально змінює традиційні підходи до аналізу та інтерпретації складних наборів даних. ML включає набір алгоритмів і методів, що дозволяють комп'ютерам навчатися на основі даних, виявляти значущі закономірності та здійснювати прогнози або приймати рішення без необхідності явного програмування. У контексті діагностики серцево-судинних захворювань машинне навчання відіграє важливу роль, оскільки дає змогу працювати з великими обсягами даних для виявлення прихованих закономірностей і покращення точності діагностики.

Розуміння основних принципів та понять машинного навчання є необхідним для ефективного застосування цієї технології в медицині. Цей розділ надає всебічне визначення машинного навчання та досліджує його основні компоненти, типи та важливість даних у процесі навчання.

Машинне навчання є підгалуззю штучного інтелекту, яка зосереджується на розробці алгоритмів і моделей, здатних вивчати закономірності у даних і використовувати ці знання для здійснення обґрунтованих прогнозів або прийняття рішень. На відміну від традиційного програмування, де кожна дія чітко визначена,

в ML алгоритми вивчають дані через ітераційний процес, дозволяючи комп'ютеру самостійно адаптуватися та генерувати рішення на основі наданої інформації.

Основними компонентами машинного навчання є:

- Дані: Вони є основою для навчання моделей, надаючи приклади та інформацію, що дозволяє моделі виявляти закономірності та взаємозв'язки. Дані можуть бути структурованими (наприклад, таблиці з числовими або категоріальними значеннями) або неструктурованими (зображення, аудіо, текст). Якість даних має критичне значення, оскільки шум, відсутність даних чи їхня незбалансованість можуть вплинути на ефективність моделі. Збір даних зазвичай передбачає етапи очищення, нормалізації та аугментації для підвищення їхньої корисності. Крім того, різноманітність джерел даних, таких як датчики IoT, веб-ресурси чи соціальні мережі, розширює можливості аналізу та прогнозування.
- Модель: Це структура, яку алгоритм використовує для навчання та прогнозування. Типи моделей, такі як лінійні регресії, дерева рішень, нейронні мережі, вибираються залежно від складності задачі та доступних даних. Модель має набір параметрів, які налаштовуються під час навчання для досягнення максимальної точності. Залежно від обсягу даних та потреб, використовують прості моделі (які швидше навчаються, але менш точні) або складні моделі (які вимагають більше ресурсів, але можуть враховувати більше деталей). Регуляризація та оптимізація є важливими етапами для запобігання перенавчанню та підвищення узагальнювальної здатності моделі.
- Алгоритм: Це математичний або обчислювальний метод, який визначає, як здійснюється навчання моделі. Алгоритми поділяються на типи, наприклад, супервізовані, несупервізовані, напівсупервізовані та алгоритми підкріплення. Вибір алгоритму залежить від типу задачі: наприклад, градієнтний спуск використовується для оптимізації параметрів, а методи кластеризації — для групування даних. Багато алгоритмів інтегрують концепцію ітераційного вдосконалення, що дозволяє поступово знижувати похибку. У сучасному машинному навчанні активно використовуються гібридні методи, які поєднують кілька алгоритмів для підвищення продуктивності та точності.

Завдяки цим основним компонентам машинне навчання стає потужним інструментом для вирішення складних завдань у різних сферах, зокрема в медицині, де воно дозволяє значно покращити процес діагностики та прогнозування захворювань.

Машинне навчання можна класифікувати за кількома основними типами, кожен з яких вирішує різні завдання на основі різних підходів до навчання.

1. Навчання з учителем (Supervised Learning)

Навчання з учителем є одним із найбільш широко використовуваних підходів у машинному навчанні. Цей тип навчання передбачає використання розмічених даних, тобто кожен приклад у навчальному наборі має відповідне вихідне значення або цільову змінну. Моделі, що навчаються з учителем, аналізують ці пари вхід-вихід, щоб навчитися передбачати або класифікувати нові дані. Приклади застосування включають класифікацію електронних листів як спам або ні, передбачення цін на ринку нерухомості або діагностику захворювань на основі медичних даних. Популярні алгоритми включають лінійну регресію, дерева рішень, метод опорних векторів (SVM) та нейронні мережі. Основний виклик полягає в необхідності великих обсягів якісних і розмічених даних для ефективного навчання моделі.

Завдання, де застосовується навчання з учителем, включають:

- Класифікацію: Завдання, де модель повинна класифікувати вхідні дані в одну з категорій. Наприклад, визначення, чи є пацієнт в групі ризику серцево-судинних захворювань на основі медичних показників.
- Регресію: Завдання, де модель передбачає числове значення. Наприклад, прогнозування ціни акцій на основі історичних даних.

2. Навчання без учителя (Unsupervised Learning)

У навчанні без учителя модель працює з нерозміченими даними, що означає відсутність явних цільових значень для навчання. Це дозволяє моделям виявляти приховані закономірності або структури у наборі даних без попереднього визначення результатів. Популярні алгоритми включають k-means, DBSCAN, і метод головних компонент (PCA). Головною складністю є визначення, наскільки

добре модель виявила закономірності, оскільки немає чітких “правильних” результатів.

Цей підхід особливо корисний у дослідницькому аналізі та для задач, таких як:

- Кластеризація: Виявлення груп подібних даних. Наприклад, сегментація споживачів за схожістю їх покупок.
- Асоціативні правила: Виявлення відносин між елементами в даних, наприклад, товарів, які часто купуються разом.

3. Навчання з підкріпленням (Reinforcement Learning)

Навчання з підкріпленням – це тип навчання, де агент (комп’ютерна програма) вчиться приймати рішення через взаємодію з оточенням. Агент виконує певні дії і отримує зворотний зв’язок у вигляді винагороди або штрафу. Це дозволяє йому покращувати свої стратегії на основі досвіду. Основними компонентами системи є агент, оточення, винагорода та політика дій. Застосування включає управління роботами, гри (наприклад, шахи або го) та оптимізацію бізнес-процесів. Основні виклики включають довгий час навчання, необхідність симуляції реального середовища та складність балансування між дослідженням нових стратегій і використанням наявних знань.

Задачі, які можуть бути вирішені за допомогою навчання з підкріпленням, включають:

- Навчання в іграх: Наприклад, алгоритми, які навчаються грати в шахи або го, отримуючи винагороди за виграші та штрафи за програші.
- Робототехніка: Агент може навчатися шляхом взаємодії з реальним середовищем, наприклад, робота, який навчається маніпулювати предметами.

Ці типи машинного навчання мають різні підходи до вирішення проблем і використовуються в залежності від завдання, доступних даних і вимог до результату. Також вказані підходи нерідко комбінуються, створюючи гібридні системи, які забезпечують більшу гнучкість і точність у виконанні завдань. Наприклад, у складних сценаріях, таких як автономне водіння, можуть використовуватися одночасно алгоритми класифікації, кластеризації та

підкріплення. Вибір підходу також залежить від кількості та якості даних, доступності обчислювальних ресурсів і складності задачі. В майбутньому розвиток машинного навчання може призвести до створення універсальних моделей, здатних вирішувати завдання будь-якого типу з мінімальним втручанням людини.

Дані є критично важливим компонентом у машинному навчанні, оскільки їх кількість, якість і репрезентативність прямо впливають на ефективність навченої моделі. Перш ніж моделі можна буде навчати, дані потребують підготовки, що включає етапи “чистки” та нормалізації. На цьому етапі здійснюється видалення помилкових, неповних або зашумлених даних, а також трансформація даних в єдиний формат для подальшої обробки. Якщо дані є некоректними, неповними або нерепрезентативними для поставленої задачі, це значно знижує точність прогнозів, а модель може виявитися неадекватною або навіть помилковою.

Процес підготовки даних включає кілька важливих етапів:

1. Очищення даних: Це включає виявлення та усунення помилок, таких як дублікати, відсутні значення або невірні записи. Для цього можуть використовуватися алгоритми заповнення пропусків, наприклад, метод середнього значення або моделювання пропущених даних. Очищення також передбачає видалення шуму, тобто нерелевантних або випадкових даних, які можуть спотворити результати аналізу. У деяких випадках дані можуть бути збагачені шляхом інтеграції з іншими джерелами, такими як відкриті бази даних чи API. Крім того, очищення допомагає вирішувати проблему некоректного форматування, наприклад, розбіжностей у записах дат чи числових значень.

2. Нормалізація даних: Цей етап передбачає масштабування числових значень до одного діапазону для уникнення домінування певних ознак під час навчання. Нормалізація може виконуватися за допомогою різних методів, таких як мінімакс-скейлінг або стандартизація. Вона також дозволяє забезпечити стабільність і швидкість алгоритмів оптимізації, які працюють краще з однорідними даними. Наприклад, у випадку зображень нормалізація може враховувати середнє значення пікселів у всьому наборі даних. Важливо також

враховувати, що неправильна нормалізація може вплинути на точність моделі, тому потрібно ретельно підбирати метод залежно від завдання.

3. Репрезентативність даних: Модель повинна бути навчена на даних, які адекватно представляють усі можливі варіанти ситуацій, які модель повинна буде обробляти в реальному світі. Це включає забезпечення балансу між різними класами, щоб уникнути упередженості моделі. Наприклад, якщо в наборі даних для класифікації є значний дисбаланс між класами, це може призвести до некоректних результатів. Крім того, репрезентативність може покращуватися за рахунок синтетичного збільшення даних, наприклад, за допомогою генерації нових прикладів через методи аугментації. Також важливо враховувати контекстні фактори, такі як час, місце чи умови, в яких збиралися дані, щоб модель була адаптована до реального середовища.

Неадекватні дані можуть призвести до неправильної інтерпретації закономірностей, що в результаті значно знижує ефективність машинного навчання та якість отриманих результатів. Тому коректне підготовлення даних є одним із ключових етапів у розробці надійних моделей машинного навчання.

2.3 Підходи до створення систем прогнозування ССЗ

Розробка системи прогнозування ССЗ базується на ряді принципів, які забезпечують її ефективність, точність та інтеграцію в медичну практику. Нижче наведено основні принципи:

1. Наукова обґрунтованість. Використання перевірених наукових даних та методик для створення моделей прогнозування. Підтримка доказової медицини: залучення великих клінічних досліджень і даних реальної практики. Адаптація системи до рекомендацій медичних організацій, таких як Європейське товариство кардіологів (ESC) або Американська асоціація серця (AHA).

2. Мультифакторний підхід. Аналіз великої кількості факторів ризику, а саме: вік, стать, спадковість; рівень холестерину, глюкози, артеріальний тиск; спосіб життя (куріння, фізична активність, стрес); дані генетичного тестування;

інтеграція структурованих (аналізи, обстеження) та неструктурованих даних (записи лікарів, зображення).

3. Використання сучасних технологій. Машинне навчання (ML) для точного прогнозування на основі складних взаємозв'язків між факторами ризику. Big Data для роботи з великими обсягами даних. Інтернет речей (IoT) для постійного моніторингу пацієнтів через носимі пристрої.

4. Персоналізація. Індивідуальний підхід до кожного пацієнта. Адаптація прогнозів та рекомендацій залежно від специфічних параметрів і стану пацієнта.

5. Раннє виявлення та попередження. Розробка алгоритмів для визначення ризику ще до появи симптомів. Інформування пацієнтів про ризики та рекомендації для їх зниження.

6. Інтерпретованість. Забезпечення прозорості та зрозумілості рішень системи для лікарів і пацієнтів. Використання методів інтерпретації моделей, таких як LIME або SHAP, для пояснення прогнозів.

7. Інтеграція в медичну екосистему. Взаємодія з електронними медичними записами (EMR) для отримання та зберігання даних. Сумісність із телемедичними платформами та мобільними додатками.

8. Модульність і масштабованість. Побудова системи як набору модулів, які можна вдосконалювати без шкоди для інших компонентів. Можливість розширення для обробки даних більшої кількості пацієнтів чи нових факторів ризику.

9. Забезпечення конфіденційності та безпеки даних. Дотримання регуляторних норм (HIPAA, GDPR) для захисту персональної інформації. Використання сучасних протоколів кібербезпеки для зберігання та обробки медичних даних.

10. Ефективність і економічна доцільність. Зменшення витрат на лікування завдяки ранньому виявленню ризиків. Забезпечення швидкої та ефективної обробки великих обсягів даних.

11. Постійне навчання системи. Регулярне оновлення моделей машинного навчання на основі нових даних. Врахування зворотного зв’язку від лікарів та пацієнтів.

12. Універсальність та доступність. Можливість використання системи в різних медичних установах, включаючи віддалені регіони. Забезпечення інтерфейсу для лікарів і пацієнтів з різними рівнями технічних навичок.

Таким чином реалізація цих принципів допоможе створити систему, яка буде точно прогнозувати ризики ССЗ, забезпечувати високий рівень медичної допомоги та сприяти зниженню захворюваності та смертності від серцево-судинних захворювань.

2.4 Оцінка безпеки системи прогнозування захворювань на основі технологій Big Data та машинного навчання

Системи прогнозування захворювань на основі технології Big Data дозволяють аналізувати великі обсяги медичних і поведінкових даних для виявлення тенденцій у розвитку захворювань, визначення факторів ризику та своєчасного виявлення можливих епідемій або спалахів інфекційних захворювань, наприклад, BlueDot. Технологія Big Data в прогнозуванні захворювань сприяє ефективнішому моніторингу стану здоров’я населення та допомагає у своєчасному вжитті організації заходів для запобігання серйозних наслідків для системи охорони здоров’я [19]. Але системи прогнозування захворювань на основі Big Data стикаються з рядом проблем кібербезпеки, оскільки вони оперують великим обсягом персональних та конфіденційних даних, які є цікавими для зловмисника. Тому виникає завдання метою якого є визначення актуальних проблем безпеки програмних засобів інтелектуальних систем прогнозування захворювань на основі технології Big Data та знаходження раціональних шляхів їх усунення.

Системи прогнозування захворювань на основі технології Big Data створюються для ефективного аналізу великих обсягів медичних, соціальних та

екологічних даних, що дозволяє краще розуміти, передбачати та запобігати різним захворюванням. Загалом, основна мета систем прогнозування захворювань на основі Big Data – це забезпечення більш ефективної системи охорони здоров'я, яка може реагувати на виклики вчасно, попереджувати захворювання, скорочувати витрати та сприяти персоналізованій і доказовій медицині [4].

Оцінка безпеки системи прогнозування захворювань на основі технологій Big Data та машинного навчання є важливою темою. Ці системи можуть допомогти у ранньому виявленні захворювань, але вони також стикаються з викликами, пов'язаними з конфіденційністю даних та точністю прогнозів.

Основні аспекти оцінки безпеки включають:

- Конфіденційність даних: Забезпечення захисту персональних даних пацієнтів.
- Точність прогнозів: Визначення рівня точності та надійності прогнозів системи.
- Відповідність законодавству: Дотримання всіх необхідних правил та нормативів у сфері охорони здоров’я та даних.
- Етика: Врахування етичних аспектів використання таких технологій.

Головним аспектом безпеки систем прогнозування захворювань на основі технологій Big Data та машинного навчання є забезпечення захисту та конфіденційності даних. Ось ключові моменти, на які слід звернути увагу:

Захист даних: Використання методів шифрування та анонімізації для збереження конфіденційності медичних даних пацієнтів. Це допомагає запобігти несанкціонованому доступу та витоку інформації.

Точність моделей: Ретельне тестування та валідація моделей машинного навчання для забезпечення їх надійності та точності прогнозів. Це включає перевірку на відповідних наборах даних та регулярне оновлення моделей.

Відповідність законодавству: Дотримання міжнародних та місцевих нормативних актів щодо захисту даних, таких як GDPR у Європі або HIPAA у США.

Етичні міркування: Забезпечення прозорості в тому, як дані використовуються та які рішення приймаються на їх основі. Важливо також враховувати можливі упередження та дискримінацію, які можуть виникнути при використанні машинного навчання.

Безпека системи: Захист серверів та інфраструктури, на яких зберігаються та обробляються дані, від кібератак та інших загроз.

Всі ці аспекти разом сприяють створенню надійної та безпечної системи прогнозування захворювань, яка може приносити користь без порушення прав та приватності пацієнтів.

2.5 Оцінка продуктивності системи прогнозування захворювань на основі технологій Big Data

Продуктивність системи прогнозування захворювань на основі технологій Big Data залежить від багатьох факторів, які визначають якість, швидкість, точність та ефективність системи. Розглянемо основні аспекти:

1. Якість даних:

- Достовірність даних: Використання надійних джерел, таких як електронні медичні записи, генетичні дослідження, носимі пристрої. Мінімізація помилок у зібраних даних.
- Повнота даних:Наявність усіх необхідних параметрів для аналізу (вікові, демографічні, клінічні дані тощо).
- Структура даних: Коректна організація даних (структуровані/неструктуровані).
- Обробка пропущених даних: Використання методів для заповнення прогалин (імпутація).

2. Потужність інфраструктури:

- Обчислювальні ресурси: Швидкодія процесорів, графічних прискорювачів (GPU), кластерів.

- Хмарні технології: Використання AWS, Azure, Google Cloud для масштабованості й обробки великих обсягів даних.
- Мережеві ресурси: Висока пропускна здатність для передачі даних.

3. Ефективність алгоритмів:

- Вибір моделей: Продуктивність залежить від правильного вибору алгоритмів машинного навчання (наприклад, Random Forest, XGBoost, нейронні мережі).
- Оптимізація моделей: Налаштування гіперпараметрів (learning rate, depth trees тощо) для підвищення точності прогнозування.
- Інтеграція новітніх технологій: Використання інноваційних алгоритмів і підходів, таких як глибоке навчання (Deep Learning).

4. Обробка та аналіз великих даних:

- Big Data інструменти: Використання Apache Spark, Hadoop для паралельної обробки великих обсягів даних.
- Реалізація ETL-процесів: Ефективне перетворення, очищення та завантаження даних.
- Масштабованість системи: Здатність обробляти дедалі більші обсяги даних без зниження продуктивності.

5. Актуальність та оновлення даних:

- Реальний час: Продуктивність системи значно покращується, якщо вона здатна працювати з потоками даних у реальному часі (real-time analytics).
- Оновлення моделей: Регулярне навчання моделей на нових даних для забезпечення точності прогнозів.

6. Інтерпретація та зручність використання:

- Зрозумілість прогнозів: Надання користувачам інтерпретованих результатів (наприклад, ризик у відсотках).
- Візуалізація: Ефективні інструменти для представлення результатів (Tableau, Power BI).
- Інтерфейси: Зручний та швидкий доступ до системи через API, мобільні додатки чи веб-платформи.

7. Безпека та конфіденційність:

- Захист даних: Використання шифрування для зберігання та передачі даних.
- Дотримання регуляторних стандартів: Відповідність вимогам GDPR, HIPAA для забезпечення конфіденційності.

8. Людський фактор:

- Експертна оцінка: Співпраця з лікарями для коригування моделей прогнозування.
- Навчання користувачів: Здатність медичного персоналу ефективно використовувати систему.

9. Модульність і масштабованість:

- Модульний дизайн: Гнучкість у додаванні нових функцій без шкоди для продуктивності.
- Готовність до зростання навантаження: Можливість розширення інфраструктури без значного перепроектування.

10. Метрики ефективності:

- Точність прогнозів: Високий рівень показників AUC-ROC, Precision, Recall.
- Швидкість роботи: Час від запиту до отримання результату.
- Надійність системи: Здатність функціонувати без збоїв навіть за високих навантажень.

Таким чином усі ці фактори разом визначають, наскільки ефективною і корисною буде система для прогнозування захворювань у реальній медичній практиці.

2.6 Визначення метрик ефективності продуктивності системи прогнозування захворювань на основі технологій Big Data

Для оцінки ефективності та продуктивності системи прогнозування захворювань, побудованої на основі технологій Big Data, використовуються

кількісні та якісні метрики. Вони забезпечують об’єктивне вимірювання точності прогнозів, швидкості обробки даних, а також надійності системи.

1. Метрики точності прогнозування. Ці метрики визначають, наскільки добре система передбачає захворювання [12, 14, 16].

- Accuracy (Точність): Частка правильних прогнозів серед усіх.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

Де:

- TP (True Positive) – правильні передбачення захворювання.
- TN (True Negative) – правильні передбачення відсутності захворювання.
- FP (False Positive) – хибні спрацьовування.
- FN (False Negative) – пропущені випадки захворювання.

- Precision (Прецизійність): Частка правильних прогнозів захворювання серед усіх прогнозів, які система визначила як позитивні.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

- Recall (Чутливість або Повнота): Частка виявлених випадків захворювання серед усіх реальних випадків.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

- F1-Score: Гармонійне середнє між Precision і Recall.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

- ROC-AUC (Area Under Curve): Площа під кривою робочих характеристик, що відображає співвідношення чутливості та специфічності.

2. Метрики продуктивності системи. Ці метрики визначають швидкість і ефективність роботи системи з великими даними.

- Throughput (Пропускна здатність):Кількість оброблених записів або прогнозів за одиницю часу.
- Latency (Затримка): Час від надходження даних до видачі результату прогнозу.
- Response Time (Час відповіді): Загальний час обробки запиту та отримання відповіді.
- Scalability (Масштабованість): Здатність системи ефективно обробляти збільшений обсяг даних без втрати продуктивності.
- Resource Utilization (Використання ресурсів): Відсоток використання обчислювальних ресурсів (ЦП, пам’ять, мережа).

3. Метрики якості даних. Ці метрики оцінюють якість даних, які використовуються для навчання системи.

- Completeness (Повнота): Відсоток наявних даних порівняно з очікуваними, що визначає рівень їх повного охоплення.
- Accuracy (Точність даних): Частка даних без помилок, що відображає їхню відповідність реальним значенням.
- Timeliness (Актуальність): Час, протягом якого дані оновлюються та залишаються корисними для аналізу.
- Consistency (Послідовність): Відсутність суперечливих даних у наборі, що забезпечує узгодженість між елементами.

4. Метрики надійності. Ці метрики визначають стабільність і безпеку системи.

- Availability (Доступність): Відсоток часу, протягом якого система працює коректно.
- Fault Tolerance (Стійкість до збоїв): Здатність системи функціонувати в умовах відмови частини компонентів.
- Recovery Time (Час відновлення): Час, потрібний системі для відновлення після збою.

5. Метрики користувацького досвіду. Оцінюють зручність використання системи лікарями та пацієнтами.

- Satisfaction Score (Оцінка задоволеності): Опитування користувачів щодо зручності роботи із системою.
- Adoption Rate (Рівень прийняття): Частка медичних установ або користувачів, які впровадили систему.
- Error Rate (Рівень помилок користувача): Кількість помилок через некоректну взаємодію з інтерфейсом.

6. Економічні метрики. Оцінюють фінансову ефективність впровадження системи.

- Cost Savings (Економія коштів): Зниження витрат на лікування завдяки ранньому виявленню захворювань.
- Return on Investment (ROI): Відношення отриманого доходу до витрат на розробку та впровадження системи.

$$ROI = \frac{\text{Повернення} - \text{Витрати}}{\text{Витрати}} \times 100\% \tag{2.5}$$

Таким чином використання цих метрик дозволяє комплексно оцінити ефективність і продуктивність системи прогнозування захворювань. Важливо враховувати всі аспекти, щоб забезпечити її точність, швидкість, надійність і прийнятність для кінцевих користувачів.

2.7 Визначення метрик ефективності продуктивності системи прогнозування захворювань на основі технологій машинного навчання

Для оцінки ефективності систем прогнозування захворювань, побудованих на основі машинного навчання (ML), використовуються різноманітні метрики. Вони дозволяють оцінити точність, швидкість, надійність, а також практичну корисність системи.

1. Метрики точності прогнозування. Ці метрики визначають, наскільки система правильно передбачає захворювання.

- Accuracy (Точність): Загальна частка правильних прогнозів.
- Precision (Прецизійність): Частка правильних прогнозів захворювання серед усіх прогнозів, які система визначила як позитивні.
- Recall (Чутливість або Повнота): Частка виявлених випадків захворювання серед усіх реальних випадків.
- F1-Score: Середнє значення між Precision та Recall, корисне при дисбалансі класів.
- Specificity (Специфічність): Частка правильно визначених негативних випадків.
- ROC-AUC (Area Under ROC Curve): Площа під кривою, яка показує баланс між чутливістю та специфічністю. Значення ближче до 1 вказує на кращу продуктивність.

2. Метрики продуктивності системи. Ці метрики оцінюють швидкість і ефективність роботи системи.

- Latency (Затримка):Час, потрібний для отримання результату прогнозу після подачі даних.
- Throughput (Пропускна здатність): Кількість оброблених запитів або прогнозів за одиницю часу.
- Inference Time (Час прогнозування): Час, необхідний моделі ML для обробки одного набору вхідних даних.

- Scalability (Масштабованість): Здатність системи ефективно обробляти зростаючий обсяг даних або кількість користувачів.

3. Метрики узгодженості даних

- Data Quality (Якість даних): Чіткість, актуальність, повнота та точність даних, використаних для навчання.
- Class Balance (Баланс класів): Важливо, щоб модель враховувала можливий дисбаланс між позитивними та негативними класами.
- Outlier Detection (Виявлення викидів): Аналіз впливу нетипових даних на прогнозування.

4. Метрики надійності

- Robustness (Стійкість): Можливість системи правильно працювати навіть у разі шуму в даних або відхилень у входах.
- Fault Tolerance (Толерантність до збоїв): Здатність системи продовжувати роботу у випадку часткових збоїв.
- Recovery Time (Час відновлення): Час, необхідний для відновлення після збою.

5. Метрики економічної ефективності

- Cost Efficiency (Ефективність витрат): Аналіз співвідношення вартості реалізації системи до її користі (зменшення витрат на лікування, зниження захворюваності).
- Return on Investment (ROI): Оцінка прибутковості системи:

$$ROI = \frac{\text{Збережені витрати або прибуток} - \text{Витрати на систему}}{\text{Витрати на систему}} \times 100\% \tag{2.6}$$

6. Метрики користувацького досвіду

- User Satisfaction (Задоволеність користувачів): Результати опитувань лікарів і пацієнтів щодо зручності та корисності системи.
- Adoption Rate (Рівень впровадження): Частка медичних установ або лікарів, які активно використовують систему.

7. Специфічні метрики для машинного навчання

- Confusion Matrix (Матриця помилок): Таблиця, що показує кількість TP, FP, TN, FN і допомагає зрозуміти продуктивність моделі.
- Cross-validation Accuracy: Середня точність моделі за результатами крос-валідації.
- Log Loss (Логарифмічна втрата): Метрична оцінка ймовірнісних прогнозів, де менше значення вказує на більш точні передбачення.

8. Інтегровані метрики

- Cumulative Gain (Кумулятивний приріст): Відображає, наскільки добре модель сегментує пацієнтів із високим ризиком.
- Lift Chart: Показує ефективність моделі в порівнянні з випадковим прогнозом.

Таким чином використання зазначених метрик дозволяє отримати всебічну оцінку системи прогнозування захворювань на основі технологій машинного навчання. Це допомагає визначити сильні та слабкі сторони моделі, забезпечуючи можливість її подальшого вдосконалення.

3 РОЗРОБКА ПРОПОЗИЦІЇ ЩОДО ПОБУДОВИ СИСТЕМИ ПРОГНОЗУВАННЯ ССЗ НА ОСНОВІ ТЕХНОЛОГІЇ BIG DATA

3.1 Огляд набору даних та його стандартизація

У цій секції ми надамо детальний огляд набору даних (невеликий масив наведено в Додатку А), який використовується в нашому дослідженні. Набір даних містить різноманітні ознаки, пов’язані з серцево-судинним здоров’ям. Ознаки, включені в набір даних, наступні.

1. Деперсоналізований ідентифікатор пацієнта (*iPatient_ID_fk*)
2. Етнічність (*sPatient_Ethnicity*). Етнічність може впливати на серцево-судинні захворювання через генетичні фактори, що визначають схильність до таких станів, як гіпертензія або діабет, які є ключовими ризиками. Соціально-економічні умови, характерні для певних етнічних груп, можуть впливати на доступ до медичної допомоги, здорового харчування та можливостей для фізичної активності. Крім того, культурні особливості, такі як традиційна дієта або рівень стресу, також можуть сприяти розвитку серцево-судинних захворювань.
3. Стать (*sPatient_Gender_Cde*). У чоловіків ризик розвитку ССЗ, таких як інфаркт міокарда, зазвичай вищий у молодому і середньому віці, тоді як у жінок він зростає після менопаузи через зниження рівня естрогенів, які мають захисний ефект. Також існують відмінності у симптомах: наприклад, жінки частіше відчувають нетипові прояви серцевого нападу, що може ускладнити діагностику та своєчасне лікування [8].
4. Ліпопротеїн високої щільності (*HDL*). Часто званий «хорошим холестерином», має важливу роль у зниженні ризику серцево-судинних захворювань, оскільки він допомагає виводити надлишок холестерину з артерій, знижуючи ймовірність розвитку атеросклерозу. Високий рівень є показником здоров’я серцево-судинної системи і може зменшити ризик серцевих нападів і інсультів.

5. Ліпопротеїн низької щільності (*LDL*). Ця ознака вказує рівень холестерину низької щільності в крові. Підвищені рівні холестерину низької щільності пов’язані зі збільшеним ризиком хвороби коронарних артерій.

6. Місце надання послуг (*p_sPx_Place_Of_Service*). Амбулаторно, під час огляду в лікаря, в аптеці, вдома.

7. Назва медичного рецепту (*sRx_Nme*). Стандартизована номенклатура медичних рецептів, підтримувана Національною бібліотекою медицини (США).

8. Приймаючі антигіпертензивні препарати (*Antihypertensive_Medication*). Ліки, що використовуються для зниження артеріального тиску у пацієнтів з гіпертонією. Вони включають кілька класів медикаментів, які допомагають контролювати тиск і зменшувати ризик серцево-судинних захворювань.

9. К-сть зафіксованих інфарктів міокарда (*Total_MI_Events*).

10. К-сть зафіксованих ішемічних нападів (*Total_Stroke_Events*).

11. Наявність нефротичного синдрому (*NEPHROTIC_SYNDROME*). Клінічний комплекс симптомів, що включає значну втрату білка через нирки (протеїнурія), набряки, гіпоальбумінемію та підвищений рівень холестерину в крові. Цей синдром є результатом пошкодження ниркових клубочків, що веде до порушення фільтраційної функції нирок.

12. Наявність хвороби Крона (*CROHNS_DISEASE*). Хронічне запальне захворювання шлунково-кишкового тракту, яке може вражати будь-яку частину від рота до ануса, часто у вигляді виразок, вузлів і рубців. Симптоми включають біль у животі, діарею, втрату ваги та загальну слабкість, а лікування зазвичай передбачає використання медикаментів, таких як імуносупресанти, або хірургічне втручання при ускладненнях.

13. Наявність неспецифічного виразкового коліта (*ULCERATIVE_COLITIS*). Хронічне запальне захворювання, яке вражає товсту кишку, викликаючи запалення, виразки та ерозії слизової оболонки. Основними симптомами є біль у животі, діарея, наявність крові в стільці та загальна

слабкість, а лікування включає медикаментозну терапію (антибіотики, протизапальні препарати) та, за потреби, хірургічне втручання.

14. Наявність псоріатичного артиту (*PSORIATIC_ARTHRITIS*). Хронічне запальне захворювання, яке поєднує ознаки псоріазу та артриту. Воно викликає запалення в суглобах, що може супроводжуватися болем, набряками та обмеженням рухливості. Крім того, у пацієнтів можуть з’являтися псоріатичні висипи на шкірі. Лікування зазвичай включає нестероїдні протизапальні препарати, біологічні терапії та інші імуносупресивні ліки для контролю симптомів і запобігання ускладненням.

15. Наявність стенозу аортального клапана (*AORTIC_VALVE_STENOSIS*). Захворювання, при якому звуження аортального клапана обмежує потік крові з лівого шлуночка в аорту, що може призвести до підвищеного навантаження на серце. Це може викликати симптоми, такі як задишка, біль у грудях та запаморочення, і вимагає лікування, яке часто включає хірургічне втручання, зокрема заміну клапана.

16. Наявність хронічної хвороби нирок (*CHRONIC_KIDNEY_DISEASE*). Поступове і незворотне погіршення функції нирок, яке може призвести до ниркової недостатності. Причини включають цукровий діабет, артеріальну гіпертензію та інші захворювання, що пошкоджують ниркові функції, і часто потребує лікування, яке може включати діаліз або трансплантацію нирки на пізніх стадіях.

17. Наявність гіперліпідемії (*HYPERLIPIDEMIA*). Стан, при якому рівень ліпідів, таких як холестерин і тригліцериди, в крові підвищений, що збільшує ризик розвитку серцево-судинних захворювань. Цей стан може бути викликаний генетичними факторами, нездоровим способом життя, такими як неправильне харчування або недостатня фізична активність, і часто потребує медикаментозного лікування для нормалізації рівня ліпідів і зниження ризиків для здоров'я.

18. Наявність гіпертензії (*HYPERTENSION*). Високий артеріальний тиск, є хронічним захворюванням, при якому тиск крові на стінки артерій постійно підвищений, що збільшує ризик розвитку серцево-судинних захворювань,

інсультів і ниркової недостатності. Основними факторами ризику є нездоровий спосіб життя, генетична схильність та інші захворювання, як цукровий діабет, що потребують контролю через медикаментозну терапію та зміни в стилі життя.

19. Наявність менопаузи (*MENOPAUSE*). Через зниження рівня естрогену, який має захисний ефект для серцево-судинної системи. Після менопаузи жінки мають підвищений ризик гіпертензії, атеросклерозу та інсультів, що частково пояснюється змінами в метаболізмі ліпідів і судинному тонусі.

20. Наявність ревматоїдного артиту (*RHEUMATOID_ARTHRITIS*). Хронічне аутоімунне захворювання, яке викликає запалення суглобів, болі, набряки та поступову втрату рухливості. Цей стан також може призводити до системних уражень, таких як пошкодження органів, і потребує комплексного лікування, що включає протизапальні препарати та імуносупресивну терапію.

21. Наявність гіпертригліцеридемії (*HYPERTRIGLYCERIDEMIA*). Стан, при якому рівень тригліцеридів у крові підвищений, що є фактором ризику для розвитку серцево-судинних захворювань, таких як атеросклероз і інсульти. Вона може бути спричинена такими факторами, як ожиріння, малорухливий спосіб життя, неправильне харчування, а також деякими хронічними захворюваннями, як цукровий діабет.

22. Вік (*labAge*). Ця ознака представляє вік людей (у роках). Вік є важливим фактором для оцінки ризику серцево-судинних захворювань, оскільки їхня поширеність зростає з віком.

23. Приймаючі препарати ліпідо-знижуючої терапії (*LLT_Medication*). Препарати ліпідо-знижуючої терапії використовуються для зниження рівнів холестерину та тригліцеридів у крові, що допомагає зменшити ризик розвитку серцево-судинних захворювань. Основними класами цих препаратів є статини, які знижують рівень LDL, а також фібрати та інгібітори PCSK9, які зменшують рівень тригліцеридів і покращують ліпідний профіль.

Набір даних було зібрано з Медико-дослідницького центру Family Heart Foundation (США), і він складається з 142674 спостережень. Дані попередньо довелося очищувати, оскільки вибірка містила відсутні значення чи дублікати.

Перед тим, як описувати дослідницьку реалізації алгоритмів, опишемо метрики, які будуть використані для оцінки якості навчених моделей. Оскільки наша цільова змінна є булевою (відповідь Так/Ні), існує всього 4 можливих варіанти:

1) True Negative (TN), вірно негативний (реальне значення негативне, передбачене значення негативне);

2) False Negative (FN), хибно негативний (реальне значення позитивне, передбачене значення негативне);

3) False Positive (FP), хибно позитивний (реальне значення негативне, передбачене значення позитивне);

4) True Positive (TP), вірно позитивний (реальне значення позитивне, передбачене значення позитивне).

Також використаємо метрику AUC (Area Under the Curve – площа під кривою). Це статистичний показник, який часто використовується для оцінки якості класифікаційних моделей, особливо в задачах бінарної класифікації. Він представляє площу під кривою ROC (Receiver Operating Characteristic), яка відображає співвідношення між чутливістю (True Positive Rate) та специфічністю (1 - False Positive Rate) при різних порогах моделі. Значення AUC коливається між 0 і 1: показник ближче до 1 вказує на високу точність моделі, тоді як значення ближче до 0.5 вказує на випадкове прогнозування. У дослідженнях AUC використовується для порівняння ефективності різних моделей та вибору найкращої для прогнозування, оскільки він враховує баланс між помилками першого та другого роду. AUC також є важливим інструментом для визначення межової чутливості моделі, що дозволяє покращувати її практичну застосовність у реальних сценаріях, особливо в контексті кривої характеристики роботи класифікатора (ROC-кривої).

ROC-крива – це графік який відображає ефективність класифікаційної моделі, зіставляючи показники чутливості (True Positive Rate) та специфічності (1 - False Positive Rate) для різних порогів класифікації. Вона дозволяє візуально оцінити, наскільки модель здатна правильно розрізняти позитивні та негативні

класи. Чим ближче крива до верхнього лівого кута, тим кращою є модель, оскільки це вказує на високу чутливість та специфічність. ROC-крива використовується для порівняння кількох моделей, дозволяючи обрати ту, яка має кращу дискримінаційну здатність. Крім того, аналіз ROC-кривої допомагає визначити оптимальний поріг класифікації, який забезпечить баланс між правильною ідентифікацією позитивних випадків і мінімізацією помилкових спрацьовувань [17].

3.2 Навчання моделі та аналіз отриманих результатів

Для навчання моделі було використано метод градієнтного бустингу на основі дерев рішень (LightGBM). Мета LightGBM – зменшити час тренування моделей, забезпечуючи високу точність та можливість роботи з великими наборами даних. Однією з ключових особливостей LightGBM є використання Leaf-wise growth (ріст за листками) замість Level-wise (рівномірного росту дерев), що забезпечує кращу мінімізацію втрат [22].

Level-wise growth (ріст за рівнями) має свої особливості. У цьому підході дерево будується рівномірно, рівень за рівнем. На кожному кроці всі вузли поточного рівня розбиваються на два дочірніх вузли (або більше, залежно від алгоритму). Цей метод забезпечує баланс між усіма шляхами від кореня до листків, тобто дерево має однакову глибину для всіх гілок.

Основними перевагами є:

- Легше контролювати розмір дерева (балансованість).
- Зменшує ризик перенавчання, особливо при невеликій кількості даних.
- Часто використовується у традиційних алгоритмах, таких як XGBoost.

Основний недолік визначено те, що ріст дерева відбувається менш гнучко, оскільки багато вузлів можуть створюватися в місцях, де вони мало сприяють зменшенню втрат.

Leaf-wise growth (ріст за листками) має певні відмінності. У цьому підході на кожній ітерації розбивається лише один вузол – той, який має найбільший потенціал для зменшення втрат. Це дозволяє дереву рости асиметрично, концентруючи ресурси на найбільш значущих розбиттях. Як результат, дерево може бути глибшим у одних гілках і неглибоким у інших.

- Основні переваги:
  - Ефективніше мінімізує втрати, оскільки обчислення зосереджуються на найбільш корисних розбиттях.
  - Може досягати кращої продуктивності (точності) на великих наборах даних.
- Основні недоліки:
  - Може призводити до перенавчання, особливо на невеликих даних.
  - Дерева можуть бути сильно незбалансованими, що іноді впливає на інтерпретованість моделі.

Тобто, основна різниця між методами полягає в тому, що при Level-wise дерево зростає рівномірно, розбиваючи всі вузли одного рівня. І тому основний акцент – на контрольованій структурі дерева. Натомість при Leaf-wise дерево зростає асиметрично, вибираючи найбільш корисний вузол для розбиття. Відповідно, основний акцент на мінімізації втрат.

Leaf-wise growth, який використовується в LightGBM, є більш ефективним для великих даних, оскільки він краще адаптується до складності даних і досягнення точності, хоча потребує додаткового налаштування для уникнення перенавчання [15]. Цей підхід дозволяє моделі концентруватися на більш значущих вузлах, що сприяє підвищенню продуктивності в умовах неоднорідного розподілу даних. Водночас, асиметрична структура дерева може призводити до значного збільшення його глибини, що може впливати на час виконання запитів. Вибір між Level-wise та Leaf-wise методами залежить від доступних ресурсів і завдань моделі, де Leaf-wise зазвичай підходить для більш складних сценаріїв. Оптимізація параметрів Leaf-wise, таких як максимальна глибина чи мінімальний обсяг вузла, є важливим етапом для забезпечення стабільної роботи моделі.

LightGBM також використовує Histogram-based splitting для зменшення обчислювальної складності розбиттів вузлів дерева. Це техніка, яка використовується для швидкого та ефективного обчислення розбиття даних у процесі створення дерева. Замість перевірки кожного значення в наборі даних, LightGBM групує числові значення у бінові сегменти (групи) і обчислює кращі точки розбиття на основі цих груп. Це дозволяє значно зменшити обчислювальні витрати і прискорити тренування моделі, особливо на великих наборах даних. Такий підхід також знижує вимоги до пам'яті, оскільки замість збереження всіх значень потрібно лише зберігати агреговані гістограми. Histogram-based splitting підтримує баланс між точністю моделі та ефективністю обчислень, що робить його ідеальним для роботи з великими обсягами високовимірних даних.

Бібліотека підтримує Exclusive Feature Bundling **(**EFB**)**, який об'єднує рідкісні ознаки для зменшення розмірності. Це оптимізаційний метод, який зменшує розмірності вхідних даних без втрати інформації. Метод групує взаємовиключні ознаки (exclusive features), тобто ті, які рідко або ніколи не приймають ненульові значення одночасно, в один стовпець. Це дозволяє зменшити кількість ознак і знизити обчислювальні витрати під час навчання моделі. За допомогою EFB модель може працювати з великими наборами даних з високою кількістю ознак без значного збільшення обсягу пам’яті. Такий підхід забезпечує високу ефективність LightGBM у задачах з великою розмірністю даних [22].

Отож, метод LightGBM оптимізовано для роботи з великими наборами даних і здатний працювати у розподіленому середовищі. Впроваджено методи, які забезпечують паралелізм і масштабованість алгоритму. Що найкраще підходить для предмету дослідження.

Експерименти демонструють, що LightGBM тренується швидше, займає менше пам’яті та часто забезпечує вищу точність. LightGBM також має вбудовані механізми для уникнення перенавчання, наприклад, обмеження глибини дерев. Впроваджено функціонал для автоматичного вибору оптимальних параметрів моделі.

Особлива увага приділяється підтримці обробки дисбалансованих даних завдяки параметрам для ваги класів. Бібліотека включає велику кількість гіперпараметрів, що дає змогу адаптувати модель до різних задач. LightGBM є відкритим програмним забезпеченням, що робить його доступним для широкої спільноти. Робота LightGBM доводить важливість інновацій у структурі дерев та алгоритмах розбиття для покращення продуктивності.

Результати тестування показують продуктивність моделі LightGBM порівняно з градієнтним бустингом (GB), який також розглядався у вигляді основної моделі, у завданні прогнозування показника Lp(a) 125 nmol/L. Нижче наведено огляд основних метрик та інтерпретацію результатів:

1. P1 Scores

- LGBM mean P1: 47.53% (95% CI: 46.22% - 48.84%)
- GB mean P1: 47.68% (95% CI: 46.57% - 48.79%)

Результати P1 показують дуже схожі середні значення для LightGBM і GB, з невеликою перевагою у GB. Різниця знаходиться в межах статистичної похибки, що вказує на порівнянну ефективність обох моделей у цьому аспекті. Також це свідчить про стабільність обох алгоритмів у прогнозуванні базового рівня. У реальних умовах обидві моделі можуть бути використані як взаємозамінні залежно від додаткових обмежень або ресурсів.

2. P2 Scores

- LGBM mean P2: 45.67% (95% CI: 39.83% - 41.51%)
- GB mean P2: 41.18% (95% CI: 40.28% - 42.08%)

LGBM трохи випереджає від GB у прогнозуванні за P2, але різниця не є суттєвою. Показник P2 для LGBM демонструє тенденцію до кращого визначення ключових особливостей у даних. Однак, широкий довірчий інтервал може свідчити про потенціал варіативності залежно від набору даних. Оптимізація гіперпараметрів може ще більше збільшити точність моделі в цьому аспекті.

3. R1 Scores

- LGBM mean R1: 3.59% (95% CI: 0.81% - 2.37%)
- GB mean R1: 2.04% (95% CI: 1.35% - 2.73%)

LGBM показує вищий середній показник R1 порівняно з GB, що може свідчити про більшу стабільність або точність у специфічних задачах, пов’язаних із цією метрикою. Вищий показник R1 для LGBM свідчить про його здатність краще вловлювати окремі важливі закономірності. Проте, інколи ширший довірчий інтервал LGBM може вказувати на потребу в додатковій валідації моделі. Висока стабільність при різних експериментах робить LGBM перспективною технологією для довгострокових сценаріїв.

4. ROC AUC Scores

- LGBM mean ROC AUC: 63.76% (95% CI: 61.55% - 61.97%)
- GB mean ROC AUC: 62.05% (95% CI: 61.77% - 62.33%)

LGBM демонструє трохи кращу площу під кривою ROC (AUC), що вказує на перевагу в загальній класифікаційній здатності. Результати AUC для LGBM показують, що модель більш збалансована між точністю та повнотою. Це робить її більш надійною для сценаріїв, де обидві метрики мають вирішальне значення. Невелика перевага може бути суттєвою в завданнях із критичною важливістю передбачень. Відповідно до описаних характеристик і був використий спосіб LightGBM для проведення подальшого дослідження.

Тепер розглянемо описові характеристики та ефективність навченої моделі за допомогою LightGBM для гіперпараметру значеннь (Lp(a)) = 125.

Feats shape: (142674, 65301). Ця ознака свідчить про розмірність вхідного набору даних (матриці ознак), де 142674 – це кількість рядків, що відповідає числу прикладів або записів у наборі даних. У нашому випадку це к-сть пацієнтів у вихідному наборі даних. 65301 – це кількість унікальних значень стовпців, які представляють ознаки (features) для кожного запису. Ознаки можуть включати числові, категоріальні або бінарні дані, що використовуються для навчання моделі.

Labels shape: (142674, 1). Вказує на розмірність міток (labels) або цільової змінної. 142674 – кількість міток відповідає числу прикладів у наборі даних, що гарантує узгодженість між ознаками та мітками. 1 – це кількість цільових змінних,

що вказує на те, що задача, є задачею з однією вихідною змінною. Для нашого дослідження ми визначали чи є у пацієнтів схильність до підвищеного рівня Lp(a).

Діаграма на рис 3.1 показує два графіки: ROC-криву та графік метрик продуктивності при зміні порогового значення класифікації. ROC-крива порівнює чутливість (справжні позитиви) та специфічність (хибні позитиви) для моделі “No Skill” і моделі “Classifier”, при цьому ROC-крива моделі класифікатора знаходиться вище кривої без навичок, що свідчить про кращу прогностичну ефективність.

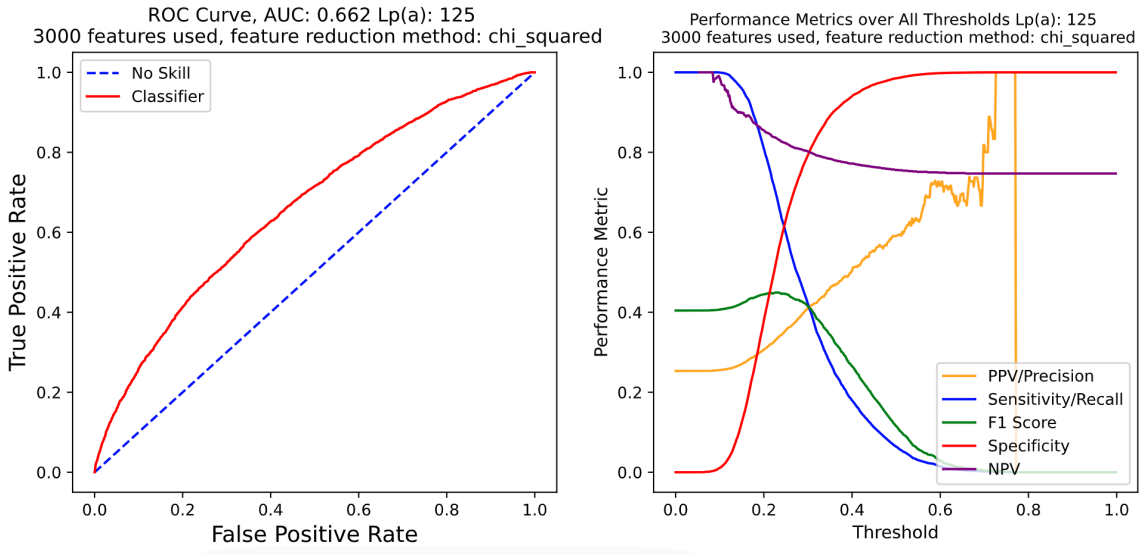


Рис 3.1 Крива ROC та метрики ефективності

Площа під ROC-кривою (AUC) для класифікатора становить 0,662, що вказує на хорошу загальну якість моделі. Другий графік показує різні метрики продуктивності (точність, чутливість, F1-оцінка, специфічність, NPV) при варіації порогового значення. Зі збільшенням порогу точність зростає, але повнота зменшується, і навпаки. Максимальна F1-оцінка досягається при порозі 0,4, а криві точності та чутливості перетинаються, вказуючи на гарну збалансованість моделі. Специфічність максимальна при порозі 0,7. Загалом, ці графіки надають всебічний аналіз продуктивності моделі, що дозволяє вибрати оптимальний поріг для реального застосування. При аналізі NPV (негативної прогностичної цінності) спостерігається стабільне зростання при низьких порогах, що важливо для мінімізації помилкових негативів. Баланс між метриками є критичним для забезпечення точності прогнозування в різних сценаріях. Отримані дані свідчать

про можливість адаптації моделі до потреб користувачів залежно від специфіки застосування. Детальний аналіз кривих також дозволяє оцінити компроміс між чутливістю та специфічністю для різних типів пацієнтів. Наприклад, в умовах високої вартості помилкових негативів важливо вибирати поріг, який мінімізує ризик пропуску хвороби. Таким чином, модель може бути налаштована для різних клінічних сценаріїв, забезпечуючи більш точне прогнозування.

У таблиці 3.1 представлені показники продуктивності моделі LightGBM для різних значень гіперпараметра k (кількість обраних ознак). Метрики включають P1, P2, recall, precision, R1, ROC AUC та score_sum.

Таблиця 3.1

Показники продуктивності моделі LightGBM

k	fold	lpa	P1	recall1	P2	recall2	precision3	R1	ROCAUC	new_idx	k	score_sum
5	4.5	200.0	0.2561	0.1116	0.2323	0.2120	0.3956	0.0044	0.6106	0	5	0.4928
10	4.5	200.0	0.2961	0.1090	0.2500	0.2143	0.4238	0.0113	0.6347	1	10	0.5574
50	4.5	200.0	0.3335	0.1050	0.2816	0.2084	0.4136	0.0341	0.6600	2	50	0.6492
100	4.5	200.0	0.3236	0.1054	0.2825	0.2127	0.4077	0.0399	0.6647	3	100	0.6460
500	4.5	200.0	0.3517	0.1048	0.2969	0.2123	0.4085	0.0527	0.6719	4	500	0.7013
1000	4.5	200.0	0.3515	0.1051	0.3006	0.2071	0.4126	0.0470	0.6726	5	1000	0.6991
2000	4.5	200.0	0.3518	0.1070	0.2918	0.2147	0.4109	0.0570	0.6725	6	2000	0.7006
3000	4.5	200.0	0.3595	0.1056	0.2943	0.2156	0.4105	0.0504	0.6720	7	3000	0.7042
4000	4.5	200.0	0.3579	0.1045	0.2937	0.2140	0.4096	0.0582	0.6716	8	4000	0.7098
5000	4.5	200.0	0.3543	0.1056	0.2930	0.2145	0.4079	0.0565	0.6724	9	5000	0.7038
6381	4.5	200.0	0.3558	0.1048	0.2939	0.2128	0.4136	0.0600	0.6722	10	6381	0.7097

Значення P1 (перший показник точності) зростають із збільшенням k, досягаючи максимальних значень при k=5000–50000, після чого стабілізуються. P2 (другий показник точності) демонструє схожу динаміку, але пікове значення досягається при k=5000.

Recall1 (перша оцінка повноти) та Recall2 поступово зростають із k, демонструючи стійкі тенденції покращення. Precision3 (точність) має коливання при малих k, стабілізуючись після k=3000.

Метрика R1 (перший специфічний показник) стрімко зростає при переході від малих k (5, 10, 50) до 1000. Значення ROC AUC (область під кривою)

поступово зростає та стабілізується при k>1000, вказуючи на високу якість класифікації.

Графіки коробкових діаграм на рис 3.2 та рис 3.3 показують розподіл значень метрик для кожного k. У правій частині графіків спостерігається зменшення варіативності результатів із більшими k, що вказує на стабільність моделі.

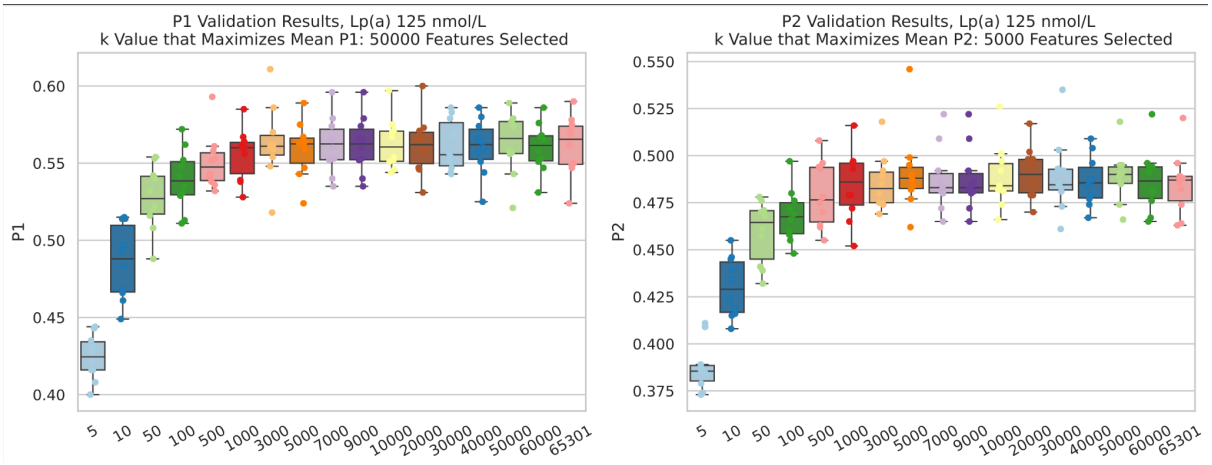


Рис 3.2 Діаграми розподілу значень P1 та P2

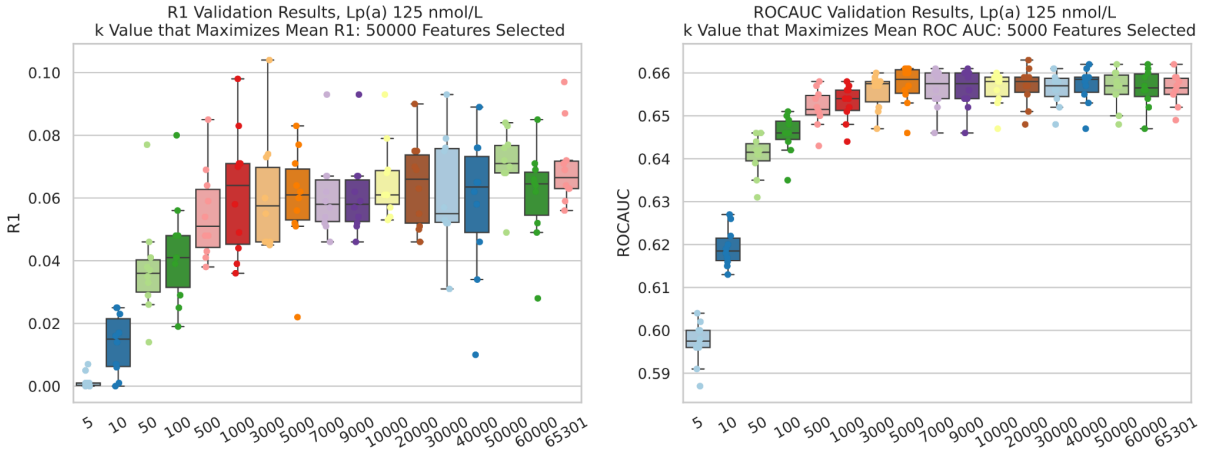


Рис 3.3 Діаграми розподілу значень R1 та ROCAUC

Найвищі значення P1 і P2 досягаються при k=5000 і k=50000, відповідно. Вибір k впливає на продуктивність, оскільки менші k призводять до нестабільних результатів. Результати підтверджують, що збільшення кількості ознак сприяє підвищенню точності до певного рівня.

Значення score_sum демонструє зростання досягненням k=3000 та продовжує залишатися високим при більшому k. Максимальна стабільність

досягається при k>3000, хоча збільшення не дає значного приросту продуктивності. Модель демонструє стійкість до великої кількості ознак, але наявний певний рівень перенавчання на дуже великих k. Графіки підтверджують, що оптимальними є k у діапазоні 3000–5000.

Метрики, такі як ROC AUC, свідчать про високу якість класифікації. Різниця між P1 та P2 вказує на певну залежність від визначення позитивного класу. Вибір оптимального k забезпечує баланс між продуктивністю, стабільністю та ефективністю моделі. Зі збільшенням k модель стає менш чутливою до випадкових змін у вибірці, що покращує її узагальнюючу здатність. Однак надмірно великі значення k можуть збільшити час обчислень, що важливо враховувати в реальних застосуваннях. Подальший аналіз оптимізації k може допомогти знайти ідеальний компроміс між продуктивністю та витратами обчислювальних ресурсів.

Модель LightGBM, застосована для прогнозування рівня Lp(a) у пацієнтів, показала високу ефективність, залежно від вибору кількості ознак (k). Збільшення k до 3000–5000 забезпечує оптимальне поєднання точності, стабільності та швидкості моделі, тоді як більші значення k призводять до перенавчання. Результати демонструють, що модель ефективно справляється з великою кількістю ознак, хоча на певному етапі збільшення розмірності не покращує продуктивність. Оцінка метрик, таких як ROC AUC, показує високу якість класифікації, що робить модель придатною для медичних прогнозів. Для створення додатків прогнозування захворювань важливо інтегрувати оптимальний вибір параметрів, таких як k, щоб уникнути перенавчання і досягти стабільних результатів. Система повинна бути здатною до обробки великих масивів даних, що можуть включати різноманітні характеристики пацієнтів, від віку до результатів лабораторних аналізів. У таких додатках також слід враховувати зручний інтерфейс для користувачів, оскільки медичний персонал потребує зрозумілих та точних результатів для прийняття рішень. Використання методів зниження розмірності може підвищити інтерпретованість моделі. Також важливо враховувати етичні аспекти, зокрема захист даних пацієнтів та правильне використання прогнозів. Це

дозволить створювати точні й безпечні медичні прогнози, що можуть стати основою для ефективної діагностики та лікування захворювань.

3.3 Розробка додатків прогнозування захворювань на основі платформи HPE Haven OnDemand

Компанія Hewlett Packard Enterprise (HPE) досягла значного прогресу у використанні великих даних і своєї платформи HPE Haven OnDemand для прогнозної аналітики в охороні здоров'я, зокрема у прогнозуванні захворювань. HPE Haven OnDemand є сервіс-платформою, яка відноситься до технології Big Data.

HPE Haven OnDemand – це хмарна платформа для обробки великих обсягів даних, Платформа надає інструменти для обробки та аналізу великих обсягів неструктурованих і структурованих даних. У сфері охорони здоров'я вона може об'єднувати записи пацієнтів, дані досліджень і показники здоров'я в реальному часі, щоб передбачити спалахи захворювань або індивідуальні ризики для пацієнтів.

HPE Haven OnDemand – це інструмент сервісу, який використовує потужності машинного навчання та аналізу неструктурованих даних для вирішення різноманітних завдань. Він надає інструменти для аналізу великих обсягів даних, включаючи текстові, аудіо-, відео- та інші формати. Haven OnDemand базується на хмарних обчисленнях і пропонує Application Programming Interface (API) для інтеграції в різні додатки, що дозволяє розробникам працювати з Big Data у сфері машинного навчання, аналізу текстів, пошуку та обробки даних.

Розробку діагностичного інструменту HPE Haven OnDemand для оцінки ризику захворювань робиться на основі прогнозування рівня (Lp(a)) в крові. Ліпопротеїни в крові – це комплекси білків і ліпідів, які транспортують жири, такі як холестерин і тригліцериди, в організмі. Вони відіграють важливу роль у метаболізмі жирів, але їх дисбаланс може призводити до розвитку серцево-судинних захворювань. Тому за його допомогою можна здійснювати

прогнозування ризиків рівня захворювань ССЗ. Це досягається створенням наступних інструментів [19, 20, 21]:

1. Для прогнозування рівня Lp(a) в крові сервіс-платформа HPE Haven OnDemand підтримує такі функції: збір та обробка даних (обробка текстів, розпізнавання емоцій, категоризація даних); обробка мультимедіа (розпізнавання зображень, аудіо та відео); інтеграція з іншими технологіями (платформа може використовуватися в бізнес-аналітиці, автоматизації процесів і прийнятті рішень).

2. За допомогою прогнозної аналітики, моделювання та машинного навчання сервіс-платформа HPE Haven OnDemand інтегрує вдосконалені алгоритми машинного навчання для виявлення шаблонів у медичних даних. Може виявляти кореляцій між рівнем Lp(a) і такими факторами, як генетичні дані, рівень холестерину, дієта, фізична активність тощо. Створює аналітичні моделі для прогнозування на основі історичних даних пацієнтів. Ці моделі є вирішальними для ранньої діагностики та індивідуального планування лікування.

3. На основі спостереження та моніторингу захворювань сервіс-платформа HPE Haven OnDemand може прогнозувати рівень Lp(a), враховуючи індивідуальні характеристики пацієнта. Може аналізувати різноманітні джерела даних, такі як соціальні мережі, стан навколишнього середовища та лікарняні записи, щоб передбачити потенційні спалахи захворювань і реагувати на них у режимі реального часу.

4. Сервіс-платформа HPE Haven OnDemand здійснює співпрацю з постачальниками медичних послуг (з медичними установами) для інтеграції прогнозної аналітики в існуючі системи охорони здоров’я. Може виконувати візуалізація результатів, тобто інтерпретацію результатів через аналітичні панелі та графіки для медичного персоналу, що спрощує прийняття рішень щодо лікування або додаткових досліджень. Це допомагає покращити процеси прийняття рішень і розподіл ресурсів.

5. Сервіс-платформа HPE Haven OnDemand забезпечує масштабованість та хмарну інтеграцію: підтримує масштабовані хмарні рішення, що робить їх доступними як для великих систем охорони здоров’я, так і для невеликих клінік.

Таким чином, HPE Haven OnDemand – це інструмент, який спрощує використання великих масивів даних, що допомагає не тільки виявляти можливі спалахи захворювань, але й прогнозувати індивідуальні ризики для пацієнтів на основі великого обсягу даних. Розробка таких діагностичних інструментів може значно покращити раннє виявлення ризику серцево-судинних захворювань і, відповідно, забезпечити своєчасну профілактику і лікування пацієнтів.

Впровадження сервіс-платформи HPE Haven OnDemand в систему прогнозування захворювань має потенціал для трансформації медицини та охорони здоров'я, роблячи профілактику захворювань точнішою та ефективнішою. Система, заснована на технологіях машинного навчання та використанні найкращих практик аналізу великих даних, дозволяє прогнозувати ризик підвищеного рівня Lp(a) у пацієнтів. Це надає можливість як пацієнтам, так і лікарям чи страховим компаніям звернути увагу на потенційні ризики, які можуть бути підтверджені лабораторними дослідженнями. Відповідно, пацієнти, які ще не зіткнулися з серцево-судинними захворюваннями, можуть коригувати свої показники здоров'я, щоб уникнути подальших проблем. Таким чином, можемо покращити рівень задоволеності системою охорони здоров'я, досягти значного економічного ефекту через зменшення тимчасової або постійної непрацездатності населення та використати ці напрацювання для подальшого розширення застосування в інших галузях медицини [18].

Дослідження представляє складний підхід до прогнозування серцево-судинних захворювань з використанням передових методів машинного навчання, зокрема зосереджуючи увагу на інноваційному застосуванні алгоритму LightGBM та платформи HPE Haven OnDemand. Аналізуючи комплексний набір даних із 142674 записів пацієнтів та 65301 унікальною ознакою, у дослідженні був розроблений прогностичний модель для оцінки ризику підвищення рівня Lp(a), важливого маркера серцево-судинного здоров'я. Модель LightGBM показала видатну продуктивність з оптимальним вибором ознак у межах 3000-5000, забезпечуючи баланс між точністю, стабільністю та обчислювальною ефективністю при мінімізації ризиків перенавчання.

Методологія використовувала складні техніки, такі як Leaf-wise ріст та гістограмне розбиття, що дозволило створити більш точну та обчислювально ефективну модель машинного навчання. Дослідження виявило, що збільшення кількості вибраних ознак до певного порогу значно покращує продуктивність моделі, при цьому такі метрики, як ROC AUC, свідчать про високоякісні можливості класифікації. Здатність моделі працювати з великими та складними наборами даних, що містять різноманітні характеристики пацієнтів – від віку до результатів лабораторних тестів – підкреслює її потенціал для передових медичних прогнозів та персоналізованих медичних втручань.

Інтеграція платформи HPE Haven OnDemand представляє трансформаційний підхід до аналізу медичних даних, використовуючи хмарні обчислення та передові алгоритми машинного навчання для обробки як структурованих, так і неструктурованих даних. Ця платформа дозволяє здійснювати комплексне прогнозування медичних ризиків, аналізуючи різноманітні джерела даних, включаючи записи пацієнтів, генетичну інформацію, фактори способу життя та екологічні дані. Завдяки наданню реальних інсайтів та інструментів для співпраці серед медичних працівників, система сприяє більш обґрунтованому прийняттю рішень, що може призвести до революційних змін у стратегіях профілактики захворювань та розподілі ресурсів у медичних установах.

Ширші наслідки цього дослідження виходять за межі безпосередніх медичних застосувань, пропонуючи зсув парадигми в превентивній медицині. Завдяки можливості раннього виявлення ризиків серцево-судинних захворювань, розроблена прогностична модель надає пацієнтам та медичним працівникам можливість вживати проактивних заходів, що потенційно знижує довгострокові проблеми зі здоров’ям та економічні витрати. Дослідження підкреслює критичну роль технологій Big Data та машинного навчання в трансформації охорони здоров’я, показуючи, як передовий аналіз даних може забезпечити нюансовані, персоналізовані медичні інсайти, які раніше були недоступними, та закладає основи для майбутніх інновацій у прогностичній та прецизійній медицині.

ВИСНОВКИ

Серцево-судинні захворювання є значним глобальним навантаженням на національні системи охорони здоров'я. Ці захворювання створюють серйозні соціальні та економічні проблеми через підвищену смертність та інвалідність. Ситуація ускладнюється старінням населення та зростаючою поширеністю факторів ризику, таких як ожиріння, малорухливий спосіб життя та стрес. Основними чинниками цього навантаження є ішемічна хвороба серця (ІХС), цереброваскулярні захворювання та гіпертонія. Ця ситуація підкреслює нагальну потребу в ефективних стратегіях профілактики, раннього виявлення та лікування для боротьби з підвищенням рівня ССЗ в усьому світі.

Основними факторами ризику для ССЗ є нездорові звички, такі як неправильне харчування (з високим вмістом насичених жирів, цукру та солі), відсутність фізичної активності, куріння та надмірне споживання алкоголю. Хронічні захворювання, такі як діабет, гіпертонія та розлади ліпідного обміну, додатково підвищують ризик. Крім того, соціальні фактори, такі як стрес, бідність та обмежений доступ до якісної медичної допомоги, погіршують ситуацію. Генетична схильність також має значний вплив, оскільки певні спадкові мутації, такі як ті, що стосуються гена APOE, збільшують вразливість до ССЗ. Однак багато випадків ССЗ можна запобігти за допомогою змін у способі життя, ранньої діагностики та своєчасного втручання. Взаємодія генетичних та поведінкових факторів підкреслює необхідність комплексного підходу до профілактики, що поєднує генетичний аналіз з корекцією поведінкових факторів. Такі заходи можуть значно знизити ризик розвитку ССЗ.

Складність діагностики ССЗ, особливо на ранніх стадіях, вимагає впровадження передових діагностичних технологій. Інновації, такі як кардіологічні методи візуалізації, біомаркери, генетичне тестування та обчислювальні підходи, включаючи машинне навчання (ML), змінюють діагностику та управління ССЗ. Ці інструменти надають точні та персоналізовані

інсайти щодо серцево-судинного здоров’я, що дозволяє раннє виявлення аномалій, прогнозування ризиків та індивідуалізовану допомогу пацієнтам. Інтеграція традиційних діагностичних методів, таких як клінічні огляди, з сучасними техніками, такими як МРТ, КТ та алгоритми ML, підвищує точність та ефективність діагностики. Однак впровадження цих технологій потребує вирішення проблем, таких як економічна ефективність, доступність і потреба в кваліфікованих медичних працівниках для управління передовими інструментами.

Машинне навчання стало потужним інструментом у діагностиці та управлінні ССЗ. Алгоритми ML можуть аналізувати величезні обсяги клінічних даних для виявлення складних патернів і прогнозування ризиків захворювань. Наприклад, моделі ML використовуються для виявлення аномалій у медичних зображеннях, таких як ехокардіограми та ЕКГ, що сприяє ранньому виявленню аритмій та інших серцевих захворювань. Крім того, ML дозволяє проводити персоналізовану оцінку ризиків, аналізуючи індивідуальні дані пацієнтів, включаючи генетичні маркери та фактори способу життя. Дослідження показали високу ефективність моделей на основі ML у покращенні точності діагностики та підтримці клінічного прийняття рішень. Ці досягнення підкреслюють потенціал ML для революціонізації діагностики та лікування ССЗ через дані, орієнтовані на індивідуалізацію.

Інтеграція технологій Big Data в системи прогнозування захворювань відкрила нові можливості для профілактики та управління ССЗ. Аналітика Big Data дозволяє обробляти величезні обсяги структурованих і неструктурованих даних, надаючи цінні інсайти щодо тенденцій здоров’я та факторів ризику. Наприклад, платформи, такі як Apache Spark та Hadoop, сприяють аналізу даних у реальному часі, покращуючи прогнозування ризиків ССЗ. Застосування Big Data також підтримує оптимізацію медичних ресурсів, персоналізовану медицину та раннє виявлення епідемій. Однак важливо вирішити проблеми кібербезпеки, пов'язані з системами Big Data, для захисту конфіденційної інформації пацієнтів. Надійне шифрування даних, багаторівнева автентифікація та регулярні оновлення

програмного забезпечення є необхідними для забезпечення безпеки та надійності таких систем.

Алгоритм LightGBM виявився високоефективною моделлю машинного навчання для прогнозування ризиків ССЗ. Аналізуючи набір даних з понад 142674 записів пацієнтів із 65301 унікальною ознакою, LightGBM продемонстрував кращу продуктивність порівняно з іншими методами градієнтного бустингу. Його інноваційні характеристики, такі як зростання за принципом leaf-wise та гістограмне розбиття, оптимізують обчислювальну ефективність та точність. Модель досягла високих показників продуктивності, зокрема за ROC AUC, що свідчить про її надійність у класифікації ризиків ССЗ. Крім того, здатність LightGBM обробляти великі набори даних з різноманітними характеристиками пацієнтів підкреслює його потенціал для розробки персоналізованих рішень у галузі охорони здоров’я. Однак правильне налаштування параметрів є необхідним для уникнення перенавчання та забезпечення надійних прогнозів.

Платформа HPE Haven OnDemand пропонує передову основу для створення додатків для прогнозування ССЗ. Ця платформа використовує хмарні обчислення та алгоритми машинного навчання для аналізу медичних даних, включаючи записи пацієнтів, генетичну інформацію та фактори способу життя. Вона підтримує моніторинг у реальному часі та надає медичним працівникам дієві інсайти через інтуїтивно зрозумілі панелі моніторингу та візуалізації. Крім того, HPE Haven OnDemand сприяє співпраці між медичними працівниками шляхом інтеграції прогностичної аналітики в існуючі медичні системи. Її масштабованість та хмарна архітектура роблять її доступною як для великих медичних мереж, так і для менших клінік. Завдяки точному та своєчасному прогнозуванню ризиків ССЗ ця платформа дозволяє клініцистам приймати більш обґрунтовані рішення та покращувати результати для пацієнтів.

Це дослідження підкреслює трансформаційний вплив технологій Big Data та ML на профілактику та управління ССЗ. Модель LightGBM та платформа HPE Haven OnDemand демонструють, як передова аналітика може надати нюансовані,

персоналізовані медичні інсайти, відкриваючи шлях до прецизійної медицини. Раннє виявлення ризиків ССЗ дозволяє пацієнтам та медичним працівникам вжити проактивних заходів, що потенційно знижує довгострокові проблеми зі здоров'ям та пов'язані витрати. Ці результати підкреслюють необхідність подальших інвестицій в інноваційні медичні технології для вирішення зростаючого навантаження від ССЗ. Майбутні дослідження повинні зосередитись на покращенні інтерпретованості моделей, забезпеченні безпеки даних та інтеграції етичних міркувань для створення довіри та максимізації переваг прогнозних медичних систем.

ПЕРЕЛІК ПОСИЛАНЬ

1. World Health Organisation. Cardiovascular diseases. URL: https://www.who.int/health-topics/cardiovascular-diseases
2. Серцево-судинні захворювання – головна причина смерті українців. Висновки з дослідження Глобального тягаря хвороб у 2019 році | Центр громадського здоров'я. URL: https://phc.org.ua/news/sercevo-sudinni-zakhvoryuvannya-golovna-prichina-smerti-ukrainciv-visnovki-z-doslidzhennya
3. Armitage P., Berry G., Matthews J. N. S. Statistical Methods in Medical Research, 4th Edition. 2001. 832 p.
4. VanderPlas J. Python Data Science Handbook. O'Reilly Media, Inc., 2016. 548 p.
5. Stouffer G. A., Runge M. S., Patterson C., Rossi J. C. Netter's Cardiology, 3rd edition. Philadelphia, PA, Saunders/Elsevier, 2019. 576 p.
6. Netter F. H. Atlas of Human Anatomy. Philadelphia, PA, Saunders/Elsevier, 2014. 5. World Health Organisation. Cardiovascular diseases. URL: https://www.who.int/health-topics/cardiovascular-diseases
7. The American Heart Asociation. Heart Disease and Stroke Statistics–2016 Update. 2016. URL: https://www.ahajournals.org/doi/10.1161/cir.0000000000000350
8. Maas A., Appleman Y. Gender differences in coronary heart disease. URL: https://pubmed.ncbi.nlm.nih.gov/21301622/
9. American Diabetes Association. Cardiovascular Disease and Risk Management: Standards of Medical Care in Diabetes. 2020. URL: https://diabetesjournals.org/care/article/43/Supplement_1/S111/30374/10-Cardiovascular-Disease-and-Risk-Management
10. Poplin R., Varadarajan A. V., Blumer K., Liu Y., McConnell M. V., Corrado G. S., Peng L., Webster D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. 2018. URL: https://www.nature.com/articles/s41551-018-0195-

11. Subramani S., Varshney N., Anand M. V., Soudagar M. E. M., Al-keridis L. A., Upadhyay T. K., Alshammari N., Saeed M., Subramanian K., Anbarasu K., Rohini K. Cardiovascular diseases prediction by machine learning incorporation with deep learning. 2023. URL: https://www.frontiersin.org/articles/10.3389/fmed.2023.1150933/full
12. Dritsas E., Trigka M. Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction. 2023. URL: https://www.mdpi.com/1424-8220/23/3/1161
13. Barda N., Dagan N., Stemmer A., Yuval J., Bachmat E., Elnekave E., Balicer R. Improving Cardiovascular Disease Prediction Using Automated Coronary Artery Calcium Scoring from Existing Chest CTs. 2022. URL: https://pubmed.ncbi.nlm.nih.gov/35296940/
14. Narsullah N., Sang J., Alam M. S., Mateen M., Cai B., Hu H. Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies. 2019. URL: https://www.mdpi.com/142 4-8220/19/17/3722
15. Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. 2020. URL: https://www.scirp.org/journal/paperinformation.aspx?paperid=99402
16. McKinney W. Python for Data Analysis, 2nd Edition. O'Reilly Media, Inc., 2017. 547 p.
17. Accuracy, Precision, Recall or F1? | by Koo Ping Shung | Towards Data Science. URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
18. TRENDS OF THE FUTURE: RISKS, OPPORTUNITIES, TASKS Collection of Materials of the Multidisciplinary Scientific and Practical Conference Kyiv, December 23th, 2016/ [Електронний ресурс] / режим доступу: http://futurolog.com.ua/publish/3/Zbirnyk.pdf/ (date of access: 20.10.2024

19. Scipione CA, Koschinsky ML, Boffa MB. Lipoprotein(a) in clinical practice: New perspectives from basic and translational science. Crit Rev Clin Lab Sci. 2018 Jan;55(1):33-54.
20. Marcovina SM, Albers JJ. Lipoprotein (a) measurements for clinical application. J Lipid Res. 2016 Apr;57(4):526-37.
21. Kostner KM, März W, Kostner GM. When should we measure lipoprotein (a)? Eur Heart J. 2013 Nov;34(42):3268-76.
22. Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30, 2017.

ДОДАТОК А

iPatient_ID_fk,sPatient_Ethnicity,sPatient_Gender_Cde,HDL,LDL,p_sPx_Place_Of_Service,sRx_Nme,Antihypertensive_Medication,Total_MI_Events,Total_Stroke_Events,NEPHROTIC_SYNDROME,CROHNS_DISEASE,ULCERATIVE_COLITIS,PSORIATIC_ARTHRITIS,AORTIC_VALVE_STENOSIS,CHRONIC_KIDNEY_DISEASE,HYPERLIPIDEMIA,HYPERTENSION,MENOPAUSE,RHEUMATOID_ARTHRITIS,HYPERTRIGLYCERIDEMIA,labAge,LLT_Medication

1,Caucasian,Male,,,[Inpatient,Outpatient],[Potassium Chloride,Ibuprofen],,,,False,False,False,False,False,False,False,False,False,False,False,63,False

2,Unknown,Female,,,[Inpatient],[],,,1,False,False,False,False,False,False,False,False,False,False,False,79,False

3,Caucasian,Male,,79,[Inpatient,Outpatient,Office],[],,,,False,False,False,False,False,False,False,False,False,False,False,62,False

4,Unknown,Male,,59,[Inpatient],[],,,,False,False,False,False,False,False,False,False,False,False,False,66,False

5,Unknown,Female,,92,[Clinic,Other],[],,,,False,False,False,False,False,False,False,False,False,False,False,52,False

6,Unknown,Male,,65,[Outpatient],[],,,,False,False,False,False,False,False,False,False,False,False,False,58,False

7,Unknown,Female,,,[Outpatient,Inpatient,Office,Other],[],,,,False,False,False,False,False,False,False,False,False,False,False,43,False

8,Caucasian,Male,,,[Inpatient,Outpatient,Office,Other],[Hydrocodone/Acetaminophen,Penicillin V Potassium],,,,False,False,False,False,False,False,False,False,False,False,False,69,False

9,Other,Male,,,[Outpatient,Inpatient,Office,Other],[],,,12,False,False,False,False,False,False,False,False,False,False,False,False,32,False

10,Caucasian,Male,,,[Inpatient,Outpatient,Office,Unknown,Other],[],,,,False,False,False,False,False,False,False,False,False,False,False,70,False

11,Caucasian,Male,,,[Inpatient,Outpatient],[Hydrocodone/Acetaminophen,Cephalexin],,,,False,False,False,False,False,False,False,False,False,False,False,70,False

12,Unknown,Female,,,[Other],[],,,,False,False,False,False,False,False,False,False,False,False,False,58,False

13,Caucasian,Male,,,[Inpatient],[],,,,False,False,False,False,False,False,False,False,False,False,False,77,False

14,Unknown,Male,,,[Outpatient,Office,Other],[Pravastatin Sodium,Dipyridamole],,,,False,False,False,False,False,False,False,False,False,False,False,52,True

15,Caucasian,Female,,100,[Outpatient,Inpatient,Office,Other],[Flu Vaccine Ts 2013-14(5yr,Up)],,,,False,False,False,False,False,False,False,False,False,False,False,60,False

16,Unknown,Male,,,[Inpatient,Office],[],,,1,False,False,False,False,False,False,False,False,False,False,False,42,False

17,Caucasian,Male,,,[Outpatient,Inpatient,Office,Other],[],,,3,False,False,False,False,False,False,False,False,False,False,False,55,False

18,Unknown,Female,,120,[Outpatient,Office],[],,,2,False,False,False,False,False,False,False,False,False,False,False,68,False

19,Unknown,Male,,,[Outpatient,Inpatient],[],,,,False,False,False,False,False,False,False,False,False,False,False,63,False

20,Unknown,Male,,,[Inpatient,Outpatient,Office],[],,1,,False,False,False,False,False,False,False,False,False,False,False,47,False

ДОСЛІДЖЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ ЗАХВОРЮВАНЬ НА ОСНОВІ ТЕХНОЛОГІЇ BIG DATA

КНДМ-62 Гоменюк Андрій

Мета дослідження – узагальнення наявних теоретичних підходів використання технологій Big Data для прогнозування ССЗ.

Об'єкт дослідження – процес прогнозування серцево-судинних захворювань.

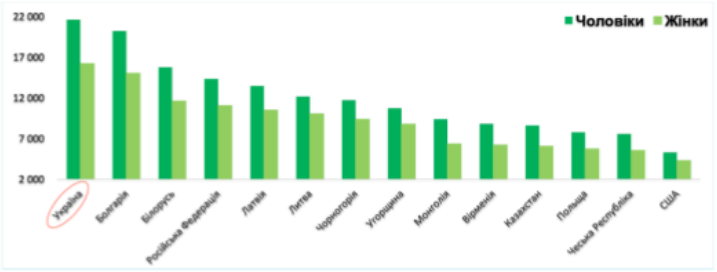
Предмет дослідження – системи прогнозування ССЗ за допомогою технологій Big Data.

Наукові завдання дослідження: проаналізувати наявні технологічні інструменти прогнозування ССЗ; виявити релевантні технології Big Data для побудови системи прогнозування ССЗ; розробити систему прогнозування ССЗ за допомогою технологій Big Data.

Актуальність дослідження

ССЗ – одна з основних причин смертності в усьому світі, а отже це актуальна і глобальна проблема

Медіанний вік групи ризику знизився, тобто ССЗ стали частіше діагностувати у пацієнтів молодшого віку. Україна – лідер за кількістю втрачених років життя через ССЗ



Методи діагностики ССЗ

Традиційні:

- аускультація серцевих звуків
- пальпація периферичних пульсів
- лабораторні аналізи крові
- ЕКГ
- ЕхоКГ
- стрес-тести

Інноваційні:

- МРТ
- КТА
- ПЕТ
- генетичні тести
- обчислювальні підходи за допомогою машинного навчання

Big Data

Volume - відображає величезний обсяг даних, який генерується щодня

Variety - великий обсяг даних може бути представленим у різних форматах: від структурованих (таблиці баз даних) до напівструктурованих (JSON, XML) та неструктурованих даних (тексти, відео, зображення).

Velocity - Швидкість, з якою генеруються, передаються та обробляються дані, є ключовим аспектом Big Data.

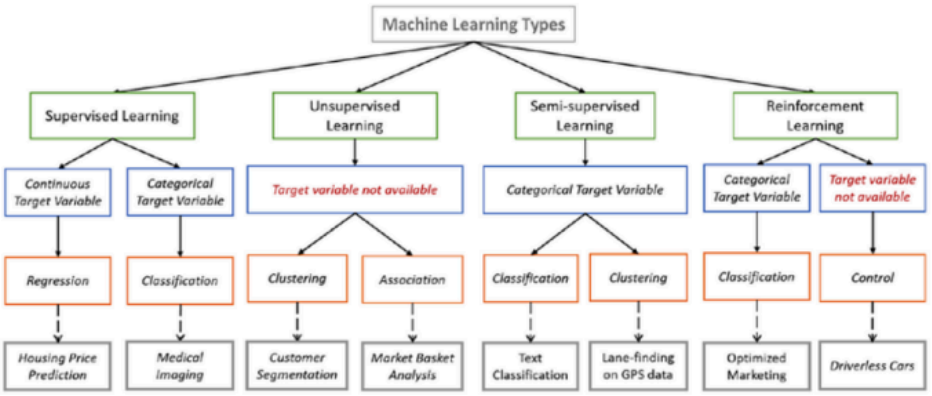
Veracity - відображає якість та правдивість даних.

Value - цінність для прийняття рішень є найбільш важливою характеристикою.

Hadoop

- **Hadoop Distributed File System (HDFS)**. Це файлова система, що дозволяє зберігати великі обсяги даних, розподілених по багатьох комп'ютерах (вузлах).
- **MapReduce**. Це модель програмування, що дозволяє обробляти великі дані паралельно на розподілених комп'ютерах. Вона складається з двох етапів: Map (розподіл даних) і Reduce (агрегація результатів).
- **YARN (Yet Another Resource Negotiator)**. Це система управління ресурсами в Hadoop, що відповідає за координацію та керування розподіленими обчислювальними ресурсами.

Машинне навчання - Типи



Набір даних

iPatient_ID_fk	sPatient_Ethnicity	sPatient_Gender_Cde	HDL	LDL	Total_MI_Events	Total_Stroke_Events	labAge
5900251	CAUCASIAN	MALE					65
5614411	CAUCASIAN	MALE					73
29635233	CAUCASIAN	MALE					54
598242	CAUCASIAN	FEMALE					56
9097361	BLACK/AFRICAN AMERICAN	MALE	66.0	125.0	1		75
4123741	CAUCASIAN	MALE	33.0	53.0			72
82276082	BLACK/AFRICAN AMERICAN	MALE	29.0	134.0			61
25174303	CAUCASIAN	MALE					80
28613110	CAUCASIAN	MALE			1		62

Метрики

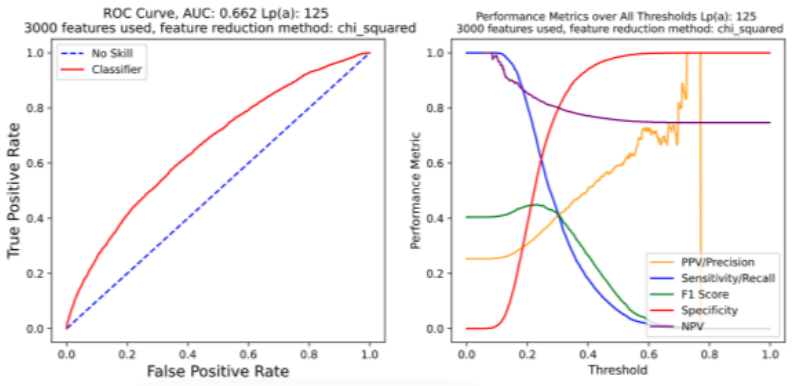
$$\text{Accuracy} = \frac{TN + TP}{TN + FN + FP + TP} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

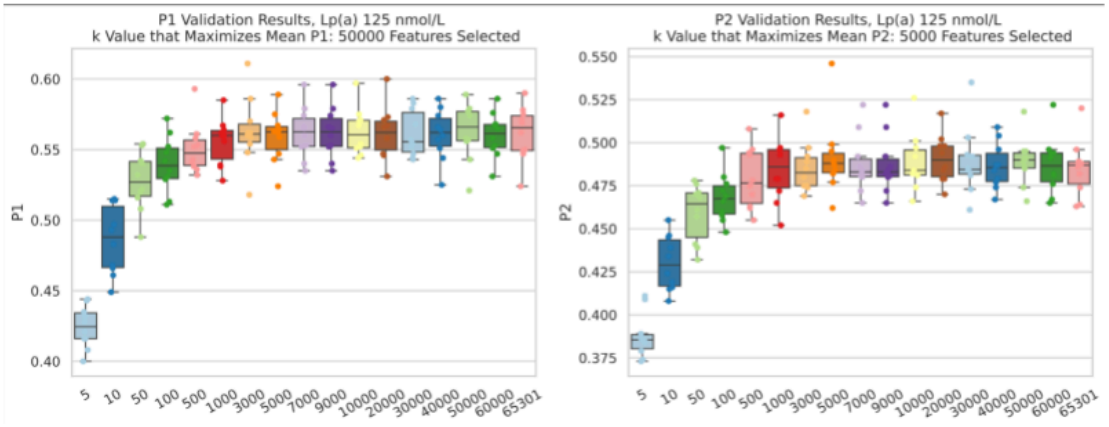
Крива ROC та метрики ефективності моделі



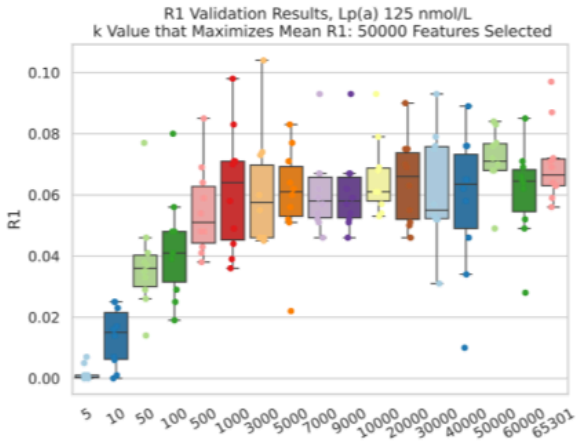
Показники продуктивності моделі LightGBM

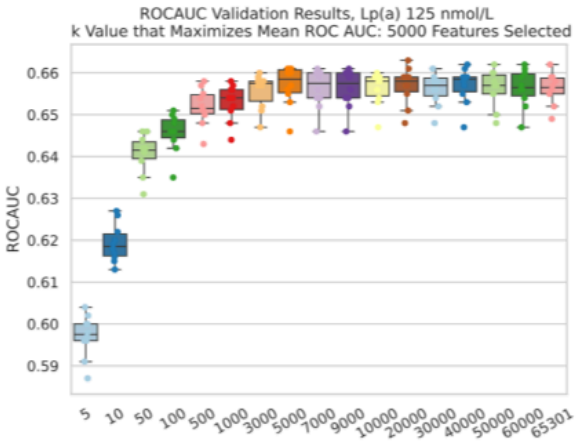
k	fold	lpa	P1	recall1	P2	recall2	precision3	R1	ROCAUC	new_idx	k	score_sum
5	4.5	200.0	0.2561	0.1116	0.2323	0.2120	0.3956	0.0044	0.6106	0	5	0.4928
10	4.5	200.0	0.2961	0.1090	0.2500	0.2143	0.4238	0.0113	0.6347	1	10	0.5574
50	4.5	200.0	0.3335	0.1050	0.2816	0.2084	0.4136	0.0341	0.6600	2	50	0.6492
100	4.5	200.0	0.3236	0.1054	0.2825	0.2127	0.4077	0.0399	0.6647	3	100	0.6460
500	4.5	200.0	0.3517	0.1048	0.2969	0.2123	0.4085	0.0527	0.6719	4	500	0.7013
1000	4.5	200.0	0.3515	0.1051	0.3006	0.2071	0.4126	0.0470	0.6726	5	1000	0.6991
2000	4.5	200.0	0.3518	0.1070	0.2918	0.2147	0.4109	0.0570	0.6725	6	2000	0.7006
3000	4.5	200.0	0.3595	0.1056	0.2943	0.2156	0.4105	0.0504	0.6720	7	3000	0.7042
4000	4.5	200.0	0.3579	0.1045	0.2937	0.2140	0.4096	0.0582	0.6716	8	4000	0.7098
5000	4.5	200.0	0.3543	0.1056	0.2930	0.2145	0.4079	0.0565	0.6724	9	5000	0.7038
6381	4.5	200.0	0.3558	0.1048	0.2939	0.2128	0.4136	0.0600	0.6722	10	6381	0.7097

Діаграми розподілу значень P1 і P2



Діаграми розподілу значень R1 та ROCAUC





Висновки

- Основними факторами ризику для ССЗ є нездорові звички, такі як неправильне харчування (з високим вмістом насичених жирів, цукру та солі), відсутність фізичної активності, куріння та надмірне споживання алкоголю.
- Складність діагностики ССЗ, особливо на ранніх стадіях, вимагає впровадження передових діагностичних технологій.
- Інтеграція технологій Big Data в системи прогнозування захворювань відкрила нові можливості для профілактики та управління ССЗ. Аналітика Big Data дозволяє обробляти величезні обсяги структурованих і неструктурованих даних, надаючи цінні інсайти щодо тенденцій здоров'я та факторів ризику.
- Алгоритм LightGBM виявився високоефективною моделлю машинного навчання для прогнозування ризиків ССЗ. Аналізуючи набір даних з понад 142674 записів пацієнтів із 65301 унікальною ознакою, LightGBM продемонстрував кращу продуктивність порівняно з іншими методами градієнтного бустингу.

Результати апробації:

- АКТУАЛЬНІ ПРОБЛЕМИ БЕЗПЕКИ ІНФОРМАЦІЙНО-ТЕЛЕКОМУНІКАЦІЙНИХ СИСТЕМ (5.11.2024). Дослідження безпеки системи прогнозування захворювань на основі технології Big Data

- СУЧАСНІ ДОСЯГНЕННЯ КОМПАНІЇ HEWLETT PACKARD ENTERPRISE В ГАЛУЗІ ІТ ТА НОВІ МОЖЛИВОСТІ ЇХ ВИВЧЕННЯ І ЗАСТОСУВАННЯ (14.12.2024).