Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Застосування методів Explainable AI (XAI) для підвищення прозорості ухвалення рішень у медичних інформаційних системах
(тема)

Виконав:
здобувач другого року навчання,
групи ДСм-24-1
Цопа М.Д.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Науки про дані (Data Science)
(повна назва спеціалізації)
Керівник доц. Магдаліна І. В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Лариса ЧАЛА
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Науки про дані (Data Science)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Цопі Микиті Дмитровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Застосування методів Explainable AI (XAI) для підвищення прозорості ухвалення рішень у медичних інформаційних системах

затверджена наказом університету від 24 листопада 20 25 р. № 1057Ст
2. Термін подання студентом роботи до екзаменаційної комісії 23 грудня 20 25 р.
3. Вихідні дані до роботи macOS Tahoe, середовище розробки PyCharm, бібліотеки: pandas, numpy, matplotlib, sklearn, xgboost, shap, фреймворк streamlit

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі
2) Теоретичні основи Explianble Artificial Intelligence
3) Аналіз даних та алгоритмів машинного навчання
4) Практична реалізація проєкту
5) Практична цінність та подальша реалізація проєкту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	24.11.2025	виконано
2	Аналіз предметної галузі	28.11.2025	виконано
3	Огляд методів Explainable AI	29.11.2025	виконано
4	Розробка медичної інформаційної системи з XAI	05.12.2025	виконано
5	Оформлення пояснювальної записки	15.12.2025	виконано
6	Підготовка презентації та доповіді	18.12.2025	виконано
7	Попередній захист	20.12.2025	виконано
8	Захист перед ЕК	23.12.2025	

Дата видачі завдання 24 листопада 20 25 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Магдаліна І.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 58 с., 24 рис., 1 дод., 12 джерел.

ІНТЕРПРЕТОВАНІСТЬ МОДЕЛЕЙ, МЕДИЧНІ ІНФОРМАЦІЙНІ СИСТЕМИ, ПІДТРИМКА КЛІНІЧНИХ РІШЕНЬ, EXPLAINANBLE ARTIFICIAL INTELLIGENCE, SHAP, XAI, XGBOOST.

Об’єкт дослідження – процес ухвалення рішень у медичних інформаційних системах з використанням методів машинного навчання.

Предмет дослідження – методи Explainable Artificial Intelligence (XAI), що застосовуються для підвищення прозорості, інтерпретованості та обґрунтованості рішень моделей машинного навчання в медичних інформаційних системах.

Мета роботи – підвищення прозорості ухвалення рішень у медичних інформаційних системах шляхом застосування методів Explainable AI для інтерпретації результатів роботи моделей машинного навчання та забезпечення зрозумілих пояснень для медичних фахівців.

Методи дослідження – методи аналізу та обробки медичних даних, алгоритми машинного навчання (градієнтний бустинг), методи пояснюваного штучного інтелекту, зокрема SHAP-аналіз, методи локальної та глобальної інтерпретації моделей, статистичні методи оцінки якості класифікації, а також методи програмної реалізації та експериментального дослідження.

ABSTRACT

Master’s thesis contains: 58 pp., 24 fig., 1 ann., 12 references.

CLINICAL DECISION SUPPORT, EXPLAINABLE ARTIFICIAL INTELLIGENCE, MEDICAL INFORMATION SYSTEMS, MODEL INTERPRETABILITY, SHAP, XAI, XGBOOST.

The object of the research is the decision-making process in medical information systems using machine learning methods.

The subject of the research is Explainable Artificial Intelligence (XAI) methods applied to improve the transparency, interpretability, and justification of machine learning model decisions in medical information systems.

The aim of the research is to enhance the transparency of decision-making in medical information systems by applying Explainable AI methods to interpret the results of machine learning models and to provide understandable explanations for medical professionals.

The research methods include methods of medical data analysis and processing, machine learning algorithms (gradient boosting), Explainable Artificial Intelligence techniques, particularly SHAP analysis, methods of local and global model interpretation, statistical methods for evaluating classification performance, as well as software development and experimental research methods.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI – Artificial Intelligence – штучний інтелект;

CSV – Comma-Separated Values – формат зберігання табличних даних;

DL – Deep Learning – розділ машинного навчання, заснований на використанні багатошарових нейронних мереж для моделювання даних;

ML – Machine Learning – машинне навчання;

PDP – Partial Dependence Plot – графік часткової залежності;

SHAP – SHapley Additive exPlanations – метод пояснення рішень моделей машинного навчання;

UI – User Interface – користувацький інтерфейс;

XAI – Explainable Artificial Intelligence – пояснюваний штучний інтелект;

XGBoost – Extreme Gradient Boosting – алгоритм градієнтного бустингу.

ВСТУП

Штучний інтелект та методи машинного навчання відіграють важливу роль у сучасному світі та знаходять широке застосування у різних сферах діяльності людини. Однією з найбільш перспективних і водночас відповідальних галузей використання таких технологій є медицина. Сьогодні алгоритми машинного навчання застосовуються для підтримки клінічних рішень, прогнозування перебігу захворювань, автоматизованого аналізу медичних зображень, обробки результатів лабораторних досліджень, а також для оцінки ризиків і виявлення прихованих закономірностей у великих обсягах медичних даних. Використання інтелектуальних систем дозволяє підвищити ефективність діагностики, зменшити навантаження на медичний персонал та покращити якість надання медичних послуг.

Разом із тим, зростання складності моделей машинного навчання призводить до появи так званої проблеми «чорної скриньки», коли механізм ухвалення рішень моделлю є незрозумілим для кінцевого користувача. У медичних інформаційних системах така ситуація є особливо критичною, оскільки результати роботи алгоритмів можуть безпосередньо впливати на життя та здоров’я пацієнтів. Лікарі повинні мати можливість не лише отримувати прогноз або рекомендацію від системи штучного інтелекту, але й розуміти причини, які призвели до такого рішення. Відсутність прозорості знижує рівень довіри до автоматизованих систем та обмежує можливість їх практичного застосування у клінічній практиці.

У зв’язку з цим зростає актуальність напрямку Explainable Artificial Intelligence (XAI), який спрямований на розробку методів та підходів для пояснення рішень моделей штучного інтелекту. Методи XAI дозволяють інтерпретувати результати роботи складних моделей машинного навчання, визначати вплив окремих ознак на прогноз, а також аналізувати поведінку моделі як на рівні окремого пацієнта, так і на рівні всієї вибірки даних.

Застосування пояснюваного штучного інтелекту в медичних інформаційних системах сприяє підвищенню прозорості ухвалення рішень, полегшує процес клінічної інтерпретації результатів та забезпечує додаткові інструменти для контролю та валідації моделей.

Машинне навчання як галузь науки і техніки займається розробкою алгоритмів, що дозволяють комп'ютерним системам навчатися на основі даних і покращувати свою продуктивність без явного програмування. Основними компонентами машинного навчання є дані, моделі та алгоритми навчання. Дані слугують основою для побудови моделей і можуть мати різну природу, зокрема числову, текстову або категоріальну, що є характерним для медичних задач. Модель являє собою математичну або алгоритмічну структуру, яка описує залежності між вхідними ознаками та результатом, а алгоритми навчання визначають спосіб налаштування параметрів цієї моделі.

Процес навчання моделі включає підбір оптимальних параметрів з метою мінімізації помилки прогнозування, а також подальшу оцінку якості роботи моделі на незалежних даних. У медичних інформаційних системах особливе значення має не лише точність прогнозування, але й здатність моделі узагальнювати інформацію та надавати стабільні, інтерпретовані результати. Саме тому поєднання методів машинного навчання з підходами Explainable AI є перспективним напрямком досліджень, що дозволяє створювати більш надійні та прозорі системи підтримки медичних рішень.

Таким чином, застосування методів Explainable AI у медичних інформаційних системах є важливим кроком на шляху інтеграції сучасних алгоритмів машинного навчання у практичну медицину. Підвищення прозорості та інтерпретованості рішень сприяє зростанню довіри з боку медичних фахівців, забезпечує відповідність етичним та регуляторним вимогам, а також відкриває нові можливості для розвитку інтелектуальних медичних систем.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Актуальність теми

Розвиток інформаційних технологій та цифровізація охорони здоров’я призвели до значного зростання обсягів медичних даних, які потребують ефективної обробки та аналізу. Сучасні лікувально-профілактичні заклади накопичують великі масиви інформації, що включають електронні медичні записи, результати лабораторних досліджень, дані медичної візуалізації, а також показники моніторингу стану пацієнтів. Обробка таких даних традиційними методами є складною та часто не дозволяє своєчасно отримувати корисні знання для підтримки клінічних рішень.

У зв’язку з цим все більшого поширення набувають методи штучного інтелекту та машинного навчання, які здатні автоматизувати аналіз великих обсягів медичних даних і виявляти приховані закономірності. AI-системи демонструють високу ефективність у задачах діагностики, прогнозування перебігу захворювань, аналізу медичних зображень та персоналізації лікування. Проте разом із підвищенням точності моделей зростає і складність їхньої внутрішньої структури, що ускладнює розуміння логіки прийняття рішень.

У медичній сфері недостатня прозорість алгоритмів є серйозною проблемою, оскільки будь-яке рішення має бути клінічно обґрунтованим і зрозумілим для лікаря. Це зумовлює необхідність використання підходів Explainable AI, які дозволяють пояснювати результати роботи моделей без втрати їхньої ефективності.

Таким чином, тема застосування методів XAI для підвищення прозорості ухвалення рішень у медичних інформаційних системах є актуальною з наукової, практичної та соціальної точок зору.

1.2 Медичні інформаційні системи та автоматизоване ухвалення рішень

Медичні інформаційні системи являють собою сукупність програмних та апаратних засобів, призначених для автоматизації процесів зберігання, обробки та передачі медичної інформації. Вони охоплюють широкий спектр функцій – від ведення електронних історій хвороб до підтримки складних клінічних процесів і взаємодії між різними підрозділами медичних закладів. Використання МІС дозволяє підвищити якість медичного обслуговування, зменшити кількість помилок і оптимізувати роботу медичного персоналу.

Інтеграція методів штучного інтелекту в медичні інформаційні системи відкриває нові можливості для автоматизованого ухвалення рішень. Алгоритми машинного навчання використовуються для аналізу клінічних даних, прогнозування ризиків ускладнень, виявлення патологій та формування рекомендацій щодо лікування. У більшості випадків такі системи виконують роль інтелектуальної підтримки лікаря, однак їхній вплив на процес ухвалення рішень є суттєвим.

Разом із перевагами автоматизації виникають і нові виклики, пов’язані з надійністю та контрольованістю AI-рішень. Помилки вхідних даних, упередженість навчальних вибірок або некоректна робота моделей можуть призводити до хибних рекомендацій. У зв’язку з цим особливо важливим є забезпечення прозорості та можливості пояснення рішень, що приймаються автоматизованими системами у медицині.

1.3 Проблема «чорної скриньки» в медичних ML-моделях

Більшість сучасних алгоритмів машинного навчання, які застосовуються в медицині, належать до класу складних моделей з високою кількістю параметрів і нелінійних зв’язків. Такі моделі часто забезпечують

високу точність прогнозування, однак їх внутрішня логіка залишається непрозорою для користувача. Це явище отримало назву проблеми «чорної скриньки», коли результат роботи моделі відомий, але причини його отримання не можуть бути чітко пояснені (рисунок 1.1).

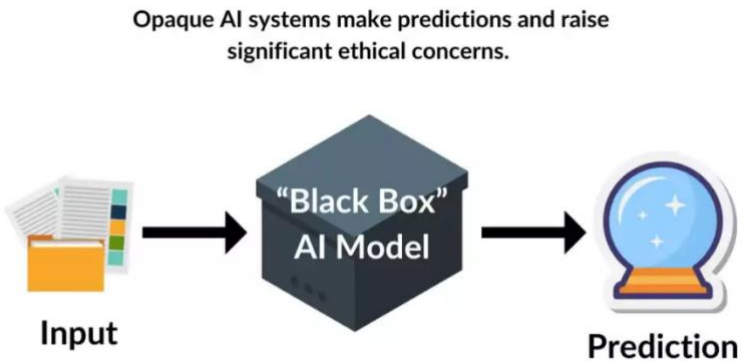

Рисунок 1.1 – Візуалізація моделі Black Box

У медичній практиці відсутність інтерпретації є суттєвим обмеженням, оскільки лікар повинен розуміти, які фактори вплинули на рішення системи. Неможливість пояснення знижує довіру до AI-рішень та ускладнює їх використання в клінічній діяльності. Крім того, непрозорість моделей унеможливлює повноцінну перевірку їх відповідності клінічним протоколам та стандартам лікування.

Проблема «чорної скриньки» також має юридичний та етичний вимір. У разі помилкового рішення складно визначити відповідальність та провести аудит роботи системи. Це створює додаткові бар'єри для сертифікації та впровадження AI-рішень у медичну практику, що підкреслює необхідність розвитку та застосування пояснюваних моделей.

1.4 Вимоги до прозорості та інтерпретованості в медицині

Прозорість і інтерпретованість є ключовими вимогами до використання штучного інтелекту в медичних інформаційних системах.

Лікарі повинні мати можливість аналізувати логіку прийняття рішень AI-системами, щоб оцінити їх клінічну доцільність та відповідність медичним стандартам. Пояснювані моделі сприяють підвищенню довіри до автоматизованих рішень і полегшують їх інтеграцію у практичну діяльність.

Інтерпретованість також є важливою для забезпечення безпеки пацієнтів та контролю якості медичних послуг. Можливість аналізу впливу окремих ознак на результат дозволяє виявляти потенційні помилки, упередженість даних та некоректну поведінку моделей. Це є особливо важливим у системах високого ризику, до яких належать медичні AI-рішення.

Крім того, сучасні регуляторні та етичні вимоги передбачають необхідність підзвітності та можливості пояснення рішень, що приймаються AI-системами. У цьому контексті методи Explainable AI виступають ефективним інструментом, який дозволяє поєднати високу точність моделей із можливістю їхнього пояснення та безпечного використання в медицині.

1.5 Постановка задачі

Метою даної роботи є побудова медичної інформаційної системи з інтеграцією методів штучного інтелекту для автоматизованого аналізу медичних даних та підтримки ухвалення клінічних рішень, а також впровадження методів Explainable AI для підвищення прозорості та інтерпретованості результатів роботи такої системи. Розроблювана система повинна поєднувати функціональність медичної інформаційної платформи з можливостями сучасних AI-алгоритмів.

У межах роботи передбачається аналіз існуючих підходів до побудови медичних інформаційних систем і використання машинного навчання у медицині. Особливу увагу планується приділити дослідженню проблеми «чорної скриньки» та визначенню вимог до прозорості AI-рішень. На основі

цього буде здійснено вибір методів Explainable AI, які дозволяють пояснювати результати роботи моделей машинного навчання.

Подальшим етапом є реалізація програмної частини медичної інформаційної системи, інтеграція AI-моделі та застосування XAI-методів для аналізу її рішень. Результатом роботи має стати медична інформаційна система, яка забезпечує автоматизоване ухвалення рішень із можливістю їх пояснення та практичного використання в клінічній діяльності, що сприятиме підвищенню довіри до AI та якості медичних рішень.

2 ТЕОРЕТИЧНІ ОСНОВИ EXPLAINABLE ARTIFICIAL INTELLIGENCE

2.1 Поняття та роль Explainable Artificial Intelligence

Explainable Artificial Intelligence є напрямом у галузі штучного інтелекту, який зосереджується на розробці методів і підходів для пояснення результатів роботи моделей машинного навчання у зрозумілій для людини формі. Основною метою XAI є забезпечення прозорості та інтерпретованості AI-рішень без суттєвого зниження точності моделей. На відміну від класичних підходів, орієнтованих виключно на оптимізацію метрик якості, Explainable AI робить акцент на можливості аналізу логіки прийняття рішень алгоритмами (рисунок 2.1).

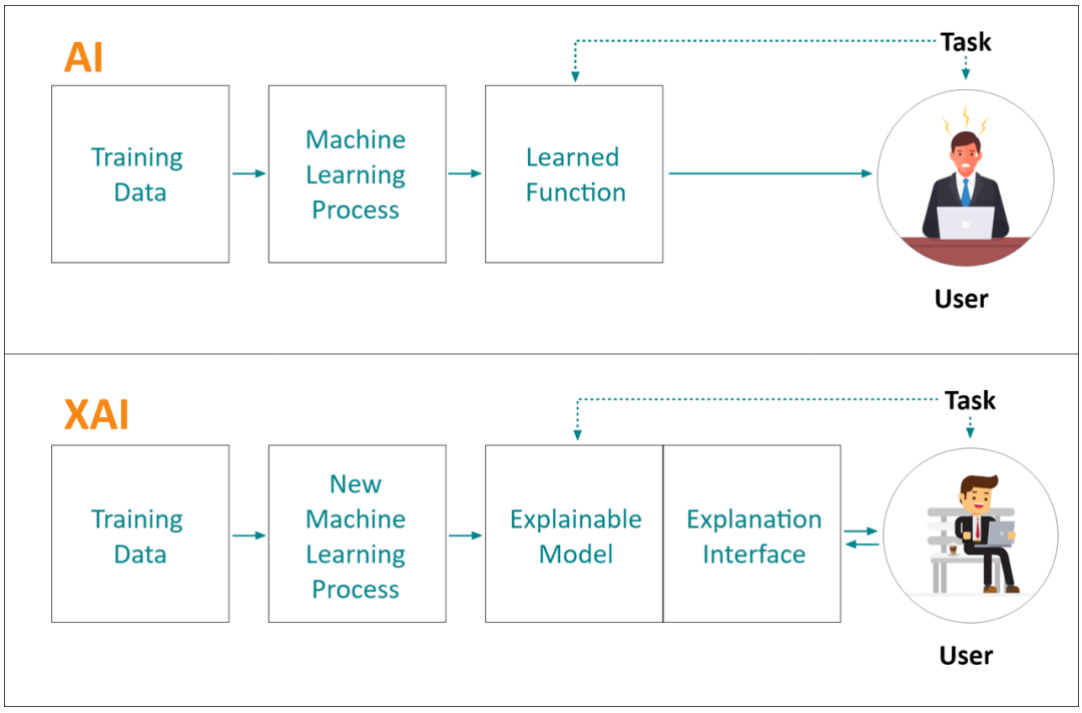

Рисунок 2.1 – AI та XAI

Поява напрямку Explainable AI зумовлена широким впровадженням складних моделей машинного навчання, зокрема глибоких нейронних

мереж і ансамблевих методів, які демонструють високу ефективність, але водночас є важко інтерпретованими. У багатьох прикладних задачах, особливо у критичних сферах, таких як медицина, фінанси чи безпека, відсутність пояснення результатів є неприйнятною. Це призвело до усвідомлення необхідності поєднання точності моделей із можливістю їх пояснення та контролю.

У медичній галузі роль Explainable AI є особливо важливою, оскільки клінічні рішення мають бути не лише правильними, а й обґрунтованими з точки зору медичних знань. Лікар повинен розуміти, які саме фактори вплинули на прогноз або рекомендацію, щоб оцінити їх відповідність клінічній картині пацієнта. Таким чином, XAI виступає ключовим елементом довіри до AI-систем і передумовою їх безпечного використання в медичних інформаційних системах.

2.2 Класифікація методів Explainable AI

Методи Explainable AI можна класифікувати за кількома ознаками, що відображають різні підходи до забезпечення інтерпретованості моделей машинного навчання. Така класифікація дозволяє систематизувати існуючі підходи та спростити вибір відповідного методу залежно від типу моделі, характеру даних і вимог до пояснюваності результатів. Одним із основних підходів до класифікації методів XAI є поділ на внутрішньо інтерпретовані (intrinsic) та post-hoc методи, який ґрунтується на способі отримання пояснення та етапі його формування.

Внутрішньо інтерпретовані моделі мають прозору або частково прозору структуру, яка дозволяє безпосередньо аналізувати логіку прийняття рішень без застосування додаткових засобів пояснення. До таких моделей належать, зокрема, лінійні моделі, дерева рішень та деякі узагальнені адитивні моделі. Їх перевагою є простота інтерпретації та висока зрозумілість для користувача, що є важливим у критичних сферах,

таких як медицина. Водночас такі моделі часто мають обмежену здатність до відображення складних нелінійних залежностей у даних, що може негативно впливати на точність прогнозування.

Post-hoc методи, на відміну від внутрішньо інтерпретованих, застосовуються до вже навчених моделей і не потребують зміни їхньої внутрішньої структури. Вони дозволяють пояснювати результати роботи навіть дуже складних моделей, таких як глибокі нейронні мережі або ансамблеві методи. Post-hoc підходи є більш гнучкими та універсальними, оскільки можуть використовуватися для різних типів моделей і задач. Проте отримані пояснення мають характер апроксимації та не завжди повністю відображають внутрішню логіку роботи алгоритму, що необхідно враховувати при їх використанні в медичних інформаційних системах.

Візуалізація відмінності двух сімейств XAI наведено на рисунку 2.2.

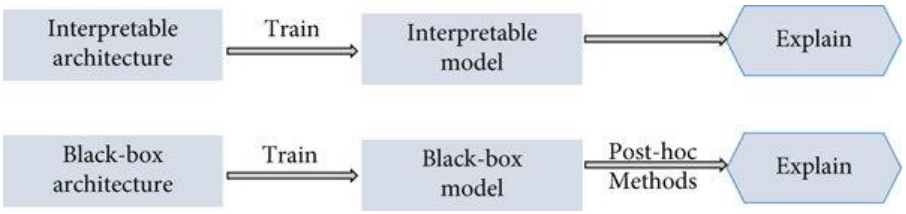

Рисунок 2.2 – Візуалізація відмінності двух сімейств XAI

Іншим важливим критерієм класифікації є поділ методів на локальні та глобальні (рисунок 2.3). Локальні пояснення спрямовані на інтерпретацію окремого прогнозу для конкретного об’єкта, наприклад, пацієнта, тоді як глобальні пояснення описують загальну поведінку моделі на всьому наборі даних. У медичних задачах обидва підходи є важливими, оскільки лікарю необхідно розуміти як загальні закономірності, так і причини конкретного рішення.

Класифікація методів інтерпретованого штучного інтелекту за критерієм їхньої залежності від архітектури базового алгоритму є одним із

найбільш фундаментальних аспектів проектування прозорих систем машинного навчання. У науковій літературі ці методи прийнято поділяти на модельно-агностичні та модельно-специфічні (рисунок 2.4), де основна відмінність полягає в способі взаємодії з внутрішньою структурою моделі.

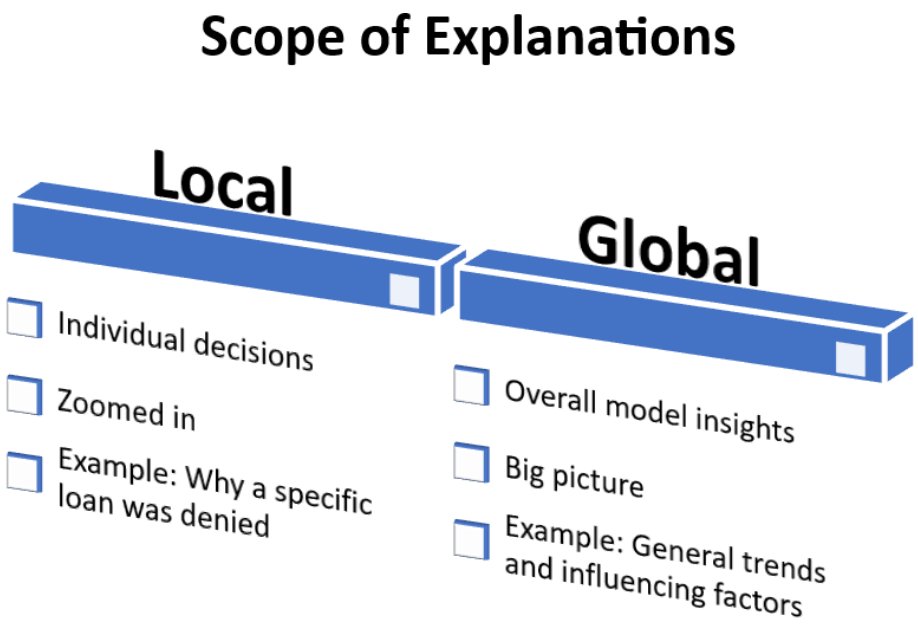

Рисунок 2.3 – Глобальні та локальні пояснення

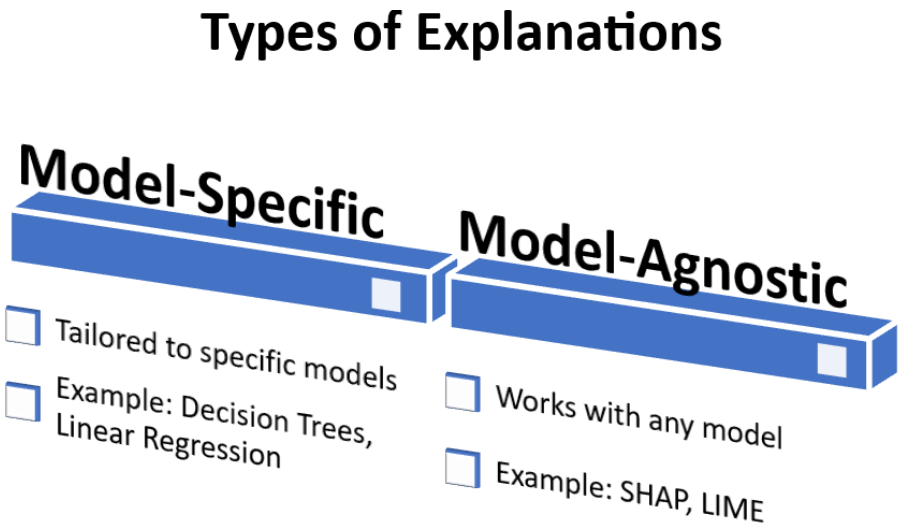

Рисунок 2.4 – Модельно-специфічні та модельно-агностичні пояснення

Модельно-агностичні підходи відзначаються своєю концептуальною універсальністю та високим рівнем адаптивності. Фундаментальна особливість таких методів полягає в тому, що вони функціонують за принципом зовнішнього спостерігача, розглядаючи будь-яку модель як закриту систему або чорну скриньку. Вони аналізують виключно динаміку зміни вихідних прогнозів залежно від варіацій вхідних даних, не вимагаючи доступу до ваг нейронних мереж, структури дерев рішень чи інших внутрішніх параметрів. Така незалежність робить їх незамінними в ситуаціях, коли дослідникові необхідно порівняти ефективність та логіку рішень кількох радикально різних архітектур, наприклад, лінійної моделі та глибокої нейронної мережі, використовуючи при цьому єдиний стандарт пояснень. Крім того, агностичні методи забезпечують стабільність системи інтерпретації в умовах постійного оновлення моделей, оскільки заміна одного алгоритму на інший не потребує перегляду всієї методології пояснення.

На противагу їм, модельно-специфічні методи розробляються з урахуванням унікальних математичних властивостей конкретних типів моделей. Вони базуються на безпосередньому аналізі внутрішніх обчислювальних процесів, що дозволяє їм виявляти складні закономірності значно швидше та точніше. Використовуючи доступ до градієнтів, коефіцієнтів або ієрархічної структури вузлів, ці методи забезпечують вищу обчислювальну ефективність, що особливо критично при роботі з великими масивами даних або складними ансамблями. Наприклад, спеціалізовані алгоритми для деревних моделей здатні за один прохід обчислити точні внески кожної ознаки, тоді як універсальні методи потребували б тисяч ітерацій для отримання наближеного результату. Однак головним обмеженням таких підходів є їхня вузька спеціалізація, адже алгоритм, ідеально адаптований під конкретну архітектуру, виявиться повністю недієздатним при спробі застосувати його до моделі іншого типу.

Розуміння цієї класифікації дозволяє обґрунтовано підходити до вибору методу інтерпретації залежно від поставленої задачі. У сфері медицини, де прогнозування критичних станів вимагає не лише високої точності, а й миттєвого пояснення, поєднання універсальних та специфічних підходів стає запорукою створення надійного цифрового інструментарію. Такий підхід забезпечує гнучкість розробки без втрати глибини аналізу, що зрештою сприяє формуванню довіри з боку лікарів та підвищує безпеку медичних рішень на основі даних.

2.3 Огляд основних методів Explainable Artificial Intelligence

Методи Explainable Artificial Intelligence охоплюють широкий спектр підходів, спрямованих на підвищення прозорості та зрозумілості рішень моделей машинного навчання. Одним із базових напрямів є аналіз важливості ознак, який дозволяє оцінити внесок кожної вхідної змінної у формування прогнозу. Такі підходи широко застосовуються для отримання глобального уявлення про поведінку моделі та визначення ключових факторів, що впливають на результат. У медичних задачах аналіз важливості ознак дозволяє ідентифікувати найбільш значущі клінічні показники, однак цей підхід часто не враховує складні взаємозв’язки між ознаками та може бути чутливим до кореляції вхідних даних.

Partial Dependence Plot є поширеним методом глобальної інтерпретації, який використовується для аналізу середнього впливу окремих ознак на результат роботи моделі. Він дозволяє візуалізувати характер залежності між вхідними змінними та прогнозом, що є корисним для виявлення нелінійних ефектів і порогових значень. У медичних інформаційних системах PDP може застосовуватися для аналізу впливу фізіологічних або лабораторних показників на ймовірність розвитку захворювання. Водночас обмеженням методу є припущення незалежності ознак, яке не завжди виконується для реальних медичних даних.

До локальних методів пояснення належить метод LIME, який є модельно-агностичним підходом і дозволяє інтерпретувати окремі прогнози складних моделей. Метод базується на локальній апроксимації поведінки моделі за допомогою простої інтерпретованої моделі в околі конкретного об’єкта. Це дозволяє визначити, які ознаки мали найбільший вплив на конкретне рішення. Попри свою універсальність, LIME може демонструвати нестабільність пояснень та залежність від параметрів методу, що обмежує його застосування в критичних медичних сценаріях.

Іншим важливим локальним методом є підхід, заснований на контрфактичних поясненнях. Контрфактичні пояснення описують мінімальні зміни у вхідних даних, які призвели б до іншого результату моделі. У медичному контексті це може бути, наприклад, зміна окремих показників пацієнта, яка знижує прогнозований ризик захворювання. Такі пояснення є інтуїтивно зрозумілими для лікарів і пацієнтів, однак їх побудова може бути складною та залежати від коректності обмежень на допустимі зміни ознак.

До групи методів, що аналізують вплив ознак на основі збурень, належать підходи, засновані на перестановці ознак. Суть цих методів полягає в оцінці зміни якості моделі при випадковій зміні значень окремих ознак. Такий підхід дозволяє оцінити глобальну важливість змінних і є модельно-агностичним. У медичних задачах він може використовуватися для попереднього аналізу даних, проте не забезпечує детальних локальних пояснень.

Метод SHAP є одним із найбільш теоретично обґрунтованих підходів до Explainable AI та базується на значеннях Шеплі з теорії кооперативних ігор. У цьому підході кожна ознака розглядається як учасник спільної взаємодії, а прогноз моделі – як результат їхньої спільної дії. SHAP забезпечує справедливий і узгоджений розподіл внеску ознак, що гарантує локальну точність і адитивність пояснень. Метод дозволяє отримувати як

локальні пояснення для окремих прогнозів, так і глобальне уявлення про поведінку моделі в цілому.

Окрім універсального підходу SHAP, існують спеціалізовані методи пояснення для нейронних мереж, зокрема Integrated Gradients та Grad-CAM. Integrated Gradients дозволяє оцінити внесок вхідних ознак у результат моделі шляхом інтегрування градієнтів уздовж шляху від базового стану до поточного входу. Grad-CAM використовується переважно для аналізу моделей, що працюють з медичними зображеннями, та дозволяє візуалізувати області зображення, які найбільше вплинули на рішення моделі. Такі методи є корисними для задач медичної візуалізації, проте мають обмежену універсальність.

На рисунку 2.5 наведено порівняння LIME та SHAP.

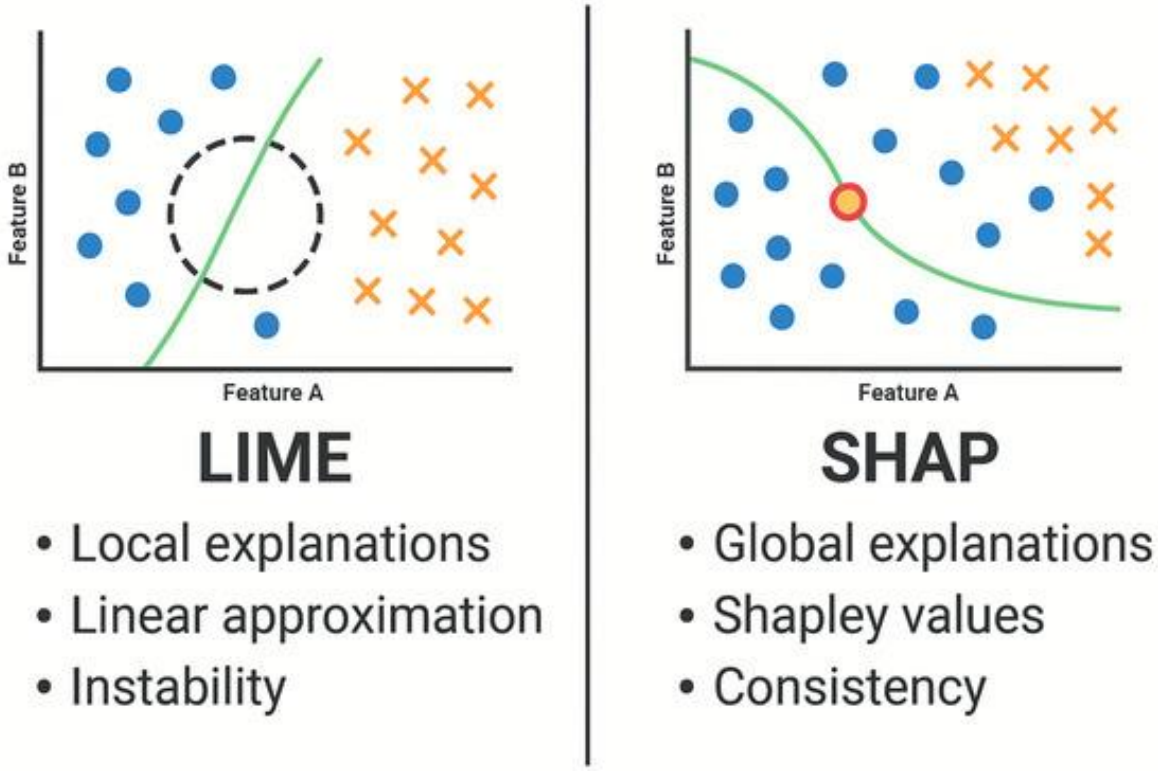

Рисунок 2.5 – Зображення порівняння LIME та SHAP

Таким чином, сучасні методи Explainable AI відрізняються за рівнем інтерпретованості, універсальністю та складністю реалізації. Кожен підхід

має власні переваги та обмеження, що необхідно враховувати при виборі методу для конкретної медичної задачі. Серед розглянутих методів SHAP вирізняється поєднанням теоретичної обґрунтованості, стабільності пояснень та практичної придатності для інтеграції в медичні інформаційні системи.

2.4 Порівняльний аналіз методів Explainable AI для медичних задач

Порівняльний аналіз методів Explainable Artificial Intelligence є необхідним етапом для обґрунтованого вибору підходу, придатного для використання в медичних інформаційних системах. Оскільки медична галузь належить до сфер підвищеного ризику, методи XAI повинні відповідати низці критеріїв, серед яких ключовими є інтерпретованість, точність пояснень, обчислювальна складність та клінічна зрозумілість отриманих результатів.

З точки зору інтерпретованості найбільш простими для сприйняття є методи аналізу важливості ознак та Partial Dependence Plot. Вони дозволяють отримати загальне уявлення про вплив окремих параметрів на поведінку моделі та добре підходять для глобального аналізу. Однак такі методи не завжди дозволяють пояснити конкретне рішення для окремого пацієнта, що є важливим у клінічній практиці. Локальні методи, зокрема LIME, контрфактичні пояснення та SHAP, забезпечують більш детальний аналіз окремих прогнозів і дозволяють лікарю зрозуміти причини конкретного рішення.

Точність пояснень є критично важливим критерієм, оскільки пояснення повинні коректно відображати реальну поведінку моделі. Методи, засновані на спрощених апроксимаціях, такі як LIME, можуть надавати лише наближені пояснення, які залежать від локальних припущень і параметрів методу. У той же час SHAP, завдяки використанню значень Шеплі, забезпечує теоретично обґрунтований і узгоджений розподіл внеску

ознак, що підвищує достовірність пояснень. Методи, засновані на перестановці ознак або PDP, не завжди враховують складні взаємодії між змінними, що може знижувати точність інтерпретації в умовах реальних медичних даних.

Обчислювальна складність методів XAI є важливим фактором при інтеграції в медичні інформаційні системи, особливо у випадках, коли пояснення необхідні в режимі, наближеному до реального часу. Прості методи, такі як аналіз важливості ознак або PDP, є відносно маловитратними з точки зору обчислень. LIME та контрфактичні методи потребують додаткових обчислень для генерації локальних варіацій даних, що може ускладнювати їх масштабування. SHAP у базовій формі є обчислювально складним, однак наявність оптимізованих реалізацій для деревоподібних і ансамблевих моделей робить його практично придатним для використання у медичних системах.

Окрему увагу слід приділити клінічній зрозумілості пояснень, тобто можливості інтерпретувати результати XAI з позиції медичних знань. Візуалізації PDP та глобальної важливості ознак є корисними для аналізу загальних закономірностей, проте не завжди дозволяють пояснити індивідуальний клінічний випадок. Контрфактичні пояснення є інтуїтивно зрозумілими, оскільки демонструють можливі зміни показників, які впливають на результат. SHAP забезпечує баланс між деталізацією та наочністю, дозволяючи представити пояснення у вигляді зрозумілих графіків і числових значень, які можуть бути безпосередньо використані лікарем при аналізі клінічної ситуації.

А це значить, що порівняльний аналіз показує, що жоден із методів Explainable AI не є універсальним для всіх задач. Вибір конкретного підходу залежить від типу моделі, характеру медичних даних та вимог до пояснюваності. Водночас метод SHAP вирізняється поєднанням високої точності пояснень, універсальності, підтримки локальних і глобальних інтерпретацій та придатності для використання в медичних інформаційних

системах, що робить його доцільним вибором для подальшої реалізації в межах даної роботи.

2.5 Обґрунтування вибору SHAP як основного методу Explainable AI

Вибір методу Explainable Artificial Intelligence для використання в медичних інформаційних системах повинен ґрунтуватися на поєднанні теоретичної обґрунтованості, практичної ефективності та відповідності специфічним вимогам медичної галузі. Проведений аналіз основних підходів до XAI показує, що багато методів мають обмеження, пов’язані з нестабільністю пояснень, низькою узгодженістю або складністю інтерпретації результатів у клінічному контексті. У цьому зв’язку метод SHAP є доцільним вибором для використання в межах даної роботи.

Однією з ключових переваг SHAP є його теоретичне підґрунтя, засноване на значеннях Шеплі з теорії кооперативних ігор. Такий підхід забезпечує справедливий і узгоджений розподіл внеску кожної ознаки у результат моделі, що гарантує локальну точність пояснень. Це означає, що сума внесків усіх ознак дорівнює різниці між прогнозом моделі для конкретного об’єкта та середнім прогнозом, що є важливою властивістю для коректної інтерпретації результатів.

SHAP добре узгоджується з сучасними моделями машинного навчання, які широко застосовуються в медицині, зокрема з ансамблевими методами та моделями градієнтного бустингу. Наявність оптимізованих алгоритмів дозволяє ефективно обчислювати значення SHAP навіть для складних моделей без значних втрат продуктивності. Це робить метод придатним для інтеграції в медичні інформаційні системи, які працюють з великими обсягами даних.

Важливою перевагою SHAP є можливість отримання як локальних, так і глобальних пояснень. Локальні пояснення дозволяють аналізувати причини конкретного прогнозу для окремого пацієнта, що є критично

важливим у клінічній практиці. Глобальні пояснення, у свою чергу, дають змогу дослідити загальну поведінку моделі, визначити найбільш значущі ознаки та перевірити відповідність роботи алгоритму медичним знанням і клінічним стандартам.

Окрему увагу слід приділити клінічній зрозумілості результатів, отриманих за допомогою SHAP. Візуалізації та числові значення, що генеруються цим методом, можуть бути представлені у формі, зручній для сприйняття лікарями, що сприяє підвищенню довіри до AI-рішень. Це особливо важливо в умовах використання автоматизованих систем ухвалення рішень, де лікар несе відповідальність за кінцевий результат лікування.

Отже, з урахуванням теоретичних переваг, універсальності, підтримки різних типів моделей та високої клінічної придатності, метод SHAP обрано як основний інструмент Explainable AI у межах даної роботи. Його використання дозволяє поєднати високу точність моделей машинного навчання з можливістю їх пояснення та безпечного впровадження у медичні інформаційні системи.

3 АНАЛІЗ ДАНИХ ТА АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

3.1 Вибір та опис датасету

Для розв’язання поставленої задачі було обрано відкритий медичний датасет Heart Failure Prediction (рисунок 3.1), який широко використовується в наукових дослідженнях для аналізу та прогнозування серцевої недостатності. Датасет містить клінічні дані пацієнтів із діагностованою серцевою недостатністю та дозволяє проводити дослідження з використанням методів машинного навчання і пояснюваного штучного інтелекту.

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	1.1	137	1	0	11	1
50	1	168	0	38	1	276000	1.1	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0
82	1	379	0	50	0	47000	1.3	136	1	0	13	1
87	1	149	0	38	0	262000	0.9	140	1	0	14	1
45	0	582	0	14	0	166000	0.8	127	1	0	14	1
70	1	125	0	25	1	237000	1	140	0	0	15	1
48	1	582	1	55	0	87000	1.9	121	0	0	15	1
65	1	52	0	25	1	276000	1.3	137	0	0	16	0
65	1	128	1	30	1	297000	1.6	136	0	0	20	1
68	1	220	0	35	1	289000	0.9	140	1	1	20	1
53	0	63	1	60	0	368000	0.8	135	1	0	22	0
75	0	582	1	30	1	263358.03	1.83	134	0	0	23	1
80	0	148	1	38	0	149000	1.9	144	1	1	23	1
95	1	112	0	40	1	196000	1	138	0	0	24	1
70	0	122	1	45	1	284000	1.3	136	1	1	26	1
58	1	60	0	38	0	153000	5.8	134	1	0	26	1
82	0	70	1	30	0	200000	1.2	132	1	1	26	1
94	0	582	1	38	1	263358.03	1.83	134	1	0	27	1
85	0	23	0	45	0	360000	3	132	1	0	28	1

Рисунок 3.1 – Датасет у Excel

Обраний датасет включає набір числових і бінарних ознак, які описують клінічний стан пацієнтів. До них належать демографічні

характеристики, такі як вік і стать, результати лабораторних аналізів, зокрема рівень креатиніну в сироватці крові та кількість тромбоцитів, а також показники, що відображають наявність супутніх захворювань. Цільова змінна відображає інформацію про летальний результат протягом періоду спостереження, що робить датасет придатним для задачі бінарної класифікації.

Age – вік пацієнта, вимірюється у роках і є числовою безперервною ознакою, що відображає демографічну характеристику пацієнта.

Anaemia – бінарна ознака, що вказує на наявність або відсутність анемії; значення 1 відповідає наявності анемії, значення 0 – її відсутності.

Creatinine_phosphokinase – рівень креатинінфосфокінази в крові, вимірюється в одиницях U/L та є числовою ознакою, що відображає стан м’язової та серцевої тканини.

Diabetes – бінарна ознака, яка вказує на наявність цукрового діабету у пацієнта; значення 1 означає наявність захворювання, 0 – його відсутність.

Ejection_fraction – фракція викиду лівого шлуночка серця, виражена у відсотках; є одним із ключових показників функціонального стану серця.

High_blood_pressure – бінарна ознака, що характеризує наявність артеріальної гіпертензії; значення 1 відповідає підвищеному артеріальному тиску, 0 – нормальному.

Platelets – кількість тромбоцитів у крові, вимірюється у кількості клітин на мікролітр; є числовою безперервною ознакою.

Serum_creatinine – рівень креатиніну в сироватці крові, вимірюється в мг/дл; показник, що відображає функцію нирок і має важливе клінічне значення.

Serum_sodium – рівень натрію в сироватці крові, вимірюється в мЕкв/л; використовується для оцінки електролітного балансу пацієнта.

Sex – бінарна ознака, що відображає стать пацієнта; значення 1 відповідає чоловічій статі, 0 – жіночій.

Smoking – бінарна ознака, яка вказує на факт куріння; значення 1 означає, що пацієнт курить, 0 – що не курить.

Time – тривалість періоду спостереження за пацієнтом, вимірюється у днях; характеризує час до настання події або завершення спостереження.

DEATH_EVENT – цільова бінарна змінна, що відображає факт летального результату; значення 1 означає настання смерті пацієнта протягом періоду спостереження, 0 – її відсутність.

Вибір саме цього датасету обґрунтований кількома факторами. По-перше, він містить реальні клінічні дані, що відповідають медичній тематиці роботи. По-друге, структура датасету є відносно простою, що дозволяє застосовувати різні алгоритми машинного навчання та методи Explainable AI без необхідності складної попередньої обробки. По-третє, дані є добре задокументованими та широко використовуються у дослідженнях, що полегшує порівняння отриманих результатів з існуючими роботами.

Аналіз якості даних показує, що датасет має обмежений обсяг, що є типовим для медичних досліджень, однак містить достатньо інформативні ознаки для побудови моделей машинного навчання. У наборі даних відсутні значні пропуски значень, що спрощує процес підготовки даних. Водночас спостерігається певний дисбаланс між класами, що необхідно враховувати при навчанні та оцінюванні моделей. Також важливо зазначити, що деякі ознаки можуть бути корельованими між собою, що потребує додаткового аналізу та врахування при інтерпретації результатів.

Таким чином, датасет Heart Failure Predictionє доцільним вибором для реалізації медичної інформаційної системи з використанням алгоритмів машинного навчання та методів Explainable AI. Його використання дозволяє не лише побудувати модель прогнозування, а й дослідити вплив окремих клінічних параметрів на результат, що є важливим з точки зору практичного застосування AI у медицині.

3.2 Постановка медичної задачі

У межах даної роботи розглядається задача автоматизованої підтримки клінічних рішень на основі аналізу медичних даних пацієнтів із серцевою недостатністю. Серцева недостатність є одним із найбільш поширених і небезпечних серцево-судинних захворювань, яке характеризується високим рівнем смертності та потребує своєчасної діагностики і прогнозування перебігу хвороби. Використання методів машинного навчання дозволяє аналізувати сукупність клінічних показників і формувати прогнози щодо ризику несприятливих наслідків.

З формальної точки зору поставлена медична задача належить до класу задач бінарної класифікації. Метою є прогнозування настання летального результату у пацієнтів із серцевою недостатністю на основі їхніх клінічних та лабораторних показників. Для кожного пацієнта необхідно визначити клас, що відповідає наявності або відсутності летального випадку протягом визначеного періоду спостереження. Така постановка задачі є типовою для медичних інформаційних систем, що використовують AI для оцінювання ризиків та підтримки клінічних рішень.

Вхідними параметрами задачі є набір клінічних ознак, що характеризують стан пацієнта, зокрема демографічні показники, результати лабораторних досліджень та інформація про супутні захворювання. Ці параметри формують вектор ознак, який використовується моделлю машинного навчання для формування прогнозу. Вихідним параметром є бінарна змінна, що відображає факт летального результату. Важливою особливістю задачі є необхідність не лише отримати точний прогноз, а й забезпечити можливість пояснення впливу кожної ознаки на результат, що є критично важливим у медичному контексті та безпосередньо пов’язане з використанням методів Explainable AI.

3.3 Огляд популярних алгоритмів машинного навчання

Для розв’язання задач класифікації в медичних інформаційних системах широко застосовуються різні алгоритми машинного навчання, які відрізняються за рівнем складності, інтерпретованості та здатності працювати з табличними даними. Основні підходи включають лінійні моделі (логістична регресія), алгоритми на основі дерев (дерева рішень, випадковийліс) та методи бустингу (XGBoost, LightGBM)

Логістична регресія (рисунок 3.2) часто використовується в медицині завдяки своїй простоті та інтерпретованості. Вона дозволяє оцінити вплив окремих ознак на ймовірність настання події, проте має обмеження у випадках, коли залежності між змінними є нелінійними.

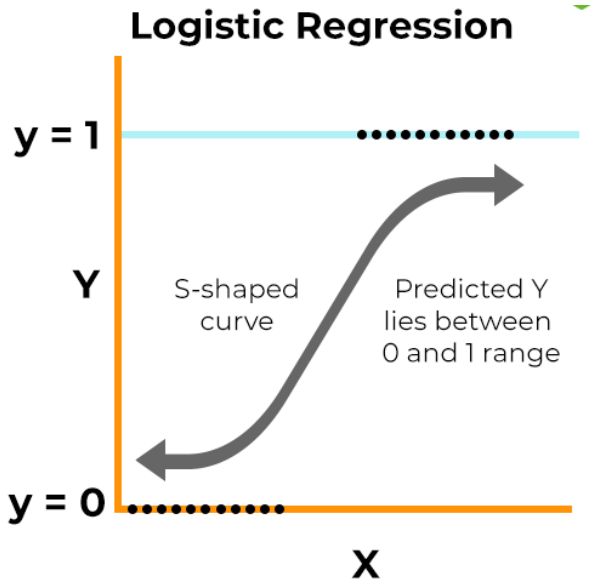

Рисунок 3.2 – Зображення логістичної регресії

Дерева рішень (рисунок 3.3) забезпечують більш гнучке моделювання та є інтуїтивно зрозумілими, однак схильні до перенавчання, особливо при роботі з невеликими датасетами.

Ансамблеві методи, такі як Random Forest, дозволяють підвищити точність прогнозування за рахунок об’єднання результатів декількох

моделей. Вони добре працюють з табличними даними та є стійкими до шуму, однак їх інтерпретованість є обмеженою. Окрему групу становлять алгоритми градієнтного бустингу, які послідовно навчають слабкі моделі для зменшення похибки. Такі підходи, зокрема XGBoost, LightGBM та CatBoost, демонструють високу ефективність у задачах медичної класифікації та прогнозування.

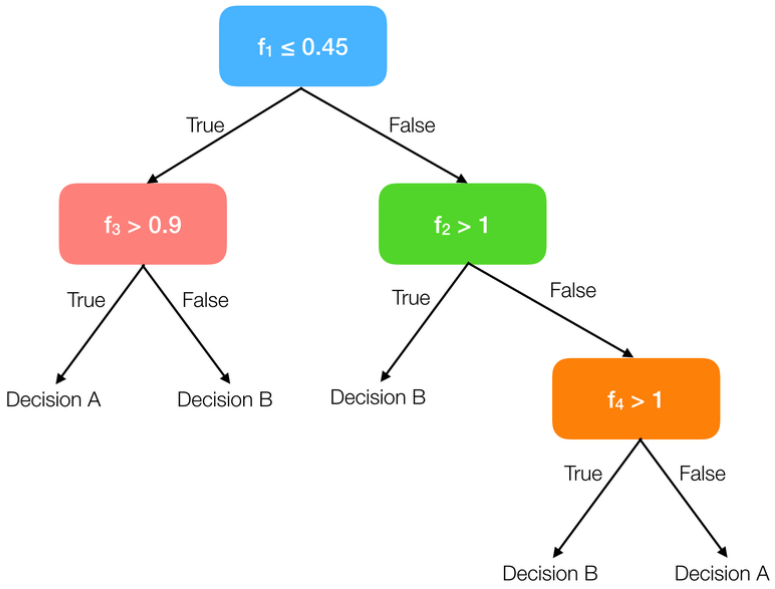

Рисунок 3.3 – Приклад дерева рішень

Таким чином, вибір алгоритму машинного навчання для медичних задач повинен враховувати не лише точність моделі, а й стабільність її роботи, здатність працювати з табличними клінічними даними та можливість подальшої інтерпретації результатів за допомогою методів Explainable AI.

3.4 Обґрунтування вибору алгоритму XGBoost

Для реалізації медичної інформаційної системи в межах даної роботи обрано алгоритм XGBoost, який належить до класу ансамблевих методів

градієнтного бустингу над деревами рішень. Однією з ключових причин вибору цього алгоритму є його висока ефективність при роботі з табличними медичними даними. XGBoost здатний автоматично враховувати нелінійні залежності та взаємодії між ознаками, що є важливим у задачах аналізу клінічних показників.

Алгоритм XGBoost (рисунок 3.4) демонструє високу стабільність та точність прогнозування навіть на відносно невеликих і зашумлених датасетах, що є типовим для медичних даних. Завдяки вбудованим механізмам регуляризації та контролю складності моделей знижується ризик перенавчання, що підвищує надійність отриманих результатів. Крім того, XGBoost дозволяє ефективно працювати з різними типами ознак без складної попередньої обробки даних.

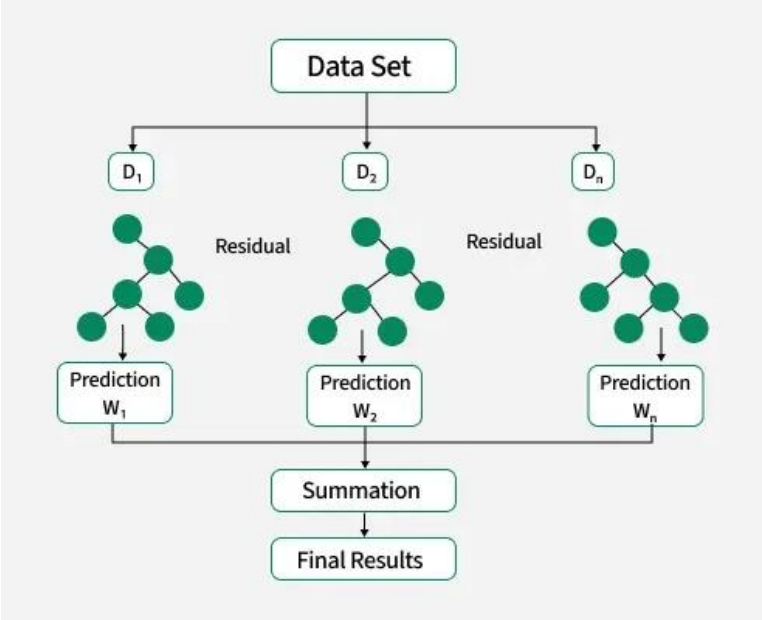

Рисунок 3.4 – Зразок XGBoost моделі

Важливою перевагою XGBoost є його сумісність з методами Explainable AI, зокрема з методом SHAP. Для моделей градієнтного бустингу існують оптимізовані реалізації SHAP, які дозволяють швидко та точно обчислювати внесок кожної ознаки у результат моделі. Це забезпечує

можливість отримання як локальних пояснень для окремих прогнозів, так і глобального аналізу поведінки моделі в цілому.

Отже, вибір алгоритму XGBoost є обґрунтованим з точки зору точності, стабільності та можливості інтеграції з методами Explainable AI. Поєднання XGBoost і SHAP дозволяє реалізувати медичну інформаційну систему, яка не лише формує якісні прогнози, а й забезпечує прозорість та інтерпретованість ухвалюваних рішень, що є критично важливим у медичній практиці.

4 ПРАКТИЧНА РЕАЛІЗАЦІЯ ПРОЄКТУ

4.1 Архітектура програмного рішення

У межах даної роботи буде спроєктовано програмне рішення у вигляді медичної інформаційної системи з підтримкою методів машинного навчання та Explainable AI, призначеної для прогнозування ризику летального результату у пацієнтів із серцевою недостатністю. Архітектура системи передбачатиме модульну структуру, що дозволяє відокремити етапи обробки даних, навчання моделі, інтерпретації результатів та взаємодії з користувачем.

Модуль машинного навчання буде реалізовано з використанням алгоритму XGBoost для задачі бінарної класифікації. Передбачається застосування бібліотек pandas та NumPy для підготовки та обробки табличних медичних даних, а також scikit-learn для розділення вибірки на навчальну та тестову і для обчислення метрик якості моделі. Навчена модель буде використовуватись для формування прогнозів як на тестових даних, так і для нових клінічних записів, введених користувачем.

Для забезпечення прозорості ухвалення рішень у систему буде інтегровано модуль Explainable AI на основі бібліотеки SHAP. Планується використання TreeExplainer, оптимізованого для моделей градієнтного бустингу, що дозволить обчислювати внесок кожної ознаки у прогноз моделі. Це забезпечить можливість отримання локальних пояснень для окремих пацієнтів, а також глобального аналізу важливості ознак і взаємодій між ними.

Крім того, передбачається реалізація додаткових методів інтерпретації, зокрема Partial Dependence Plots, для аналізу середнього впливу окремих ознак на результат прогнозування. Для побудови відповідних візуалізацій планується використання засобів бібліотеки scikit-learn разом із matplotlib.

Користувацький інтерфейс системи буде реалізовано з використанням фреймворку Streamlit, що дозволить створити інтерактивний вебзастосунок для введення клінічних даних, перегляду результатів прогнозування, метрик якості моделі та XAI-візуалізацій. Для оптимізації продуктивності системи передбачається використання механізмів кешування обчислень.

Таким чином, запропонована архітектура дозволить реалізувати повний цикл функціонування медичної інформаційної системи з інтегрованими методами машинного навчання та Explainable AI, забезпечуючи поєднання точності прогнозування з прозорістю та інтерпретованістю AI-рішень.

4.2 Реалізація прогнозування для пацієнта та локального XAI-аналізу

У програмному рішенні прогнозування для конкретного пацієнта буде реалізовано у вигляді окремої вкладки вебінтерфейсу, створеного з використанням фреймворку Streamlit. Для збору клінічних показників застосовуються стандартні елементи керування, зокрема number_input, selectbox та checkbox, що забезпечує контроль допустимих значень і відповідність типів даних ознакам навчального датасету.

Введені користувачем значення (рисунок 4.1) програмно формуються у словник Python, де ключі відповідають назвам ознак датасету. Далі цей словник перетворюється у об’єкт pandas DataFrame з одним рядком, що забезпечує повну сумісність із навченою моделлю XGBoost. Такий підхід дозволяє уникнути додаткових етапів попередньої обробки та мінімізує ризик помилок під час формування вхідних даних.

Прогнозування здійснюється за допомогою методу predict_proba навченої моделі XGBoost, що дозволяє отримати ймовірність належності пацієнта до класу летального результату. Отримане значення використовується як основний кількісний показник ризику. Додатково у коді реалізовано логіку порогової класифікації, яка на основі ймовірності

формує категоріальний рівень ризику. Для підвищення наочності результату застосовується графічна візуалізація у вигляді горизонтальної шкали ризику, побудованої за допомогою matplotlib (рисунок 4.2).

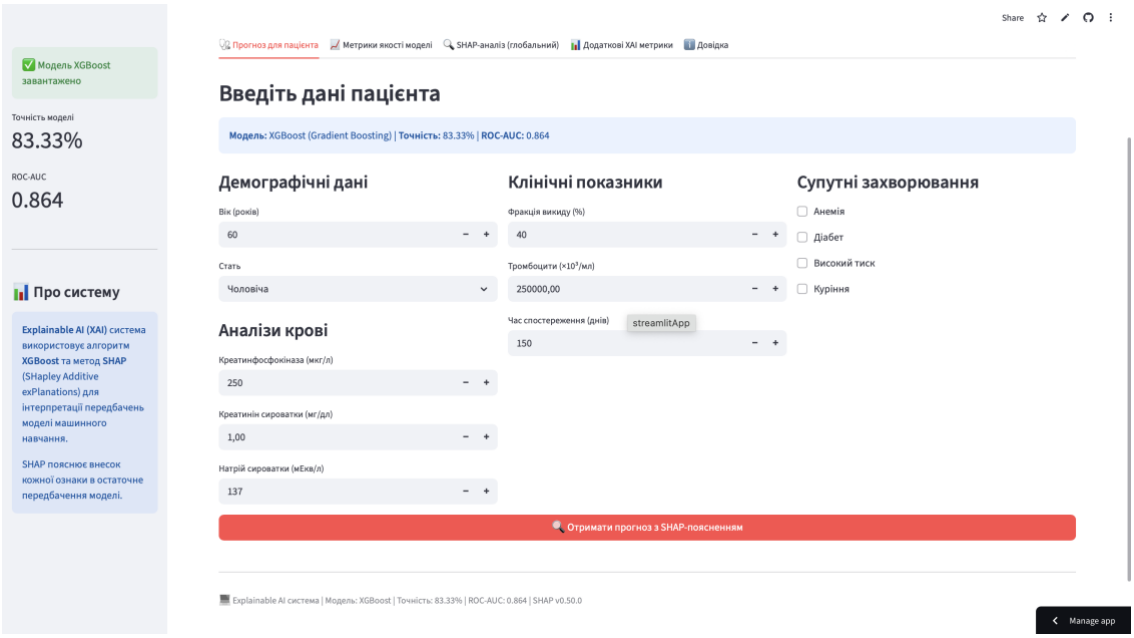

Рисунок 4.1 – Input даних пацієнта

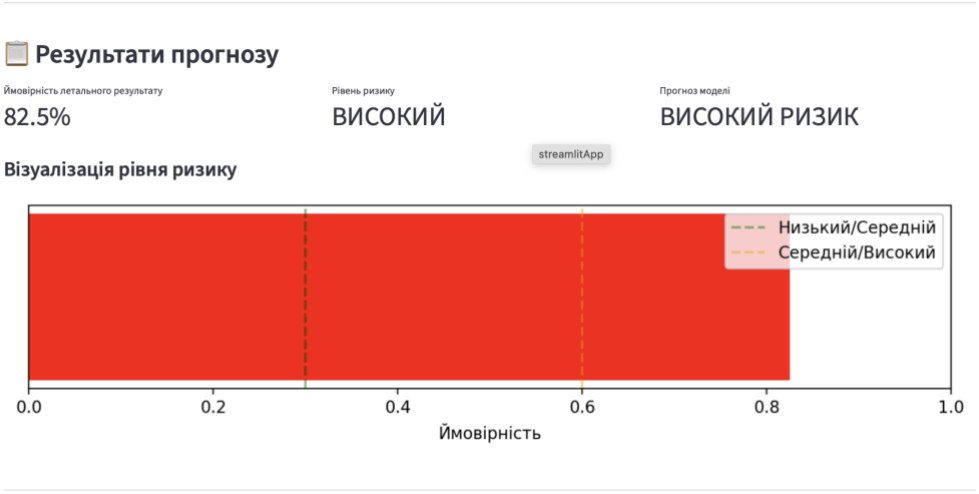

Рисунок 4.2 – Візуалізація ризику

Для пояснення конкретного рішення моделі використовується локальний Explainable AI-аналіз із застосуванням бібліотеки SHAP. У коді створюється об’єкт TreeExplainer, оптимізований для моделей градієнтного бустингу, що дозволяє ефективно обчислювати SHAP-значення для одного спостереження. SHAP-значення розраховуються безпосередньо для DataFrame з даними пацієнта, що забезпечує точне локальне пояснення прогнозу.

Отримані SHAP-значення візуалізуються за допомогою waterfall-графіка (рисунок 4.3), який будується з використанням функції shap.plots.waterfall . Даний тип візуалізації дозволяє послідовно відобразити вплив кожної ознаки на фінальне рішення моделі, починаючи з базового значення і завершуючи кінцевим прогнозом. Додатково SHAP-значення зберігаються у табличній формі, що дозволяє відсортувати ознаки за силою їх впливу та відобразити їх у зручному для аналізу вигляді.

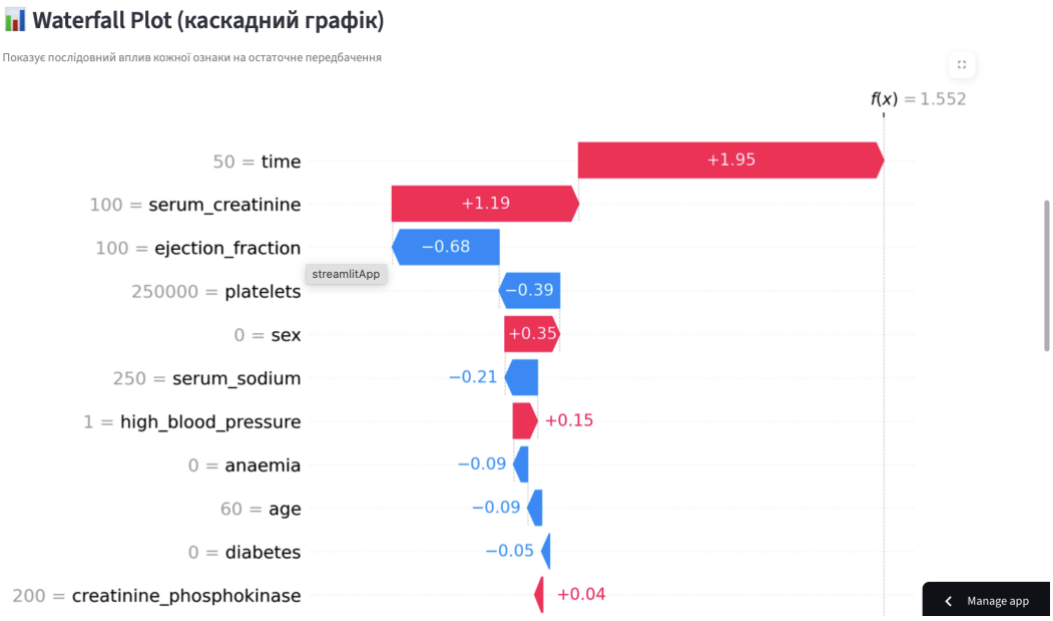

Рисунок 4.3 – Зображення як модель приймає рішення

Контрфактичні пояснення у системі реалізуються шляхом аналізу отриманих SHAP-значень (рисунок 4.4). У коді визначається перелік

незмінних ознак, які не можуть бути скориговані в клінічній практиці. Далі відбираються змінювані ознаки з найбільшим позитивним внеском у ризик летального результату. На основі цих ознак програмно формуються текстові рекомендації, що демонструють, які параметри потенційно можуть бути змінені для зниження ризику, без безпосереднього втручання у роботу моделі.

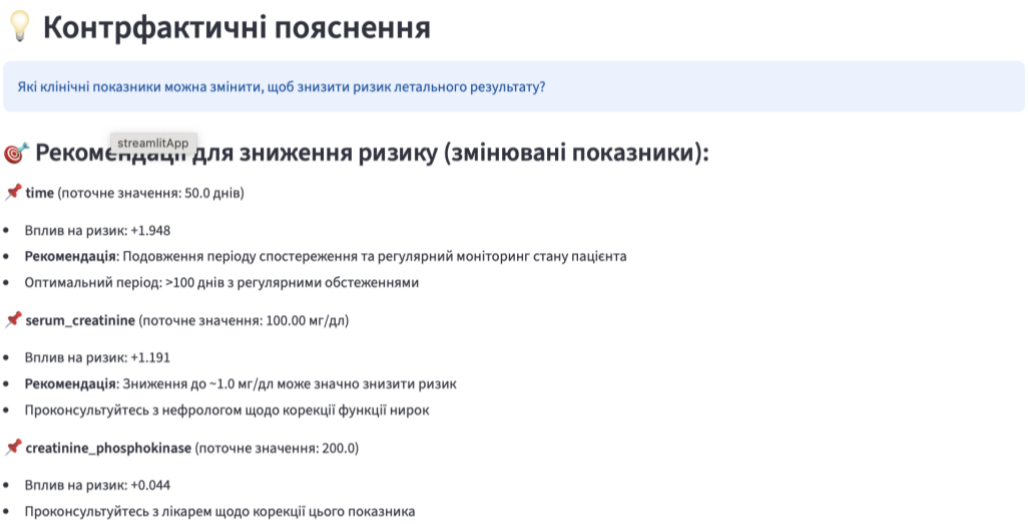

Рисунок 4.4 – Контрфактичні пояснення

А це значить, що реалізація першої вкладки системи базується на прямій інтеграції прогнозної моделі XGBoost та локального SHAP-аналізу. Такий підхід дозволяє не лише отримати числовий результат прогнозування, а й детально обґрунтувати кожне рішення моделі, що є критично важливим для використання системи у медичному контексті.

4.3 Реалізація оцінювання якості моделі машинного навчання

Для оцінювання надійності та стабільності роботи моделі машинного навчання у програмному рішенні буде реалізовано окрему вкладку, присвячену аналізу метрик якості класифікації. Даний функціональний блок дозволяє кількісно оцінити ефективність моделі XGBoost на тестовій

вибірці та є необхідним елементом для обґрунтування можливості її використання в медичній інформаційній системі.

Після навчання моделі тестова вибірка формується за допомогою функції train_test_split із параметром stratify, що забезпечує збереження співвідношення класів. На основі передбачень моделі програмно обчислюються основні метрики класифікації, зокрема accuracy, precision, recall та F1-score, з використанням функцій бібліотеки scikit-learn. Отримані значення зберігаються у структурі даних та використовуються для подальшої візуалізації у користувацькому інтерфейсі (рисунок 4.5).

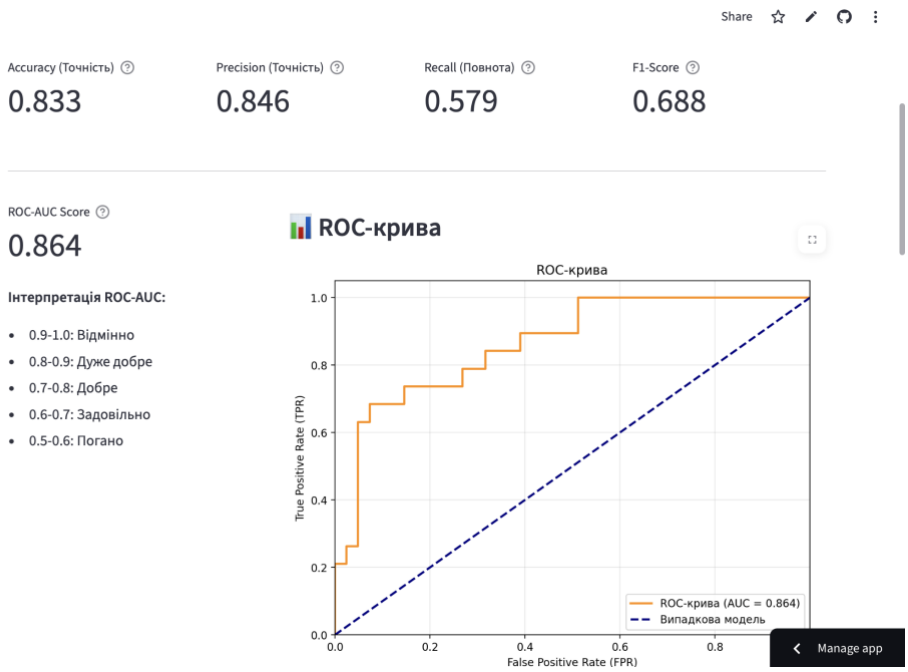

Рисунок 4.5 – Метрики оцінки якості моделі

Окрему увагу приділено обчисленню метрики ROC-AUC, яка характеризує здатність моделі розрізняти класи незалежно від порогу класифікації. Для цього використовується ймовірнісний вихід моделі, отриманий за допомогою методу predict_proba. ROC-крива будується на основі значень False Positive Rate та True Positive Rate, що дозволяє наочно оцінити якість моделі при різних порогових значеннях.

Для детального аналізу помилок класифікації у системі реалізовано побудову матриці помилок. Матриця формується з використанням функції confusion_matrix та візуалізується за допомогою бібліотеки seaborn у вигляді теплової карти. Додатково у коді виконується розклад матриці на складові TN, FP, FN та TP, що дозволяє обчислити похідні метрики, зокрема специфічність та негативну прогностичну цінність, які є важливими для медичних задач (рисунок 4.6).

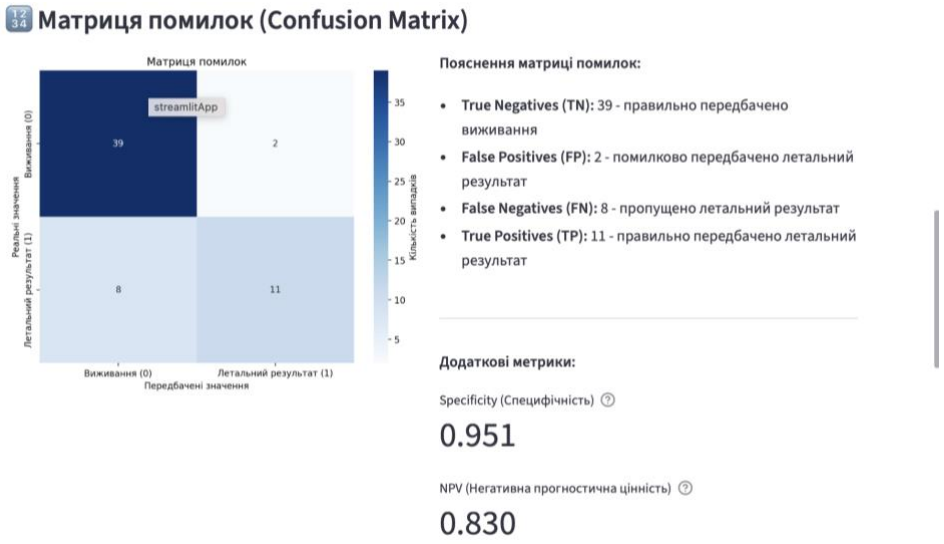

Рисунок 4.6 – Матриця помилок

Для оцінювання поведінки моделі на незбалансованих даних реалізовано побудову Precision–Recall кривої (рисунок 4.7). Даний графік дозволяє проаналізувати компроміс між точністю та повнотою при різних порогах класифікації і є більш інформативним у випадках, коли один із класів представлений меншою кількістю спостережень.

Додатково у вкладці формується детальний звіт класифікації з використанням функції classification_report, який містить значення precision, recall та F1-score для кожного класу окремо (рисунок 4.8). Результати подаються у табличному вигляді, що спрощує їх аналіз та подальше документування.

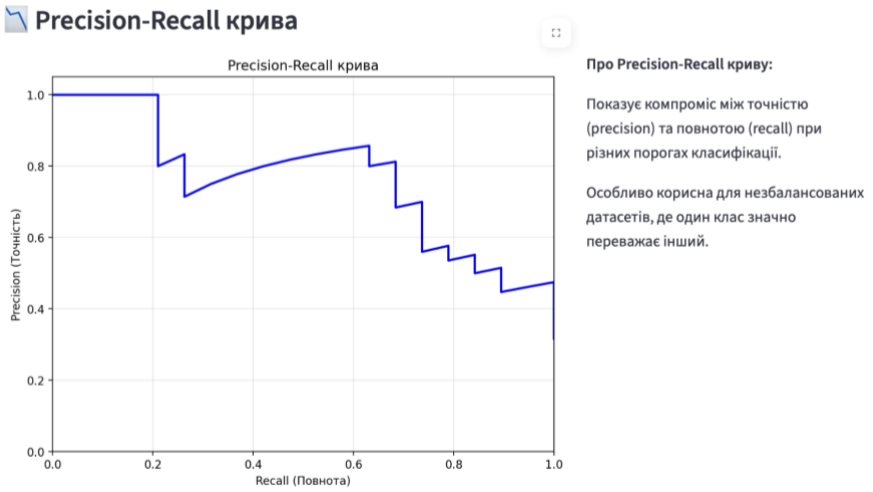

Рисунок 4.7 – Precision-Recall крива

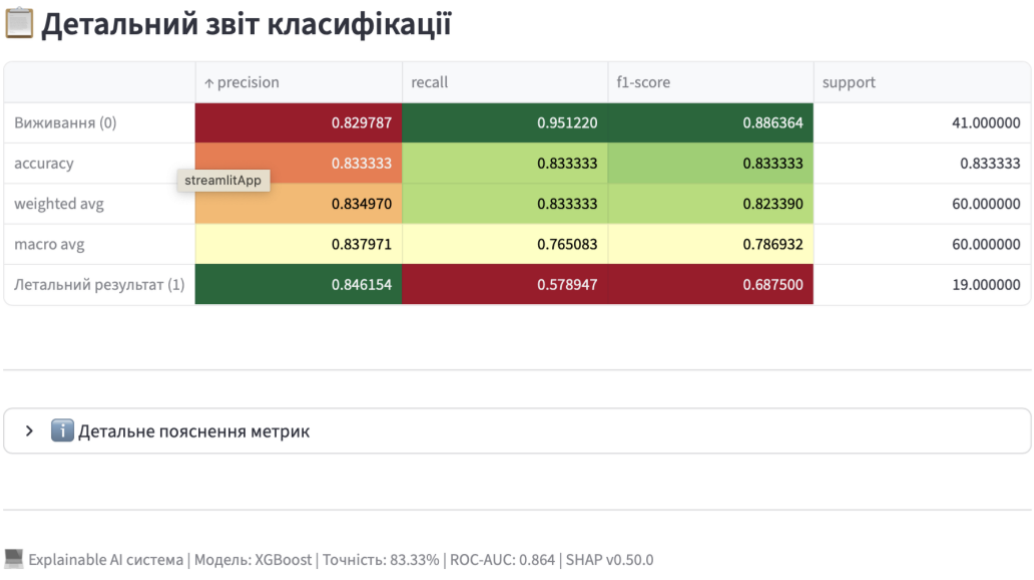

Рисунок 4.8 – Звіт характеристик

Таким чином, другий таб системи реалізує повний набір інструментів для кількісного оцінювання якості моделі XGBoost. Використання стандартних метрик машинного навчання у поєднанні з наочними візуалізаціями дозволяє обґрунтовано оцінити ефективність моделі та підтвердити доцільність її застосування у медичній інформаційній системі.

4.4 Реалізація глобального SHAP-аналізу моделі

Для аналізу загальної поведінки моделі машинного навчання та визначення найбільш впливових клінічних ознак у програмному рішенні буде реалізовано окрему вкладку, присвячену глобальному Explainable AI-аналізу. Даний таб дозволяє оцінити, які фактори в середньому найбільше впливають на рішення моделі XGBoost на всій тестовій вибірці.

Глобальний SHAP-аналіз реалізується шляхом обчислення SHAP-значень для всіх спостережень тестової вибірки з використанням об’єкта TreeExplainer. Для цього у коді передається масив ознак тестових даних, що дозволяє отримати матрицю SHAP-значень, де кожен рядок відповідає окремому пацієнту, а кожен стовпчик – окремій ознаці.

Основним елементом візуалізації глобального аналізу є SHAP Summary Plot (рисунок 4.9). Даний графік будується за допомогою функції shap.summary_plot і дозволяє одночасно відобразити важливість ознак та характер їх впливу на прогноз моделі. Положення точок по горизонтальній осі відображає напрямок і силу впливу ознаки, тоді як колір точки відповідає фактичному значенню цієї ознаки. Такий підхід дозволяє виявити нелінійні залежності та потенційні порогові ефекти в роботі моделі.

Для кількісної оцінки важливості ознак у системі додатково реалізовано побудову SHAP Bar Plot (рисунок 4.10). У коді обчислюється середнє абсолютне значення SHAP для кожної ознаки, що дозволяє ранжувати параметри за ступенем їх впливу на результати моделі. Отримані значення візуалізуються у вигляді горизонтальної гістограми, яка наочно демонструє відносну важливість усіх клінічних показників.

Застосування глобального SHAP-аналізу дозволяє перевірити, чи відповідає логіка роботи моделі медичним знанням, зокрема чи враховуються клінічно значущі показники як основні фактори ризику.

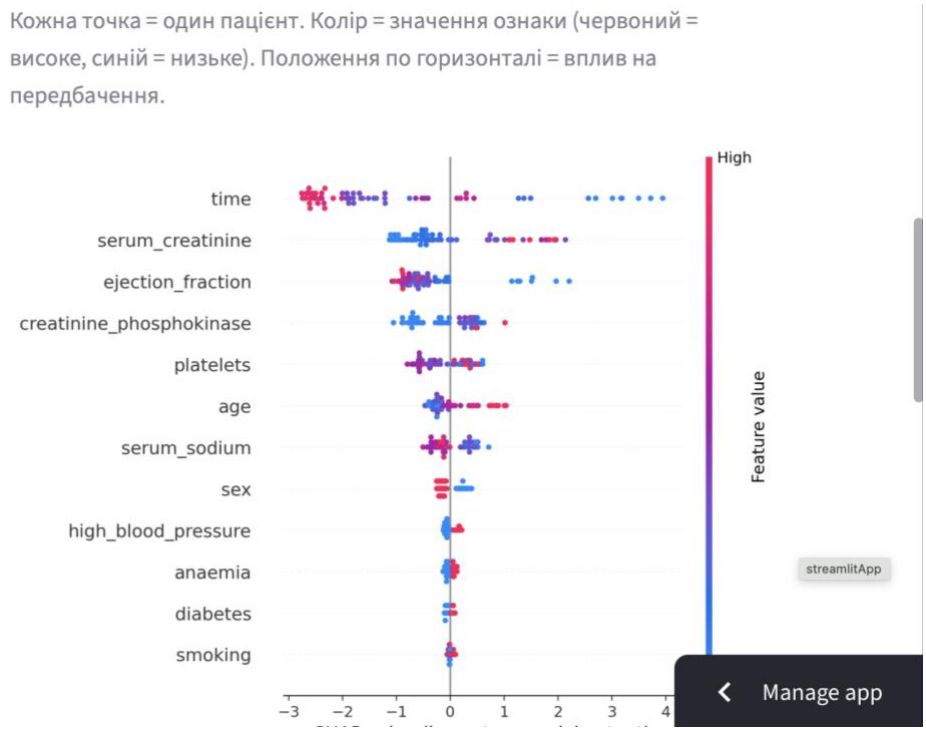

Рисунок 4.9 – SHAP Summary Plot

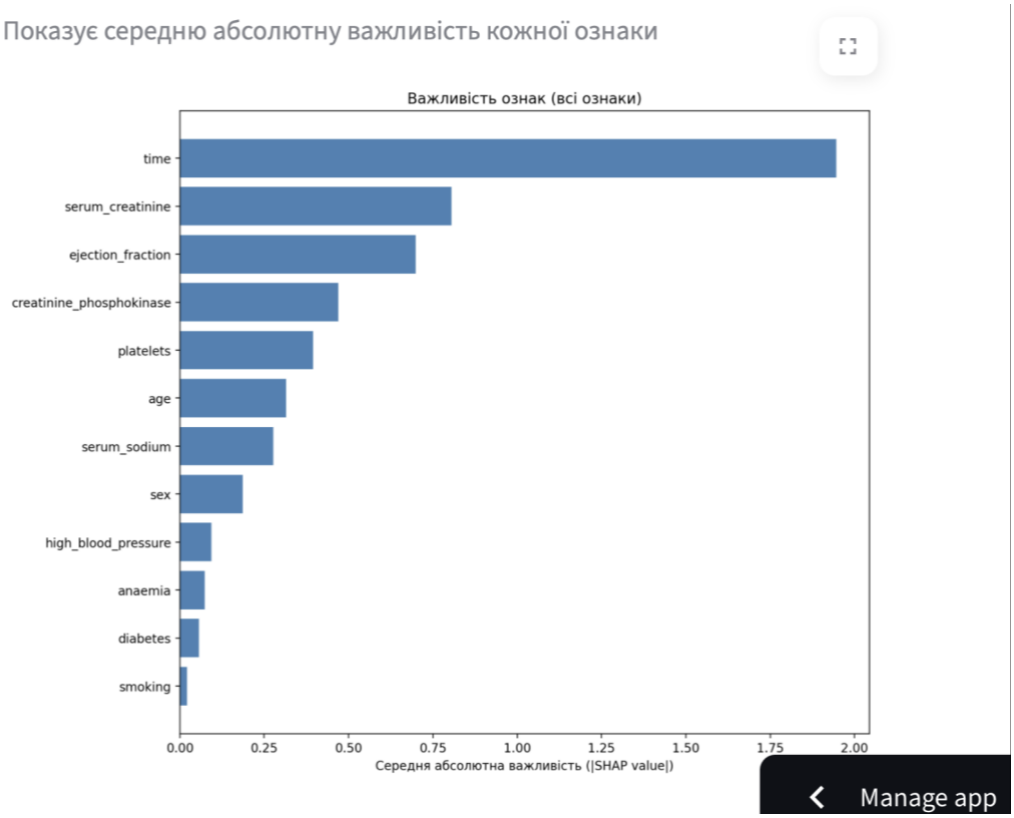

Рисунок 4.10 – SHAP Bar Plot

Крім того, результати глобального аналізу можуть бути використані для виявлення потенційних помилок у даних, надмірної залежності моделі від окремих ознак або прихованих взаємозв’язків між параметрами.

А значить, третій таб системи реалізує інструменти глобальної інтерпретації моделі XGBoost на основі SHAP. Це забезпечує прозорість роботи моделі на рівні всієї вибірки та створює основу для подальшого аналізу взаємодій між ознаками і додаткових методів Explainable AI.

4.5 Реалізація додаткових методів Explainable AI: Partial Dependence Plots та аналіз взаємодій ознак

Для поглибленого аналізу поведінки моделі машинного навчання у програмному рішенні буде реалізовано окрему вкладку, присвячену додатковим методам Explainable AI. Даний таб призначений для аналізу середнього впливу окремих ознак на результат прогнозування, а також для дослідження взаємодій між клінічними параметрами.

Одним із основних методів, реалізованих у даному табі, є Partial Dependence Plots. Для їх побудови буде використано функціональність бібліотеки scikit-learn, зокрема клас PartialDependenceDisplay. У коді формується об’єднаний набір навчальних і тестових даних, що дозволяє отримати більш стабільну оцінку середнього впливу ознак. Користувачеві надається можливість обирати одну або декілька ознак для аналізу, після чого для них будується відповідний PDP-графік.

Для неперервних ознак Partial Dependence Plot відображає зміну середньої ймовірності летального результату при зміні значення ознаки за фіксованих значень інших параметрів (рисунок 4.11). Для бінарних ознак у коді реалізовано окрему логіку, яка полягає у примусовій фіксації значення ознаки на рівні 0 та 1 з подальшим порівнянням середніх прогнозів моделі (рисунок 4.12). Це дозволяє наочно оцінити вплив наявності або відсутності певного клінічного фактору.

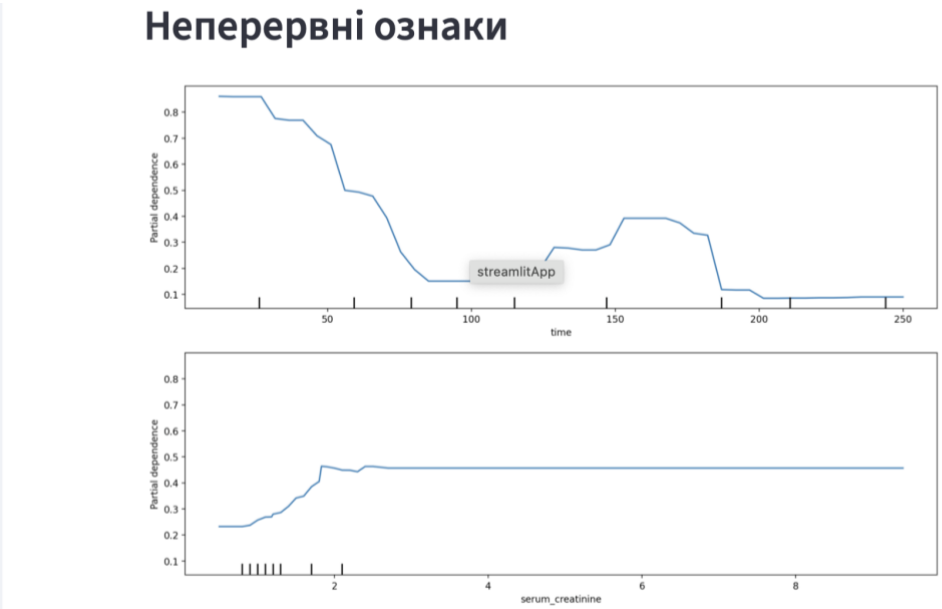

Рисунок 4.11 – PDP з неперервними ознаками

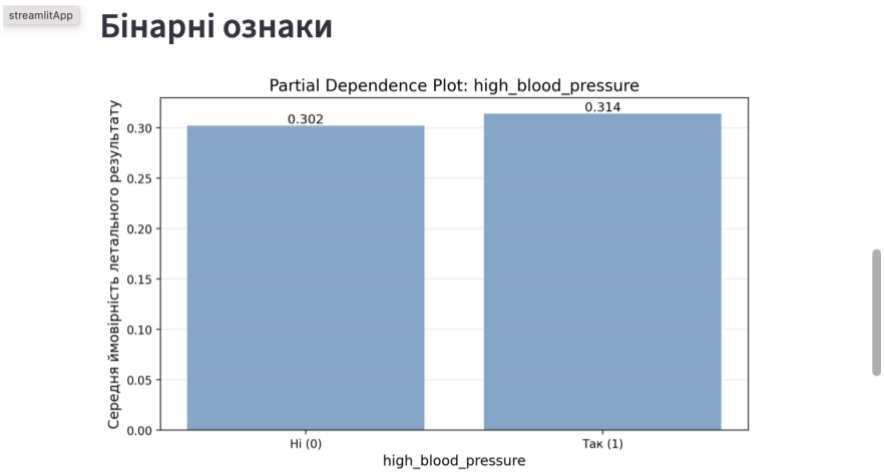

Рисунок 4.12 – PDP з бінарними ознаками

Окрім PDP, у системі буде реалізовано аналіз взаємодій між ознаками з використанням SHAP Interaction Values (рисунок 4.13). Для цього у коді використовується метод shap_interaction_values об’єкта TreeExplainer. З метою зменшення обчислювальної складності аналіз взаємодій виконується на випадковій підвибірці тестових даних. Отримані значення усереднюються за абсолютною величиною, що дозволяє кількісно оцінити силу взаємодії між кожною парою ознак.

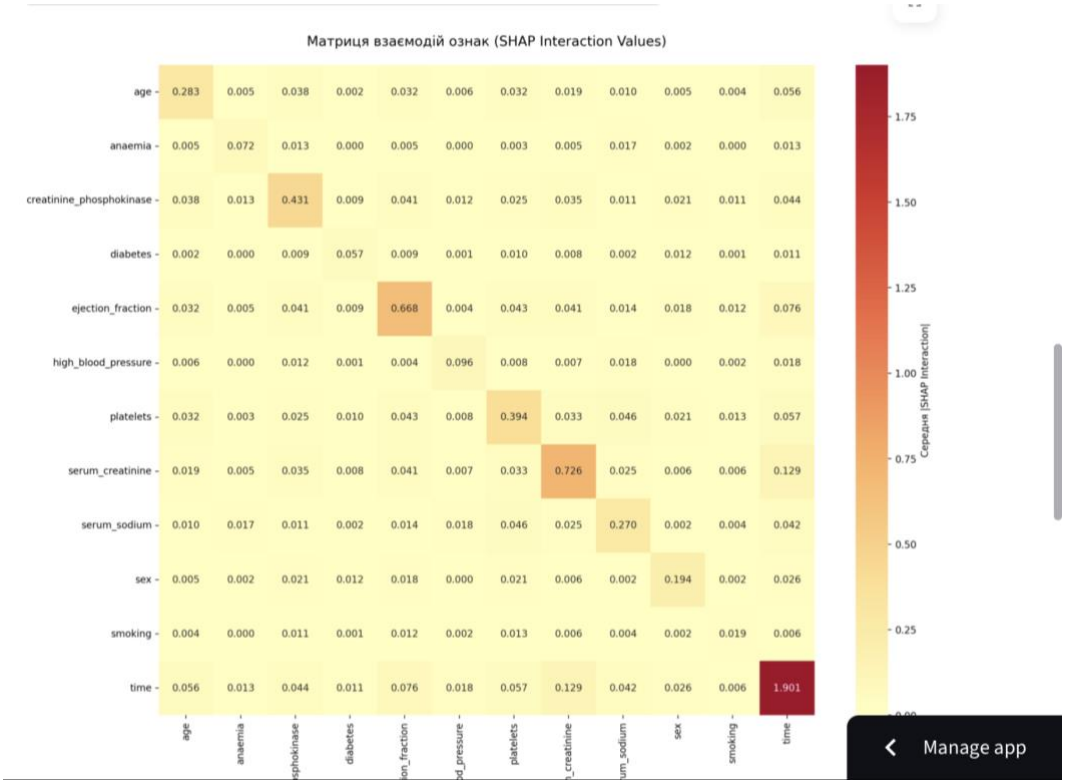

Рисунок 4.13 – Матриця взаємодій ознак (SHAP Interactions)

Результати аналізу взаємодій візуалізуються у вигляді теплової карти, побудованої з використанням бібліотеки seaborn. Діагональні елементи матриці відображають основний ефект окремих ознак, тоді як позадіагональні елементи характеризують парні взаємодії між ними. Додатково у коді формується перелік найбільш значущих взаємодій, що дозволяє виділити ключові комбінації клінічних параметрів, які спільно впливають на прогноз моделі.

Таким чином, четвертий таб системи забезпечує розширений Explainable AI-аналіз, який доповнює локальні та глобальні пояснення дозволяє глибше зрозуміти внутрішню логіку роботи моделі XGBoost та оцінити складні нелінійні залежності між клінічними показниками.

Наприкінці хочу надати посилання проєкту на гітхабі: https://github.com/Giveit2/pythonProjects та посилання на його деплой: https://beginner-level-projects-ff7f4n9jb7bcayqb8wpvno.streamlit.app

5 ПРАКТИЧНА ЦІННІСТЬ ТА ПОДАЛЬША ОПТИМІЗАЦІЯ ПРОЄКТУ

5.1 Аналіз прозорості та клінічної інтерпретованості

Розроблена в межах даної роботи система прогнозування серцевої недостатності орієнтована не лише на отримання точних передбачень, а й на забезпечення повної зрозумілості результатів для кінцевого користувача, для лікаря. На відміну від класичних реалізацій Explainable AI, де інтерпретованість обмежується візуальними графіками, у даному проєкті реалізовано комплексний підхід до прозорості, який поєднує числові метрики, візуальні пояснення та текстові інтерпретації.

Ключовою перевагою системи є те, що XAI-методи інтегровані на всіх рівнях взаємодії з користувачем. Локальні та глобальні SHAP-пояснення супроводжуються текстовими підказками, які пояснюють зміст графіків, напрямок впливу ознак та клінічне значення отриманих результатів. Такий підхід суттєво знижує поріг входу для медичних спеціалістів, які не мають глибокої підготовки у сфері машинного навчання, і сприяє формуванню довіри до системи.

Окрему увагу приділено інтерпретації метрик якості моделі. Усі основні показники – accuracy, precision, recall, F1-score, ROC-AUC, специфічність та негативна прогностична цінність – супроводжуються текстовими поясненнями, які описують їхнє практичне значення у клінічному контексті. Таким чином, лікар отримує не лише числове значення метрики, а й розуміння того, які типи помилок може допускати модель та які ризики з цим пов’язані. Це дозволяє уникнути некритичного використання автоматизованих прогнозів і сприяє усвідомленому прийняттю рішень.

Важливою складовою практичної цінності системи є реалізація контрфактичних пояснень, які подаються у формі зрозумілих текстових

рекомендацій. На основі SHAP-значень визначаються змінювані клінічні показники, що найбільше впливають на ризик, після чого користувачеві пропонуються пояснення щодо потенційних напрямків їх корекції. Це переводить модель із площини абстрактного прогнозування у площину прикладної клінічної аналітики.

Додатково у системі реалізовано окремий довідковий модуль, який виконує освітню та пояснювальну функцію (рисунок 5.1). У цьому розділі користувач може ознайомитися з базовими поняттями Explainable AI, принципами роботи алгоритму XGBoost, описом використаних XAI-методів та клінічним значенням кожної ознаки датасету. Наявність такого табу дозволяє використовувати систему не лише як інструмент прогнозування, а й як навчальне середовище для формування коректного розуміння можливостей і обмежень штучного інтелекту в медицині.

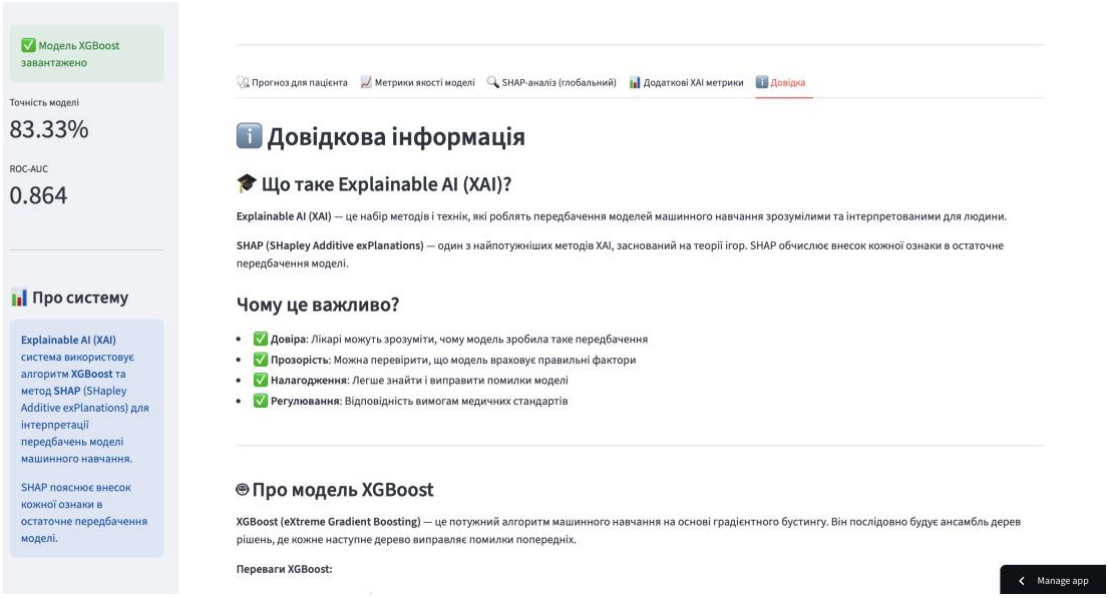

Рисунок 5.1 – Окремий довідковий модуль

У сукупності текстові підказки, вбудовані пояснення метрик, локальні та глобальні XAI-візуалізації, а також довідковий розділ формують цілісну концепцію прозорості системи. Модель не приховує свою логіку, а

навпаки – активно пояснює власні рішення, стимулюючи користувача до критичного аналізу та обґрунтованого використання результатів.

Наше розроблене програмне рішення виходить за межі формального впровадження Explainable AI і демонструє приклад комплексної, користувач-орієнтованої реалізації інтерпретованого машинного навчання. Це значно підвищує клінічну цінність системи, рівень довіри до неї та потенціал практичного впровадження у реальні медичні інформаційні середовища.

5.2 Аналіз подальшого розвитку проєкту

Розроблений програмний продукт має значний потенціал для подальшого розвитку та масштабування як у науковому, так і у практичному напрямках. Поточна версія системи є прототипом, який демонструє можливості поєднання машинного навчання та Explainable AI у задачах медичного прогнозування, однак існує низка напрямків, у яких проєкт може бути суттєво покращений. Перш за все, перспективним є розширення та ускладнення набору даних. Використаний у роботі датасет містить обмежену кількість спостережень і клінічних ознак, що обмежує узагальнювальну здатність моделі. У майбутньому доцільно використовувати більші клінічні датасети, сформовані на основі електронних медичних карток, які включають історію лікування, результати лабораторних та інструментальних досліджень, супутні захворювання та динаміку показників у часі. Це дозволить побудувати більш стійкі та клінічно релевантні моделі прогнозування.

Важливим напрямком розвитку є перехід від статичного прогнозування до аналізу часових рядів. У реальній клінічній практиці стан пацієнта змінюється з часом, тому врахування послідовності вимірювань може значно підвищити точність і практичну цінність прогнозів. Реалізація такого підходу відкриває можливість використання більш складних

моделей, здатних працювати з динамічними даними, у поєднанні з методами Explainable AI для пояснення прогнозів у часі.

Подальший розвиток системи також пов’язаний із розширенням функціоналу Explainable AI. Окрім наявних локальних та глобальних пояснень, у майбутньому доцільно реалізувати інтерактивні контрфактичні сценарії, які дозволять лікарю змінювати значення клінічних параметрів і одразу спостерігати вплив цих змін на прогноз моделі та відповідні пояснення. Такий підхід перетворює систему з аналітичного інструмента на активний засіб підтримки клінічних рішень.

Перспективним напрямком є інтеграція системи з клінічними протоколами та медичними рекомендаціями. Поєднання результатів Explainable AI з клінічними гайдлайнами дозволить не лише виявляти фактори ризику, а й надавати більш обґрунтовані рекомендації щодо можливих напрямків лікування або подальшого спостереження за пацієнтом. Це підвищить практичну значущість системи та сприятиме її впровадженню у реальну медичну практику.

З технічної точки зору доцільним є впровадження підходів MLOps, що забезпечать автоматизований процес оновлення моделей, контроль якості прогнозів та моніторинг зміни характеристик даних. Це дозволить підтримувати стабільність роботи системи при надходженні нових клінічних даних та своєчасно реагувати на можливу деградацію моделі.

Окремим напрямком подальшого розвитку є підтримка кількох алгоритмів машинного навчання в межах однієї системи. У майбутньому може бути реалізований функціонал вибору та порівняння різних моделей, таких як логістична регресія, Random Forest, LightGBM або нейронні мережі для табличних даних. Це дозволить оцінювати не лише точність прогнозів, а й стабільність та клінічну зрозумілість Explainable AI-пояснень для різних підходів, що є особливо важливим у медичних застосуваннях.

Крім того, архітектура системи дозволяє масштабувати її на інші медичні задачі, зокрема прогнозування ризику розвитку хронічних

захворювань або ускладнень. Це відкриває можливість створення універсальної медичної інформаційної платформи з підтримкою Explainable AI, яка може бути адаптована під різні клінічні сценарії. Таким чином, подальший розвиток проєкту полягає не лише у підвищенні точності моделей, а й у розширенні функціональних можливостей, поглибленні інтерпретованості рішень та адаптації системи до реальних умов медичної практики.

ВИСНОВКИ

У даній кваліфікаційній роботі було досліджено проблему прозорості та інтерпретованості рішень штучного інтелекту в медичних інформаційних системах, що є особливо актуальною в умовах зростаючого використання машинного навчання у клінічній практиці. Основну увагу зосереджено на застосуванні методів Explainable Artificial Intelligence з метою підвищення довіри до автоматизованих рішень та забезпечення їхньої клінічної обґрунтованості.

У першому розділі роботи було проаналізовано сучасний стан медичних інформаційних систем, особливості використання алгоритмів машинного навчання в медицині та основні ризики, пов’язані з непрозорістю AI-моделей. Розглянуто проблему «чорної скриньки», а також юридичні, етичні та клінічні вимоги до інтерпретованості рішень у медичній сфері. Показано, що відсутність пояснень суттєво обмежує можливість практичного використання AI-систем у медицині.

У другому розділі було розглянуто теоретичні основи Explainable AI, наведено класифікацію методів інтерпретації та виконано огляд найбільш поширених підходів, зокрема LIME, SHAP, Partial Dependence Plots та інших. Проведений аналіз показав, що метод SHAP є найбільш придатним для задач медичного прогнозування завдяки своїй теоретичній обґрунтованості, універсальності та здатності надавати як локальні, так і глобальні пояснення.

У третьому розділі було сформульовано медичну задачу класифікації ризику летального результату для пацієнтів із серцевою недостатністю, виконано аналіз обраного датасету та обґрунтовано вибір алгоритму XGBoost як основної моделі машинного навчання. Показано, що XGBoost є ефективним інструментом для роботи з табличними медичними даними та добре поєднується з методами Explainable AI.

У четвертому розділі реалізовано програмне рішення у вигляді інтерактивної медичної інформаційної системи з веб-інтерфейсом. Система забезпечує введення даних пацієнта, прогнозування ризику на основі моделі XGBoost, аналіз якості моделі за допомогою набору метрик, а також глибоку інтерпретацію рішень із використанням SHAP. Особливу увагу приділено локальним поясненням, контрфактичному аналізу, глобальному аналізу важливості ознак та взаємодій між ними. Додатково реалізовано текстові підказки, пояснення метрик та окремий довідковий розділ, що підвищує зрозумілість системи для користувача.

У п’ятому розділі проаналізовано практичну цінність розробленої системи та можливості її подальшого розвитку. Показано, що поєднання точного машинного навчання з Explainable AI дозволяє створити інструмент підтримки клінічних рішень, який не лише прогнозує ризики, а й пояснює логіку своїх рішень, сприяючи підвищенню довіри з боку лікаря. Також окреслено перспективи масштабування системи, розширення датасетів, інтеграції кількох моделей машинного навчання та поглиблення функціоналу інтерпретації.

У результаті виконання дипломної роботи досягнуто поставленої мети – розроблено медичну інформаційну систему з інтегрованими методами Explainable AI, яка забезпечує прозоре, інтерпретоване та клінічно зрозуміле ухвалення рішень. Отримані результати підтверджують доцільність використання Explainable AI у медичних системах і можуть бути використані як основа для подальших досліджень та практичних впроваджень у сфері цифрової медицини.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. IBM. What is Explainable AI (XAI)? | IBM. *IBM*. URL: https://www.ibm.com/think/topics/explainable-ai (date of access: 21.12.2025).

2. Contributors to Wikimedia projects. Explainable artificial intelligence - Wikipedia. Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence (date of access: 21.12.2025).

3. Бодянский Е. В. Искусственные нейронные сети: архитектуры, обучение, применения. Харьков ТЕЛЕТЕХ, 2004. 372 с.

4. Minsky M., Papert A. Perceptrons – Expanded Edition. Cambridge, MA: The MIT Press, 1988. 308 p.

4. NumPy. URL: https://numpy.org (date of access: 21.12.2025).

5. pandas – Python Data Analysis Library. URL: https://pandas.pydata.org (date of access: 21.12.2025).

6. scikit-learn: machine learning in Python – scikit-learn 1.5.0 documentation. URL: https://scikit-learn.org/stable/ (date of access: 21.12.2025).

7. Welcome to Python.org. URL: https://www.python.org(date of access: 21.12.2025).

8. ДСТУ 3008:2015. Документація. Звіти у сфері науки і техніки. Структура та правила оформлюванню. «УкрНДНЦ», 2016. С. 31.

9. Загальні методичні вказівки з дипломного проектування в університеті. Упоряд.: Ковтун П.С., Дудар З.В., Журавльов В.Я., Шкіль О.С. Харків: ХНУРЕ, 2003. С. 40.

10. Методичні вказівки щодо структури, змісту та оформлення навчально-методичної літератури для авторів (упорядників) оригіналів. Упорядн.: П.С.Ковтун, І.О. Мілютченко, Б.П. Косіковська. Харків: ХНУРЕ, 2002. С. 60.

11. NTA. «Консервируем» данные: сравниваем модуль pickle и альтернативные способы сериализации. URL: https://habr.com/ru/articles/766904/ (date of access: 21.12.2024).

12. *EDPS Homepage | European Data Protection* Supervisor. URL: https://www.edps.europa.eu/system/files/2023-11/23-11 16_techdispatch_xai_en.pdf (date of access: 21.12.2025).