Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Інформаційних управляючих систем
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)
Дослідження моделей автоматизованої діагностики захворювань у системі приватного медичного закладу
(тема)

Виконала:
Студентка 2 курсу, групи ІУСТм-22-1
Буцька Анастасія Сергіївна
(прізвище, ім'я, по батькові)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
Освітня програма Інформаційні управляючі системи та технології
(повна назва освітньої програми)
Керівник професор кафедри ІУС
Ірина ПАНФЬОРОВА
(Власне ім'я ПРІЗВИЩЕ)

Допускається до захисту

Зав. кафедри _____ Костянтин ПЕТРОВ
(підпис) (Власне ім'я ПРІЗВИЩЕ)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Інформаційних управляючих систем

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 Інформаційні управляючі системи та технології
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

«_20_» _листопада_ 2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентці Буцькій Анастасії Сергіївні
(прізвище, ім'я та по батькові)

1. Тема роботи Дослідження моделей автоматизованої діагностики захворювань у системі приватного медичного закладу
затверджена наказом університету від 16 листопада 2023 р. № 1359Ст
2. Термін подання студентом роботи до екзаменаційної комісії 15 січня 2024 р.
3. Вихідні дані до роботи: науково-технічна література, публікації та інтернет-ресурси, що стосуються теми кваліфікаційної роботи, фактичні дані про клієнтів приватного медичного закладу за рік

4. Перелік питань, що потрібно опрацювати в роботі: огляд та аналіз сучасного стану об'єкта дослідження, огляд і аналіз існуючих медичних інформаційних систем, опис та аналіз функціональних та структурних особливостей об'єкта приватної клініки як об'єкта автоматизації, постановка задачі кваліфікаційної роботи, теоретичне вирішення задачі автоматизованої діагностики захворювань, дослідження ансамблювання моделей машинного навчання в медичній діагностиці, розробка ансамблевої моделі машинного навчання, оцінка можливостей і обмежень впровадження ансамблевої моделі, програмна реалізація ансамблювання моделей машинного навчання, апробація запропонованої ансамблевої на реальних медичних даних

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз сучасного стану об’єкта дослідження	20.11.2023	виконано
2	Огляд існуючих моделей та методів автоматизованої діагностики захворювань у системі приватного медичного закладу	21.11.2023-22.11.2023	виконано
3	Постановка задачі дослідження	23.11.2023-24.11.2023	виконано
4	Теоретичне вирішення задачі автоматизованої діагностики захворювань	25.11.2023-29.11.2023	виконано
5	Розробка ансамблевої моделі машинного навчання	30.11.2023-03.12.2023	виконано
6	Оцінка можливостей і обмежень впровадження ансамблевої моделі	04.12.2023-05.12.2023	виконано
7	Програмна реалізація ансамблювання моделей машинного навчання	06.12.2023-09.12.2023	виконано
8	Апробація ансамблевої моделі машинного навчання на реальних медичних даних	10.12.2023-13.12.2023	виконано
9	Опис отриманих теоретичних та практичних результатів дослідження	14.12.2023-15.12.2023	виконано
10	Написання пояснювальної записки	16.12.2023-01.01.2024	виконано
11	Підготовка презентації	02.01.2023-10.01.2024	виконано
12	Надання роботи для перевірки на плагіат	11.01.2024	виконано
13	Попередній захист	11.01.2024	виконано
14	Надання роботи на рецензію	12.01.2024	виконано
15	Захист	17.01.2024	виконано

Дата видачі завдання: «20» листопада 2023 р

Студент _____
(підпис)

Керівник роботи _____ проф. каф. ІУС Ірина ПАНФЬОРОВА
(підпис) (посада, власне ім’я, ПРІЗВИЩЕ)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи: 110 сторінок, 33 рисунків, 10 таблиць, 2 додатка, 38 джерел.

АВТОМАТИЗОВАНА ДІАГНОСТИКА, АНСАМБЛЮВАННЯ, МОДЕЛІ МАШИННОГО НАВЧАННЯ, ПРОГНОЗУВАННЯ, ETL, STACKING.

Кваліфікаційна робота спрямована на вивчення моделей, методів і технологій автоматизованої діагностики захворювань в інформаційній системі приватного медичного закладу.

Мета кваліфікаційної роботи – аналіз сучасного стану систем автоматизованої діагностики, дослідження та удосконалення існуючих моделей вирішення задачі автоматизованої діагностики захворювань в приватних медичних закладах.

В роботі досліджуються теоретичні основи автоматизованих систем діагностики та їх практичне застосування в медичній галузі. Також розглядаються переваги та недоліки різних моделей і методів автоматизованої діагностики. Дослідження включає комплексний огляд відповідної літератури, приклад та аналіз даних, отриманих з автоматизованої системи діагностики приватного медичного закладу.

Теоретичними результатами дослідження є опис рекомендацій щодо розробки та впровадження ефективної автоматизованої системи діагностики в приватних медичних закладах. Запропоновані рішення спрямовані на підвищення точності та ефективності діагностики захворювань і в цілому на підвищення якості медичних послуг, що надаються приватними медичними закладами.

ABSTRACT

The explanatory note to the qualification work: 110 pages, 33 figures, 10 tables, 2 appendices, 38 sources.

AUTOMATED DIAGNOSTICS, ENSEMBLING, MACHINE LEARNING MODELS, FORECASTING, ETL, STACKING.

The purpose of this work is aimed at studying models, methods, and technologies of automated diagnosis of diseases in the information system of a private medical institution.

The goal of this work is to analyze the current state of automated diagnostic systems, research and improve existing models for solving the problem of automated disease diagnosis in private medical institutions.

The work explores the theoretical foundations of automated diagnostic systems and their practical application in the medical field. The advantages and disadvantages of different models and methods of automated diagnosis are also considered. The research includes a comprehensive review of relevant literature, examples, and analysis of data obtained from the automated diagnostic system of a private medical institution.

The theoretical results of the research include recommendations for the development and implementation of an effective automated diagnostic system in private medical institutions. The proposed solutions aim to increase the accuracy and efficiency of disease diagnosis and, overall, improve the quality of medical services provided by private medical institutions.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

АІС – автоматизована інформаційна система

БД – база даних

ІС – інформаційна система

КТ – комп’ютерна томографія

МІС – медичні інформаційні системи

МРТ – магнітно-резонансна томографія

ПМД – перша медична допомога

СМД – спеціалізована медична допомога

СУБД – система управління базами даних

ЦБД – центральна база даних

ШНМ – штучної нейронної мережі

AI – Artificial Intelligence

AJAX – Asynchronous JavaScript And XML

AWS – Amazon Web Services

CNN – Convolutional Neural Networks

EHR – Electronic Medical Records

ETL – Extract, Transform, Load

GCP – Google Cloud Platform

HDD – Hard (magnetic) disk drive

HIPAA – Health Insurance Portability and Accountability Act;

HTTP – HyperText Transfer Protocol

IDEF0 – Integrated Definition for Functional Modeling

JSON – JavaScript Object Notation

LIS – Laboratory Information Systems

LSTM – Long short-term memory

ML – Machine Learning

MLP – Multilayer Perceptron

NLP – Natural Language Processing

RAM – Random Access Memory

RNN – Recurrent Neural Networks

SADT – Structured Analysis and Design Technique

SVM – Support Vector Machine

ВСТУП

Автоматизована діагностика в медицині означає використання комп'ютерних систем для аналізу даних пацієнтів і генерування діагнозів. Одним із технологічних досягнень останніх років є впровадження автоматизованих діагностичних систем у приватних медичних закладах. Використовуючи різні алгоритми та статистичні моделі, ці системи можуть генерувати точні та своєчасні діагнози для широкого спектру захворювань.

Дослідження моделей та методів автоматизованої діагностики захворювань є важливою задачею в медичній галузі, особливо для приватних медичних закладів. Приватні медичні заклади зазвичай забезпечують високу якість медичної допомоги та забезпечують більш індивідуальний підхід до кожного пацієнта. Однак, для досягнення максимальної точності в діагностиці, вони повинні використовувати сучасні методи та моделі автоматизованої діагностики.

Метою кваліфікаційної роботи є аналіз сучасного стану систем автоматизованої діагностики, дослідження та удосконалення існуючих моделей вирішення задачі автоматизованої діагностики захворювань в приватних медичних закладах.

Кваліфікаційна робота має важливе практичне значення для приватних медичних закладів, оскільки результатами дослідження являється покращення процесу діагностики та лікування захворювань шкіри. Робота має наукове значення, оскільки спрямована на розробку ансамблевої моделі машинного навчання (machine learning – ML), що може призвести до створення більш ефективної системи діагностики та лікування.

Для досягнення поставленої мети в роботі будуть застосовані різні аспекти дослідження, такі як огляд літератури, попередній аналіз існуючих рішень, статистичний аналіз даних та візуалізація даних, опис теоретичного вирішення

задачі. Результати дослідження будуть представлені у вигляді числових показників, графіків та таблиць, які демонструють ефективність розробленої моделі діагностики.

Дослідження спрямоване на покращення процесу діагностики та лікування злоякісних утворень шкіри в системі приватного медичного закладу шляхом дослідження та розробки ансамблевої моделі ML. Результати дослідження можуть бути використані для покращення якості медичної допомоги та забезпечення більш точного та швидкого виявлення захворювань шкіри, що є особливо важливим для приватних медичних закладів, які забезпечують високу якість медичної допомоги та індивідуальний підхід до кожного пацієнта.

1 ОГЛЯД ІСНУЮЧИХ МОДЕЛЕЙ ТА МЕТОДІВ АВТОМАТИЗОВАНОЇ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ У СИСТЕМІ ПРИВАТНОГО МЕДИЧНОГО ЗАКЛАДУ

1.1 Оцінка сучасного стану об’єкта дослідження

Початок медичної автоматизованої діагностики сягає середини 20 століття. На той момент науковці та інженери розробляли перші прилади, призначені для автоматизованого аналізу медичних даних. З появою комп'ютерів і розвитком обчислювальних технологій почалася ера цифрової обробки медичних даних. Це відкрило нові можливості для автоматизації діагностики та аналізу медичних зображень та сигналів. Програмне забезпечення для обробки даних стало необхідним інструментом для лікарів та дослідників, дозволяючи їм швидше та точніше аналізувати інформацію.

Розвиток технологій спричинив значні зміни в галузі охорони здоров’я. Однією з таких змін є використання інформаційних систем (ІС) для автоматизації різних медичних процесів, у тому числі діагностики захворювань. Використання автоматизованих діагностичних систем може допомогти зменшити навантаження на медичних працівників і підвищити точність діагностики [4].

ІС, яка використовується в приватній медичній системі для автоматизованої діагностики захворювань, складається з трьох основних модулів: збору, обробки та аналізу даних. Модуль збору даних передбачає збір даних про пацієнта, включаючи історію хвороби, симптоми та результати медичних досліджень . Модуль обробки та аналізу даних реалізує організацію та структурування даних у єдиному форматі для подальшого аналізу [5].

Модуль аналізу даних передбачає використання алгоритмів ML та статистичних моделей для аналізу даних і створення діагнозу. Модуль збору даних ІС зазвичай виконується медичними працівниками, які збирають дані про пацієнтів

за допомогою різних інструментів, таких як анкети, медичні огляди та лабораторні дослідження. Дані потім зберігаються в базі даних (БД), яка служить вхідною інформацією для компонента обробки даних системи.

Модуль обробки даних системи передбачає організацію та структурування даних у форматі, який можна легко проаналізувати. Зазвичай це робиться за допомогою методів інтелектуального аналізу даних і очищення даних, щоб усунути будь-які помилки або невідповідності в даних. Автоматизувати цей процес можна за допомогою ETL (Extract, Transform, Load) – вилучення, перетворення, завантаження. Один із базових процесів управління сховищами даних, а також найменування класу утиліт автоматизації цього процесу [6]. Потім дані перетворюються у формат, який можна легко проаналізувати за допомогою алгоритмів ML та статистичних моделей.

Модуль аналізу даних системи передбачає використання алгоритмів ML та статистичних моделей для аналізу даних і створення діагнозу. Алгоритми ML, які використовуються в системі, навчаються з використанням великих наборів даних пацієнтів і медичних записів. Ці алгоритми створені для вивчення даних і визначення закономірностей і тенденцій, які можна використовувати для діагностики захворювань.

ІС, яка використовується в приватній медичній системі для автоматизованої діагностики захворювань, має бути розроблена таким чином, щоб бути зручною та доступною для медичних працівників. Для систем даного типу рекомендується використовувати веб-інтерфейс, який дозволяє медичним працівникам вводити дані пацієнтів і отримувати результати діагностики. Система має бути розроблена з високою безпекою та відповідає всім відповідним нормам захисту даних.

Для розуміння роботи системи доцільно створити контекстну діаграму. Методологія IDEF0 (Integrated Definition for Functional Modeling) представляє собою систему функціонального моделювання, яка використовується для опису виробничих функцій. Вона надає засоби моделювання для аналізу, розробки,

перепроєктування та інтеграції ІС та бізнес-процесів. IDEF0 є частиною родини методів IDEF і базується на мові функціонального моделювання SADT (Structured Analysis and Design Technique). SADT є структурованою методикою аналізу та проєктування [7–8].

На рисунку (1.1–1.2) відображено контекстну діаграму першого рівня функціонування системи в цілому після впровадження автоматизованої діагностики у моделі ТО-ВЕ. Ця схема надає докладний опис матеріалів, стандартів і даних, які використовуються та обробляються системою в цілому.

Рисунок 1.1 – Схема бізнес-процесу автоматизованої діагностики захворювань (контекстна діаграма рівня А-0)

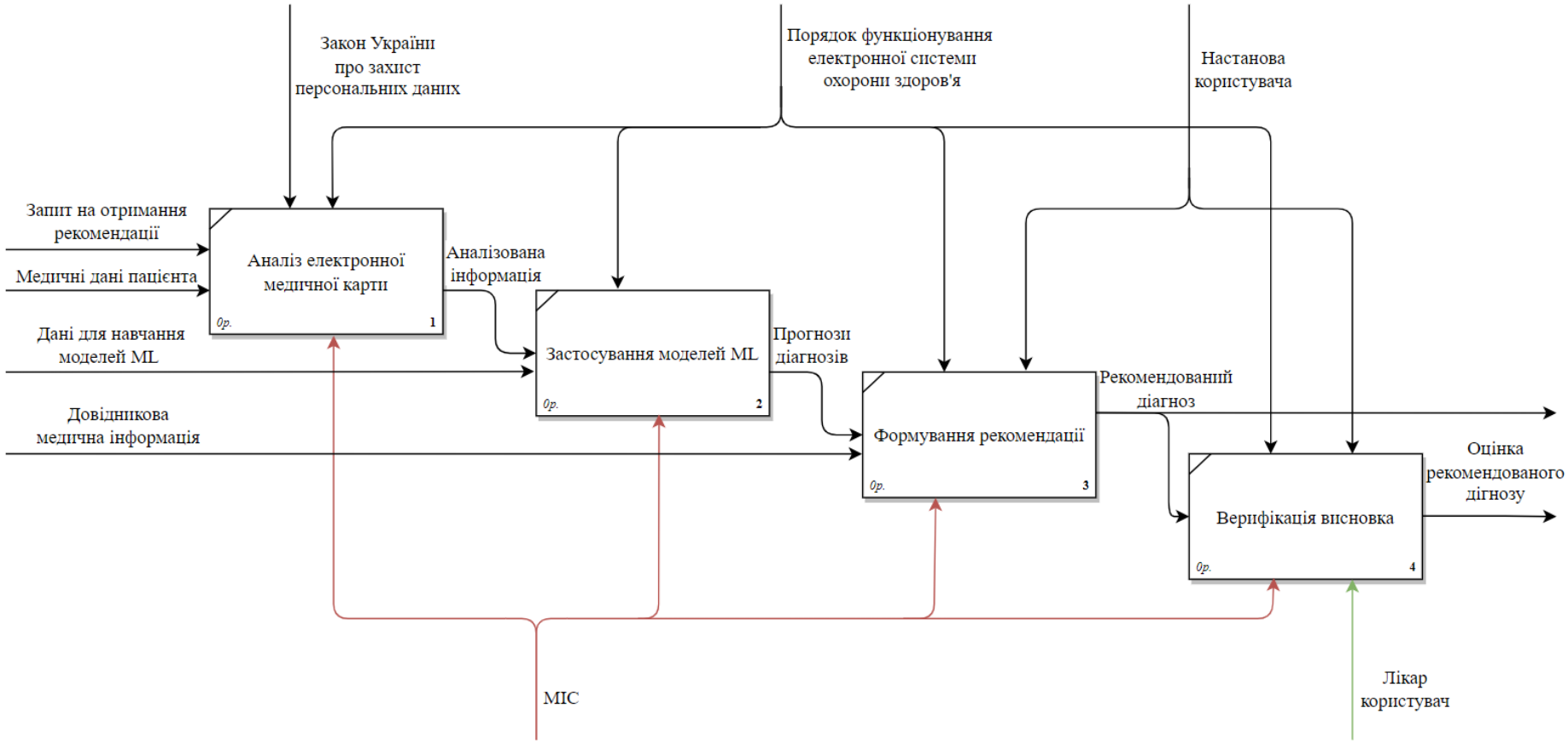

Рисунок 1.2 – Схема бізнес-процесу автоматизованої діагностики (діаграма декомпозиції першого рівня)

Для отримання більш повного уявлення про функціонування системи проводиться процес декомпозиції. В даному контексті функціональну задачу автоматизованої діагностики розглянуто як процес, розбитий на більш дрібні підпроцеси нижчого рівня.

Кожен блок, аналогічно до діаграми першого рівня, детально описується вхідними даними, використовуваними стандартами та матеріалами, а також вихідними даними.

У цьому випадку функціонування системи на першому рівні декомпозиції представлено чотирма взаємозв'язаними задачами (рис. 1.2): аналіз електронної медичної карти, застосування моделей ML, формування рекомендації та верифікація висновка.

Системи такого типу мають бути розроблені таким чином, щоб бути масштабованими та бути легко інтегрованими з іншими медичними системами, такими, як системи електронних медичних записів. Система також має бути розроблена гнучкою для налаштування відповідно до конкретних потреб різних медичних установ.

Приватна медична система «МедТек» є прикладом системи, яка використовує автоматизовані моделі діагностики для покращення результатів лікування пацієнтів. Система використовує обширну базу даних клінічних випадків та медичних досліджень для навчання своїх моделей, що дозволяє їй постійно покращувати свою ефективність і адаптуватися до нових виявлених патологій.

Шляхом аналізу інформації з результатів лікарських обстежень система «МедТек» розвиває глибокі знання у сфері діагностики та лікування різноманітних захворювань. Цей підхід дозволяє системі надійно визначати потенційні ризики для здоров'я та ранні стадії захворювань, забезпечуючи лікарям та пацієнтам можливість вчасно реагувати на проблеми зі здоров’ям.

1.2 Огляд і аналіз існуючих медичних інформаційних систем

В Україні існують медичні інформаційні системи (МІС), які спрощують різні аспекти управління медичним закладом, сприяють підвищенню ефективності догляду за пацієнтами [9].

Єдина державна електронна система охорони здоров'я (eHealth) – це комплексна система, запущена в Україні в 2018 році. Вона має на меті інтегрувати медичні дані з різних джерел, таких як лікарні, поліклініки та аптеки, в єдину цифрову платформу. eHealth складається з декількох МІС, що використовують центральну база даних (ЦБД), яка відповідає за централізоване зберігання і обробку інформації в цих МІС. Окремі лікарні та поліклініки можуть вибирати окремі МІС для власного використання [10].

На рисуноку 1.3 зображено схематичне представлення електронної системи охороно здоров'я eHealth.

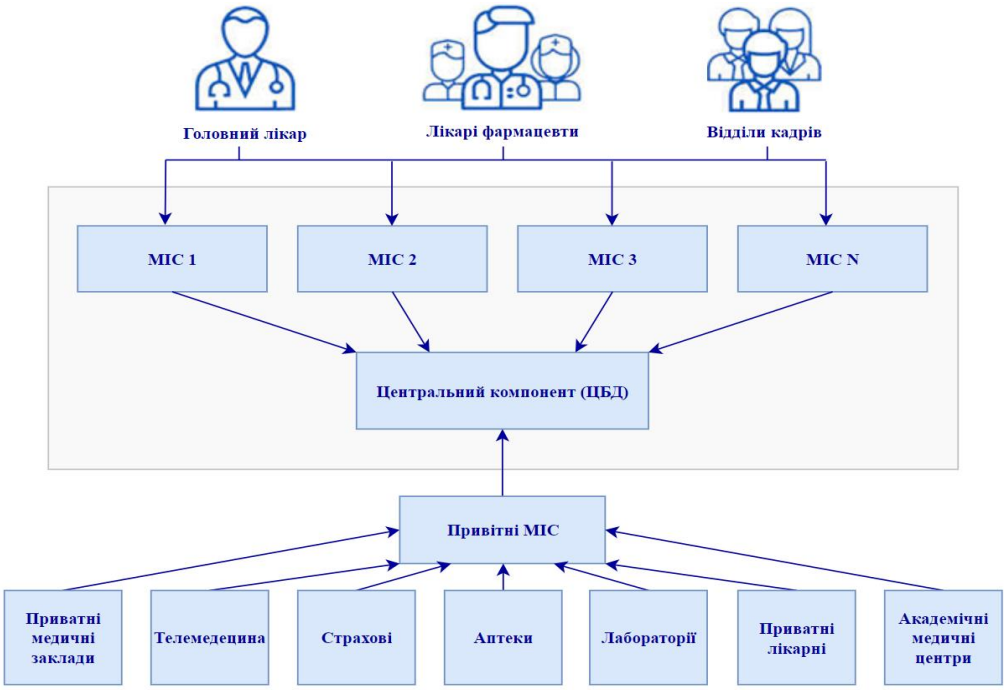

Рисунок 1.3 – Схема взаємодії державної електронної системи охорони здоров'я eHealth з приватними МІС

Нижче наведено функціонал та переваги використання eHealth.

Переваги роботи з eHealth для приватних клінік:

– медичний висновок про тимчасову непрацездатність;
– електронний рецепт;
– вакцинація;
– декларації.

Переваги роботи з eHealth для лікарів та керівників державних закладів охорони здоров'я:

– адміністративний модуль первинної медицини;
– робоче місце лікаря першої медичної допомоги (ПМД);
– електронний рецепт;
– медичний висновок про тимчасову непрацездатність;
– доступ до даних ПМД;
– функціонал обліку наданих послуг;
– адміністративний модуль вторинної медичної допомоги;
– робоче місце лікаря спеціалізованої медичної допомоги (СМД);
– робота з записами про ідентифікованих та неідентифікованих пацієнтів;
– приєднання записів неідентифікованого пацієнта до ідентифікованого;
– доступ до даних;
– медичні висновки;
– план лікування.

Приватні МІС можна класифікувати за своїм призначенням та ознаками:

– МІС базового рівня;
– консультативно-діагностичні системи;
– МІС рівня лікувально-профілактичного закладу;
– скринінгові системи;
– МІС лікувально-профілактичної установи;
– МІС поліклінічного обслуговування.

Модуль автоматизованої діагностики входить до складу консультативно-діагностичних систем. Аналіз результатів порівняння таких МІС подано в таблиці 1.1.

Таблиця 1.1 – Порівняльна характеристика існуючих МІС в Україні

Назва системи	Критерій порівняння	Характеристика
Askep	Опис системи	Askep – це МІС, яка надає електронний доступ до медичних даних, включаючи медичні записи та історію лікування. Вона спрощує процес обміну інформацією між лікарями та пацієнтами.
	Наявність автоматизації діагностики	Askep не має автоматизованих інструментів для діагностики, але надає засоби для зберігання та обміну медичними даними.
	Переваги	Askep дозволяє швидкий та зручний доступ до медичних записів, полегшуючи спілкування між медичним персоналом та пацієнтами. Вона також сприяє збереженню часу та зусиль, пов'язаних з обробкою та обміном медичних даних.
	Недоліки	Недоліком Askep є відсутність автоматизації діагностики та обмежені можливості аналізу медичних даних в системі. Крім того, інтерфейс може бути менш зручним та менш інтуїтивно зрозумілим для користувачів.

Продовження таблиці 1.1

Назва системи	Критерій порівняння	Характеристика
Helsi	Опис системи	Helsi – це інноваційна система електронного здоров'я, яка надає електронні медичні картки та ефективно обробляє медичні дані. Вона забезпечує централізований доступ до медичних записів та дозволяє швидко здійснювати електронний запис на прийом до лікарів.
	Наявність автоматизації діагностики	Helsi використовує автоматизовані методи для підтримки діагностики медичних станів та хвороб. Вона використовує аналітичні алгоритми та ML, що допомагає лікарям швидше та точніше ставити діагнози.
	Переваги	Helsi надає зручний доступ до медичних даних, забезпечуючи пацієнтам можливість переглядати свої результати аналізів, діагностику та історію лікування в онлайн-режимі. Крім того, система дозволяє легко здійснювати електронний запис на прийом до лікарів і сприяє ефективній координації медичного обслуговування.
	Недоліки	Недоліком Helsi є складність впровадження системи, оскільки вона вимагає інтеграції з існуючими медичними системами та забезпеченням високого рівня конфіденційності медичних даних.

Продовження таблиці 1.1

Назва системи	Критерій порівняння	Характеристика
Moniheal	Опис системи	Moniheal – це інноваційна система електронного здоров'я, яка надає електронні медичні картки та поліпшує обробку та управління медичними даними. Вона пропонує розширений набір функцій та інструментів для підтримки медичного персоналу та пацієнтів.
	Наявність автоматизації діагностики	Moniheal використовує інтелектуальні алгоритми та ML для автоматизації процесу діагностики. Вона допомагає лікарям аналізувати медичні дані та приймати більш точні та ефективні рішення.
	Переваги	Moniheal забезпечує простий та зручний доступ до медичних записів, включаючи результати аналізів та історію лікування. Вона також надає можливість ведення електронного журналу самостійної оцінки здоров'я, сприяючи активному залученню пацієнтів до свого здоров'я.
	Недоліки	Одним з недоліків Moniheal є високі вимоги до технічної інфраструктури та навичок користувачів, що може ускладнити впровадження системи. Крім того, можуть виникати питання щодо конфіденційності медичних даних.

Продовження таблиці 1.1

Назва системи	Критерій порівняння	Характеристика
Medcard24	Опис системи	Medcard24 – це МІС, яка надає доступ до електронних медичних карток та дозволяє зберігати та обмінюватися медичними даними.
	Наявність автоматизації діагностики	Medcard24 надає деякі автоматизовані інструменти для підтримки діагностики та аналізу медичних даних.
	Переваги	Medcard24 забезпечує зручний доступ до медичних записів та результатів аналізів. Вона також дозволяє спілкуватися з медичним персоналом в онлайн-режимі та отримувати консультації.
	Недоліки	Недоліком Medcard24 є обмежені можливості автоматизації та недостатня інтеграція з іншими медичними системами.
Health24	Опис системи	Health24 – це МІС, яка надає електронні медичні картки та дозволяє здійснювати моніторинг та управління здоров'ям.
	Наявність автоматизації діагностики	Health24 використовує автоматизовані інструменти для підтримки діагностики та аналізу медичних даних.
	Переваги	Health24 дозволяє легко вести електронний журнал здоров'я, моніторити показники здоров'я та отримувати рекомендації щодо здорового способу життя. Вона також надає можливість консультуватися з медичним персоналом в онлайн-режимі.

Кінець таблиці 1.1

Назва системи	Критерій порівняння	Характеристика
Health24	Недоліки	Недоліком Health24 є обмежені можливості обробки складних медичних даних та відсутність детального аналізу результатів аналізів.

Порівняльна таблиця 1.1 надає загальний огляд основних МІС, включаючи їх опис, наявність автоматизації діагностики, переваги та недоліки.

Вибір лідера серед цих систем залежить від конкретних потреб та вимог. Однак, з урахуванням наданих характеристик, Moniheal може вважатися потенційним лідером. Moniheal пропонує повний опис системи, має автоматизовані інструменти для діагностики, а також переваги у вигляді інтеграції з іншими медичними системами та зручного інтерфейсу. Що стосується недоліків, вони відсутні або менш помітні у порівнянні з іншими системами.

Але не зважаючи на всі переваги даної системи, розробка власної автоматизованої системи діагностики захворювань має свої переваги і може бути обґрунтована наступними аргументами:

– пристосування до потреб конкретного медичного закладу;

– контроль якості встановлення діагнозу;

– зменшення залежності від зовнішніх постачальників МІС;

– інтеграція системи діагностики з іншими модулями приватної МІС та розширення функціоналу;

– конфіденційність то безпека.

Розробка власної системи діагностики дозволяє точно врахувати особливості та потреби конкретної медичної установи чи групи пацієнтів, дозволяє забезпечити високий рівень контролю якості результатів, дозволяє зменшити залежність від зовнішніх постачальників аналогічних систем, дозволяє легко інтегрувати її з

іншими медичними системами та розширювати її функціональність згідно з потребами вашої установи. Власна система діагностики дозволяє забезпечити вищий рівень конфіденційності та безпеки медичних даних пацієнтів.

Враховуючи ці аргументи, розробка власної автоматизованої системи діагностики може бути корисною для медичної установи, яка бажає оптимізувати та покращити процес діагностики захворювань у відповідності до своїх потреб.

1.3 Огляд моделей діагностики захворювань

Діагностика захворювань відіграє вирішальну роль у галузі медицини, дозволяючи медичним працівникам точно ідентифікувати та класифікувати різні захворювання. З розвитком технологій і доступністю маштабування медичних даних моделі ML стали цінними інструментами для діагностики захворювань. Далі представлено огляд моделей діагностики захворювань, їх застосування та використання.

Існує кілька типів моделей діагностики захворювань, кожна з яких підходить для різних сценаріїв і цілей. Одним із поширених типів є модель на основі правил, яка спирається на заздалегідь визначені правила та дерева рішень для класифікації захворювань на основі симптомів і клінічних спостережень. Іншим підходом, що широко використовується, є статистичне моделювання, яке використовує статистичні методи для аналізу медичних даних і виявлення закономірностей, пов’язаних із конкретними захворюваннями.

Моделі на основі правил використовують набір правил «якщо-тоді» для класифікації захворювань. Кожне правило складається з умов, заснованих на конкретних симптомах або клінічних спостереженнях, і пов’язаної мітки захворювання. Наприклад, правило може стверджувати, що якщо у пацієнта гарячка

та біль у горлі, він класифікується як такий, що має високу ймовірність стрептококової ангіни. Ці правила зазвичай випливають із експертних знань і клінічних рекомендацій [11].

Однією з переваг моделей на основі правил є їх інтерпретація. Процес діагностики прозорий, оскільки кожне рішення базується на чітких правилах. Ця прозорість дозволяє медичним працівникам зрозуміти, як ставиться діагноз, підвищуючи довіру до прогнозів моделі. Крім того, правила можна легко оновлювати або змінювати на основі нових доказів або змін у клінічних рекомендаціях, що дозволяє моделі адаптуватися до розвитку медичних знань.

Однак моделі на основі правил мають обмеження. Вони значною мірою залежать від якості та повноти набору правил. Якщо важливі правила відсутні або неточні, діагностична точність моделі може бути скомпрометована. Крім того, якщо моделі, засновані на правилах, використовуються у випадках складних захворювань, які демонструють різноманітні симптоми або вимагають комплексної оцінки багатьох факторів, то правила прийняття рішень можуть стати надто складними, що призведе до великої кількості помилкових позитивних або помилкових негативних результатів.

Дерева рішень забезпечують альтернативний підхід до класифікації захворювань. Це ієрархічні структури, що складаються з вузлів і гілок, де кожен вузол представляє рішення, засноване на певному симптомі або спостереженні, а кожна гілка представляє можливий результат або мітку захворювання. Процес прийняття рішення починається з кореневого вузла і йде за гілками до досягнення листкового вузла, який відповідає остаточному діагнозу.

Однією з переваг дерев рішень є їх простота та легкість інтерпретації. Ієрархічна структура дозволяє медичним працівникам простежити шлях діагностики та зрозуміти аргументацію кожного рішення. Дерева рішень можуть обробляти як категоричні, так і безперервні змінні, що робить їх придатними для широкого діапазону клінічних даних.

Дерева рішень також можуть обробляти відсутні значення та викиди, оскільки вони не вимагають повних даних для всіх змінних. Крім того, дерева рішень можна легко перетворити на набір правил «якщо-тоді», забезпечуючи переваги інтерпретації та прозорості. Однак дерева рішень можуть постраждати від переобладнання, коли модель стає надто специфічною для навчальних даних і погано працює на нових, невидимих даних. Техніки стандартизації та урегулювання, такі як обрізка та встановлення мінімальної кількості зразків на лист, можуть пом’якшити переобладнання та покращити узагальнення.

Обидва підходи мають обмеження, такі як залежність від попередньо визначених правил.

У статистичному моделюванні використовуються різні статистичні методи для аналізу медичних даних і виявлення закономірностей, пов’язаних із конкретними захворюваннями. Статистичне моделювання для автоматизованої діагностики включає кілька ключових етапів. По-перше, збираються медичні дані, які можуть включати демографічні дані пацієнтів, клінічні обстеження, лабораторні результати, дані зображень комп’ютерної діагностики та іншу відповідну інформацію. Потім ці дані впорядковуються та попередньо обробляються, щоб забезпечити послідовність, точність і сумісність даних з різних джерел.

Після того, як дані підготовлені, застосовуються статистичні моделі для виявлення закономірностей і зв’язків, за якими можна відрізнити здорових людей від тих, хто має певні захворювання. Ці моделі можуть включати класичні статистичні методи, такі як логістична регресія, дискримінантний аналіз і аналіз виживання, а також більш просунуті підходи, такі як метод опорних векторів (SVM) і «Випадковий ліс» (Random forest).

Однією з переваг статистичного моделювання в автоматизованій діагностиці є його здатність обробляти великі та складні набори даних. З удосконаленням електронних медичних записів і доступністю великих даних статистичні моделі можуть аналізувати величезні обсяги інформації для отримання значущої

інформації. Виявляючи закономірності та асоціації, які можуть бути неочевидними для людей-спостерігачів, статистичне моделювання дозволяє виявити нові діагностичні маркери та фактори ризику.

Крім того, статистичне моделювання дає кількісну оцінку ймовірності наявності чи прогресування захворювання. Використовуючи статистичні методи для оцінки ймовірностей, медичні працівники можуть приймати обґрунтовані рішення щодо діагностики, планування лікування та прогнозу. Такий кількісний підхід дозволяє більш об’єктивно оцінювати наявні докази, зменшуючи залежність від суб’єктивних суджень.

Однак існують певні проблеми, пов’язані з автоматизованою діагностикою за допомогою статистичного моделювання. Якість і репрезентативність даних є критичними факторами, які можуть впливати на точність і можливість узагальнення моделей. Для забезпечення надійних результатів потрібно ретельно розглянути упередження, відсутні дані та вплив інших факторів.

Ще одна проблема полягає в інтерпретації статистичних моделей. Незважаючи на те, що статистичне моделювання дає цінну інформацію, складність моделей і велика кількість залучених змінних може ускладнити повне розуміння та інтерпретацію основних зв’язків. Вкрай важливо знайти баланс між складністю моделі та можливістю інтерпретації, гарантуючи, що отримані знання є клінічно значущими та практичними.

Крім того, інтеграція статистичних моделей у клінічну практику вимагає ретельної перевірки та оцінки їхньої ефективності. Моделі потрібно оцінювати на незалежних наборах даних, щоб переконатися в їх точності, чутливості, специфічності та клінічній корисності. Важливо встановити чіткі вказівки та протоколи для розгортання та впровадження автоматизованих діагностичних систем на основі статистичного моделювання.

1.4 Огляд методів діагностики захворювань

Методи діагностики захворювань охоплюють широкий спектр варіантів, які використовуються для ідентифікації та класифікації захворювань. Ці методи передбачають збір і аналіз даних про пацієнта, включаючи симптоми, клінічні спостереження, лабораторні тести і дослідження зображень. Нижче наведено огляд різних методів діагностики захворювань, висвітлюючи їх сильні сторони, обмеження та застосування.

Клінічна діагностика є основним методом в охороні здоров'я, який спирається на знання та досвід медичних працівників. Він включає збір анамнезу пацієнта, проведення фізичних оглядів та інтерпретацію клінічних ознак і симптомів для встановлення діагнозу. Клінічна діагностика часто доповнюється діагностичними тестами та візуальними дослідженнями, щоб підтвердити або уточнити діагноз. Клінічний діагноз є важливим, але він може бути суб’єктивним і залежати від досвіду клініциста, що призводить до мінливості діагностичної точності.

Лабораторні дослідження відіграють вирішальну роль у діагностиці захворювань, забезпечуючи об’єктивні та кількісні вимірювання різних фізіологічних і біохімічних маркерів. Тести включають аналізи крові, аналіз сечі, генетичне тестування та молекулярні діагностичні аналізи. Лабораторні дослідження можуть допомогти у виявленні інфекційних захворювань, метаболічних розладів, аутоімунних станів і генетичних аномалій. Вони необхідні для скринінгу захворювань, моніторингу ефективності лікування та відстеження прогресування захворювання. Однак лабораторні тести можуть мати обмеження щодо чутливості, специфічності та економічної ефективності.

Медичні методи візуалізації, такі як рентген, комп’ютерна томографія (КТ), магнітно-резонансна томографія (МРТ) і ультразвук, широко використовуються для діагностики захворювань. Дослідження зображень надають детальну структурну та

функціональну інформацію про організм, допомагаючи у виявленні та характеристиці різних захворювань. Такі методи особливо цінні в радіології, кардіології, онкології та неврології. Інтерпретація медичних зображень вимагає спеціальної підготовки та досвіду, а точність діагнозу залежить від якості методу візуалізації, отримання зображення та інтерпретації.

Методи штучного інтелекту (artificial intelligence – AI), включаючи машинне та глибоке навчання, набули популярності в діагностиці захворювань. Методи AI можуть аналізувати великі та складні набори даних, витягувати релевантні функції та створювати прогнозні моделі. Ці моделі можуть допомогти у класифікації захворювання, оцінці ризику та прогнозі. Методи діагностики захворювань на основі AI показали багатообіцяючі результати в різних сферах медицини, таких як радіологія, патологія та дерматологія. Однак інтеграція AI в клінічну практику вимагає вирішення проблем, пов’язаних із якістю даних, інтерпретацією та перевіркою.

У багатьох випадках діагностика захворювання вимагає інтеграції кількох методів і підходів. Поєднання клінічних результатів, лабораторних тестів, медичної візуалізації та молекулярної діагностики може забезпечити всебічний і точний діагноз. Мультимодальні підходи підвищують точність діагностики та дозволяють отримати більш цілісне розуміння захворювання. Інтегровані методи діагностики особливо цінні при складних захворюваннях, рідкісних станах і випадках, коли один метод може не надати достатньої інформації для точного діагнозу.

Інтеграція багатьох методів діагностики, розробка автоматизованих діагностичних систем і використання даних про пацієнтів є перспективними для підвищення точності діагностики та персоналізованого медичного обслуговування. Майбутні дослідження мають бути зосереджені на розробці стандартизованих протоколів, покращенні сумісності даних і забезпеченні етичного та відповідального використання діагностичних методів у практиці охорони здоров’я.

1.5 Технології в автоматизованій діагностиці захворювань

Завдяки впровадженню новітньої медичної апаратури, що має можливість аналізувати величезну кількість показників, стало можливим обробляти значно більше медичних даних. Але це не полегшило процес постановки діагнозу для медичних фахівців. Фактично, проблема обробки великого обсягу інформації, отриманої традиційними методами, ще більше наголосила на необхідності знаходити нові та більш ефективні підходи до аналізу та інтерпретації отакого обсягу інформації для постановки діагнозу.

Інтеграція інформаційних технологій у медичну практику привела до створення різноманітних автоматизованих методів медичної діагностики. Ці методи дозволяють лікарям приймати рішення про наявність або відсутність захворювань у режимі активного діалогу з комп'ютерними системами в реальному часі. І очевидно, що цей напрямок стає однією з основних тенденцій у розвитку сучасної медичної діагностики.

Проте, важливо зазначити, що перехід до використання новітніх технологій провокує нові виклики пов’язані з обробкою даних. Необхідно забезпечити надійність, точність та безпеку автоматизованих технологій, а також враховувати особливості взаємодії між лікарем та комп'ютерною системою. Дослідження у цьому напрямку сприяє подальшому розвитку медичної практики та забезпеченню якісного рівню діагностики та лікування пацієнтів.

Варто звернути увагу саме на технології, які використовуються для автоматизації постановки діагнозу та прийняття медичних рішень. Нижче наведено кілька ключових технологій, які лідирують у цій галузі:

- AI та ML;
- аналіз медичних зображень;
- обробка природної мови (Natural Language Processing NLP,);

– геномний аналіз;
– прогностична аналітика;
– робототехніка та автоматизація.

Алгоритми AI і ML впроваджуються в медичні дослідження та практику з метою аналізу великих обсягів медичних даних. Серед них знаходяться історії пацієнтів, медичні зображення та геномні дані [12]. Ці передові алгоритми можуть виявляти закономірності та кореляції, які можуть залишатися непоміченими для лікарів-людей, сприяючи ранньому виявленню захворювань, надаючи варіанти лікування та прогнозуючи результати для пацієнтів.

Засоби, базовані на AI, здатні аналізувати медичні зображення, такі як рентгенівські знімки, магнітно-резонансна томографія (МРТ) і комп'ютерна томографія (КТ). Ці інструменти забезпечують швидке і точне виявлення аномалій, уражень або пухлин, що допомагає радіологам зосередитися на більш складних випадках і підвищує загальну швидкість і точність діагностики.

Також важливим елементом є обробка NLP, які дозволяють видобувати цінну інформацію з неструктурованих текстових даних, таких як медичні звіти та клінічні нотатки. NLP допомагає узагальнити історії пацієнтів, отримувати відповідну інформацію та навіть автоматизувати процес створення медичних звітів.

Завдяки прогресу в геноміці, секвенування геному пацієнта, тобто встановлення послідовності нуклеотидних основ ДНК, стало реальністю. AI може допомогти інтерпретувати цю генетичну інформацію, виявляти потенційні генетичні маркери захворювань і розробляти індивідуальні плани лікування.

Не менш важливою є і прогностична аналітика, де моделі ML можуть аналізувати дані пацієнтів і передбачати прогресування захворювань або потенційні ризики для здоров’я. Це може призвести до раннього втручання та впровадження профілактичних заходів, що значно покращує результати лікування пацієнтів.

Роботизовані системи можуть допомогти під час хірургічних операцій, забезпечуючи більш точні рухи та потенційно зменшуючи інвазивність процедур.

Автоматизація також може покращити лабораторні робочі процеси, від обробки зразків до аналізу даних.

Хоча телемедицина та інструменти віддаленого моніторингу не є повністю діагностичною технологією, вони дозволяють постачальникам медичних послуг дистанційно контролювати стан здоров’я пацієнтів. AI може допомогти проаналізувати дані, зібрані за допомогою цих інструментів, попереджаючи медичних працівників про будь-які тривожні тенденції.

Важливо відзначити, що, хоча ці технології мають величезні перспективи, вони також пов’язані з труднощами, які необхідно розглянути. Етичні проблеми щодо конфіденційності пацієнтів, упередженості алгоритму та ролі людського досвіду в прийнятті рішень. Регуляторним органам необхідно встановити вказівки та стандарти для забезпечення відповідального розвитку та впровадження цих технологій.

Підсумовуючи, технології, що використовуються для автоматизації діагностики та прийняття медичних рішень, швидко розвиваються та мають потенціал трансформувати охорону здоров’я. Використовуючи AI, ML та інші вдосконалені інструменти, медичні працівники можуть надавати більш точні та своєчасні діагнози, що призводить до покращення результатів лікування пацієнтів і підвищення ефективності систем охорони здоров’я в цілому.

1.6 Висновки та постановка задачі дослідження

Дослідивши моделі, методи та технології в автоматизованій діагностиці захворювань виділено ряд питань, які потрібно опрацювати. По-перше, в чому недоліки існуючих рішень та як поєднувати існуючі моделі в автоматизованій діагностиці для отримання рішення задачі автоматизованої діагностики

захворювань. По-друге, необхідно спроєктувати ансамблеву модель використання ML для допомоги в діагностиці захворювань, бо це стає актуальною точкою в сучасних дослідженнях [13].

Для ефективного аналізу та діагностики захворювань необхідно зібрати медичні дані з різних джерел та привести їх до єдиного стандартного вигляду. Адже для точної діагностики необхідно мати доступ до обширних та репрезентативних медичних даних. Якщо дані недостатні або недостовірні, це може призвести до неточних результатів. В якості попередньої обробки та підготовки даних необхідно спроєктувати ETL процес, який допоможе вирішити задачу збору та приведення даних до стандартного вигляду.

Наступною проблемою, з якою слід попрацювати, є неоднозначність діагнозу. Певні захворювання мають схожі симптоми або можуть проявлятися по-різному у різних пацієнтів. Моделі ML можуть мати обмежені можливості розпізнавання таких варіацій. Тож на допомогу може прийти AI та алгоритми системи прийняття рішень, які надають лікарям рекомендації та допомагають при прийнятті клінічних рішень на основі найновіших наукових даних та стандартів.

Тож спільне застосування цих трьох категорій: моделей, методів і технологій дозволяє створювати комплексні системи автоматизованої медичної діагностики, здатне ефективно аналізувати різноманітні медичні дані та надавати цінні інформаційні ресурси для лікарів і медичного персоналу.

З математичної точки зору задача автоматизованої діагностики поєднує в собі елементи ML, статистики та патерн-розпізнавання, і її основна мета полягає у розробці та застосуванні підходів, що дозволяють системам автоматично визначати стани та аномалії на основі аналізу даних. У контексті ML завдання автоматизованої діагностики може розглядатися як завдання класифікації, де потрібно визначити клас або категорію, до якої належить поточний стан системи або об'єкта. Задача автоматизованої діагностики захворювань може бути пов'язано із задачею регресії,

де необхідно передбачити числове значення, таке як параметр чи характеристика обстеження.

Особливістю задачі дослідження є різноманітність медичних даних, необхідність інтерпретованості моделей, дотримання етичних норм конфіденційності даних пацієнтів, вирішення невизначеності та варіабельності медичних випадків та навчання на обмежених даних.

Метою магістерської кваліфікаційної роботи є дослідження та удосконалення існуючих моделей вирішення задачі автоматизованої діагностики захворювань в приватних медичних закладах.

Для досягнення мети магістерської кваліфікаційної роботи необхідно вирішити наступні задачі дослідження:

– на основі проведеного аналізу існуючих МІС та варіантів вирішення задачі дослідження вдосконалити вирішення задачі автоматизованої діагностики захворювань;

– модифікувати існуючі моделі з урахуванням характерних обмежень даної задачі;

– практична реалізація математичного механізму для вдосконалення вирішення задачі автоматизованої діагностики захворювань;

– експериментальна перевірка отриманих результатів вирішення задачі автоматизованої діагностики захворювань та порівняння з існуючими підходами до вирішення задачі.

2 ТЕОРЕТИЧНЕ ВИРІШЕННЯ ЗАДАЧІ АВТОМАТИЗОВАНОЇ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ

2.1 Визначення автоматизованої діагностики

Діагноз – це медичний висновок про стан здоров'я обстежуваного, а також сутність хвороби та стан пацієнта , виражений у прийнятій медичній термінології та заснований на всебічне систематичне вивчення пацієнта [13].

Автоматизація у сфері медичної діагностики – це створення комплексу унікальних математичних і технічних прийомів, які розробляються з однією основною метою – збільшити точність та достовірність медичних діагнозів, а також прискорити процес їх встановлення. Однак слід зазначити, що ключовою характеристикою сучасних МІС є інтелектуалізація.

Інтелектуалізація передбачає взаємодію медичних фахівців з автоматизованими ІС, щоб приймати найбільш оптимальні та ефективні медичні рішення. Такі інтелектуальні ІС, засновані основі записів в БД, надають різноманітні рішення та рекомендації.

У процесі взаємодії з цією системою, медичний фахівець може повністю прийняти запропоноване системою рішення, так і відхилити його, або ж внести корективи відповідно до свого професійного досвіду і знань. Тобто остаточне рішення завжди залишається у компетенції лікаря, і він несе відповідальність за його прийняття.

У цьому світлі автоматизація в галузі медичної діагностики за допомогою інтелектуальних ІС відіграє ключову роль в оптимізації процедур прийняття медичних рішень, проте слід підкреслити, що вона жодним чином не знижує значущості професійних медичних експертів у даному процесі.

Процес ставлення діагнозу складається з кількох важливих етапів – це збір інформації про пацієнта та симптоми хвороби, обробка та аналіз зібраних даних, встановлення діагнозу.

Перший крок полягає у зборі даних про стан пацієнта та всі ознаки, які можуть свідчити про хворобу. Ця інформація може включати медичну історію, результати обстежень, лабораторні аналізи, та інше.

На другому етапі проводиться аналіз отриманих даних. Лікар оцінює всі доступні інформаційні ресурси та діагностичні дані, щоб визначити можливі варіанти хвороби.

На останньому етапі лікар встановлює остаточний діагноз на основі попереднього аналізу та експертних знань.

Автоматизація може бути впроваджена на будь-якому з перерахованих етапів процесу або весь процес повністю. Але необхідно наголосити, що повна автоматизація можлива лише у випадках, де медична діагностика має на увазі чітко визначені алгоритми та завдання, які можуть бути вирішені без безпосередньої участі медичного персоналу.

На першому етапі проєктуються стандартизовані форми для складання медичних історій різного роду та анкет для збору інформації. Отримані відомості про пацієнта вносяться до цифрових або текстових форматів, що відповідають встановленим стандартам. Такий підхід дозволяє ефективно систематизувати та обробляти інформацію про стан пацієнта.

Результат обстеження конкретного пацієнта може бути представлений у вигляді вектора даних, який містить інформацію про його стан і симптоми:

$$f_k = (s1, s2, \dots s_n), \tag{2.1}$$

де s_i = 1, якщо є даний s_i -й симптом;

s_i = 0, якщо цей симптом відсутній;

s_i = - 1, якщо даний симптом не дослідили (i = 1, 2, ..., n).

На другому етапі процесу діагностики акцент зміщується на виділення, аналіз та оцінку зібраних даних. Далі відбувається оцінка важливості виявлених симптомів у контексті різних захворювань. Цей етап відіграє ключову роль у встановленні точного діагнозу та виборі оптимального лікувального курсу.

При вирішенні завдань технічної діагностики відіграє важливу роль опис об'єкта в системі діагностичних ознак, які мають високу діагностичну значущість.

Для кількісної оцінки діагностичної значущості ознак застосовується теорія інформації. Основний принцип полягає у тому, що діагностична цінність ознаки визначається інформацією, що вноситься як ознака оцінки стану системи. Ключове поняття теорії інформації – це ентропія системи, яка є функцією стану системи та використовується для вимірювання обсягу інформації [13].

Для вирішення завдання постановки діагнозу використовуються різноманітні джерела інформації, включаючи медичні знання та опис симптомів, які проявляються у конкретного пацієнта. Медичні знання містять в собі важливу інформацію про зв'язок між конкретними симптомами та можливими захворюваннями. Всі ці симптоми, виявлені у пацієнта, служать ключовими джерелами даних. На основі цієї інформації створюється початковий діагноз, який враховує кількість симптомів (n) та кількість можливих захворювань (m), позначимо симптоми як $S_1,S_2,...,S_n$, а хвороби – $D_1,D_2,...,D_m$. Важливо враховувати різноманітність інформації та обсяг знань, щоб точно поставити діагноз та вибрати оптимальний спосіб лікування для кожного конкретного випадку.

Щодо конкретного хворого символ S_i вказує на наявність і-го симптому, а символ D_j вказує на наявність j-го із зазначених захворювань. Загальна кількість можливих комбінацій становить 2^{n+m}. Кожну з таких комбінацій можна назвати комплексом симптомів та захворювань та позначити як C_j^i, де j - номер захворювання, i - номер симптому.

Довільну комбінацію симптомів можна назвати комплексом симптомів та позначити як s^i, а комбінацію захворювань – комплексом захворювань та позначити як D_j. За такими позначеннями можна розглядати всі можливі комбінації між комплексами симптомів та комплексами захворювань як $S^i * D_j = C^i_j$.

Потім потрібно створити унікальну матрицю, що складається з нулів та одиниць. У цій матриці кожен стовпець представляє різні можливі комбінації симптомів та захворювань. Ця матриця називатиметься логічним базисом даних для симптомів та захворювань. Визначення, які комбінації симптомів та захворювань можливі, а які ні, є завданням медицини. Нехай медичні науковці надають наступні настанови:

– якщо у пацієнта діагностовано захворювання D_1 та одночасно D_2 або D_3 , то у нього обов'язково виявлятимуться симптоми S_1 та S_3;

– якщо D_2 відсутній, то симптом S_2 не спостерігатиметься у пацієнта;

– якщо у пацієнта виявлено захворювання D_2 та D_3, і при цьому відсутній D_1, то у нього обов'язково проявиться симптом S_3;

– пацієнт завжди матиме хоча б один із симптомів, і при цьому у нього обов'язково буде хоча б одне захворювання.

Наявні симптоми можуть свідчити про наявність захворювання, але брак видимих симптомів необов'язково свідчить про відсутність захворювання. Перетворимо інструкції медичного фахівця, а саме пункти 1, 2, 3 і 4, на математичну форму запису, щоб їх можна було використовувати в інформаційній системі, що представлено формули (2.2) – (2.5) відповідно:

$$D_1 \cap [D_2 \cup D_3] \rightarrow [S_1 \cap S_3], \tag{2.2}$$

$$\overline{D_2} \rightarrow \overline{S_2}, \tag{2.3}$$

$$(\overline{D_1} \cap (D_2 \cap D_3) \rightarrow S_3, \tag{2.4}$$

$$[S_1 \cup S_2 \cup S_3] \rightarrow [D_1 \cup D_2 \cup D_3] \tag{2.5}$$

Логічний добуток цих висловлювань створює булеву функцію. Позначимо її як $E = E(S_1, \dots, S_n, D_1, \dots, D_m)$, що визначає зв'язок між захворюваннями та симптомами.

Далі розглянуто приклад діагностики конкретного пацієнта. Деякі симптоми вже виявлені у пацієнта, деякі ні. Крім того, існує третя група симптомів, про які наразі немає інформації. Ця сукупність симптомів може бути неповною і позначається як профіль симптомів пацієнта, позначений як G. Сукупність захворювань (яка також може бути неповною) – це профіль захворювань E. Обидва ці профілі разом утворюють профіль симптомів та захворювань E*G. Профіль G також можна представити як булеву функцію симптомів, $G = G(S_1, \dots, S_n)$.

Тепер можливо сформулювати логічне завдання: на основі заданого профілю симптомів G і функції E знайти профіль захворювання f як булеву функцію аргументів $D_1, \dots, D_m$. $E = E(D_1, \dots, D_m)$.

Варто відзначити, що побудова моделей захворювань ґрунтується на відповідній структурі діагнозу. При прийнятті рішень, враховуючи вагу відповідних оцінок, які базуються на інформації про пацієнта, беруться до уваги як основні, так і супутні захворювання, а також стан окремих функцій органів та регулюючих систем організму. Спеціалізовані системи автоматизованого аналізу надають різні варіанти щодо постановки діагнозу. Проте важливо пам'ятати, що остаточне вирішення завжди залишається за лікарем.

2.2 Технологія обробки даних для автоматизованої діагностики

Розвиток медичної сфери, впровадження новітніх методів діагностики та лікування призвели до швидкого зростання обсягу медичної інформації. З цього

часу стає очевидним, що фахівцям в галузі охорони здоров'я та науковцям в галузі медицини, яким необхідна комп'ютерна грамотність, необхідно адаптуватися до сучасних вимог.

Розвиток сфери комп'ютерних технологій значно впливає на область медицини та охорони здоров'я, вимагаючи від медичних фахівців нових знань і навичок у галузі систематизації та обробки медичної інформації. Сучасний етап розвитку системи охорони здоров'я пов'язаний з появою інтегрованих галузей знань, що включають загальнонаукові принципи, такі як медична кібернетика, охорона здоров'я, а також аспекти менеджменту та маркетингу. Внаслідок цього спостерігається різке збільшення обсягів інформації, що вимагає обробки під час вирішення традиційних завдань медичної сфери:

– постановка діагнозу;

– прогнозування процесу лікування, результатів лікування;

– уявлення цілісного уявлення про захворювання, що формується на основі численних даних, одержуваних за допомогою різноманітних методів обстеження різними фахівцями;

– вибір тактики лікування;

– корекція процесу лікування;

– прийняття рішень щодо управління та регулювання життєво важливими функціями організму в умовах дефіциту часу.

У медичній сфері ключовим джерелом даних є пацієнт. Вивчивши стратегії отримання надійної інформації та дослідивши процес виникнення інформації з цих джерел можна визначити необхідні знання для її правильного розуміння та розглянути варіанти збереження цих даних в МІС.

МІС визначаються великим обсягом даних, які підлягають комплексній обробці на трьох ключових етапах. На початковому етапі вся накопичена інформація про пацієнта упорядковуються у визначених структурах, що відображаються в транзакційних БД. На наступному етапі БД піддається

систематизації, що включає зміну її структури, порядку розташування інформації та визначення взаємозв'язків між елементами. Сформовані дані переносяться на третій етап, де відбувається міграція до сховища даних, на основі якого формуються тематичні вітрини даних. Інформаційна основа МІС включає в себе історії хвороби, виписки, епікризи, стандартизовані карти обстеження, критерії ефективності, діагностичні оцінки та каталог медичних термінів. Важливо відзначити, що на сьогоднішній день не доцільно використовувати автономні медичні комп'ютерні системи для рішення локальних завдань окремих медичних підрозділів.

Однією з ключових технологій, що допомагає вирішити завдання обробки даних, є технологія ETL. Цей процес включає в себе три основних етапи (рис. 2.1):

– вилучення (extract);
– трансформація (transform);
– завантаження (load).

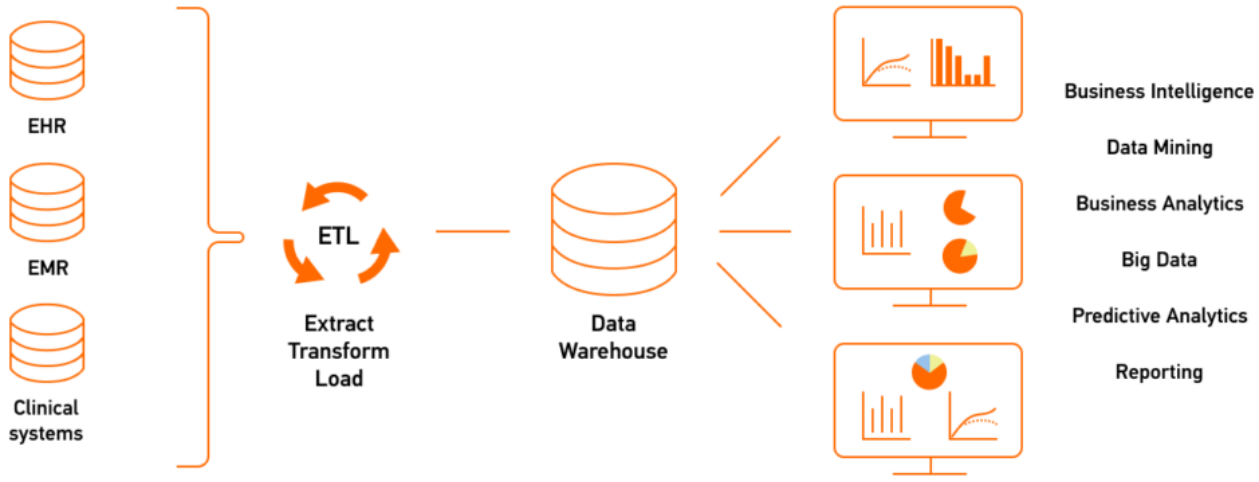

Рисунок 2.1 – Типова архітектура ETL процесу для МІС

На етапі вилучення дані отримуються з різних джерел медичної інформації, таких як електронні медичні картки, лабораторні результати, зображення з

обладнання для діагностики тощо. Ці дані можуть мати різні формати та структури. Отримані дані піддаються процесу трансформації для стандартизації.

Етап трансформації включає в себе очищення даних від помилок, стандартизацію форматів, об'єднання даних з різних джерел та виконання різних обчислень для метрик та атрибутів.

Останній етап передбачає завантаження підготовлених та трансформованих даних в медичну інформаційну систему. Це дозволяє легко доступатися до інформації та використовувати її для різних медичних потреб, включаючи діагностику та прийняття рішень.

Ефективно спроєктований ETL процес виконує завдання вивантаження даних із різноманітних джерел, гарантує цілісність та високу якість отриманих даних, інтегрує їх з різних джерел так, щоб їх можна було легко використовувати. Також, важливо, щоб ETL-процес форматував дані у такий спосіб, який оптимально відповідає подальшому їхньому використанню на презентаційному рівні [14].

На відміну від традиційного пакетного ETL, який опрацьовує значні об'єми даних у заздалегідь визначених інтервалах часу, потоковий ETL представляє собою неперервний процес обробки та аналізу даних, що генеруються у режимі реального часу. Цей підхід відкриває можливість оперативної реакції на надходження даних, що дозволяє вчасно виявляти та використовувати цінну інформацію. Потоковий ETL процес забезпечує неперервний аналіз та інтеграцію даних, що стає більш динамічним та гнучким середовищем для опрацювання інформації у реальному часі.

З використанням потокового ETL дані збираються відразу після їхнього створення та піддаються обробці невеликими порціями. Кожна порція проходить процес трансформації та завантаження в цільову систему, таку як сховище даних чи інформаційна панель у режимі реального часу. Це надає можливість практично в режимі реального часу аналізувати та отримувати розуміння з даних. Такий підхід дозволяє оперативно виявляти залежності в даних та використовувати цінну

інформацію, що сприяє ефективному прийняттю рішень та підвищує точність аналітичних процесів

Потоковий ETL зазвичай використовується в сценаріях, коли дані потрібно проаналізувати та швидко вжити заходів, як у випадку автоматизованої діагностики. Деякі популярні інструменти для потокової передачі ETL включають Apache Kafka, Apache Flink і Apache Spark Streaming.

Далі розглянуто кілька поширених проблем, які можуть виникнути під час впровадження процесу ETL для інтеграції даних охорони здоров'я:

– сумісність між вихідними та цільовими даними;
– якість даних;
– масштабованість;
– відповідність нормативним вимогам;
– безпека даних;
– автоматизація та моніторинг.

Усвідомлення цих проблем дозволить підготуватися до вирішення вищезазначених проблем під час впровадження процесу ETL у сфері охорони здоров'я.

Моніторинг процесів ETL має вирішальне значення для забезпечення якості рішення. У більшості випадків валідації даних і перевірки їх потоку між базами даних достатньо для моніторингу процесів ETL. Щоб досягти належного моніторингу всієї системи, рекомендується використовувати моніторинг зовнішніх завдань, які доставляють дані ззовні та відстеження часу, витраченого на виконання процесів ETL. Аналіз архітектури ETL процесу проводиться для виявлення потенційних загроз і проблем у майбутньому. Відстеження стандартних показників продуктивності, таких як час відповіді, підключення тощо.

Створюючи ETL процеси сховища даних з нуля, варто дотримуватись принципів FAIR, але в такому випадку виникає така проблема, що робочі процеси інтеграції даних є складними та неефективними, якщо виконувати їх вручну [15].

Як і у випадку з будь-яким складним завданням розробки програмного забезпечення, документацією часто нехтують не лише для самих програмних артефактів, але й для будь-якого виконаного робочого процесу [16]. Незважаючи на те, що дані про минулі запуски робочого процесу є дуже корисними, отримання такого типу інформації є складним завданням [17].

2.3 Моделі машинного навчання для автоматизованої діагностики

Традиційний діагностичний процес у сфері охорони здоров’я часто передбачає інтерпретацію медичних даних, таких як медичні зображення, історії пацієнтів і лабораторні результати. Незважаючи на те, що людський досвід є безцінним, він може зайняти багато часу, бути схильним до помилок і мати різну точність. Навпаки, автоматизована діагностика використовує потужність ML для швидкої та узгодженої обробки та аналізу великих обсягів даних охорони здоров’я. Це не тільки зменшує ймовірність людських помилок, але й дозволяє постачальникам медичних послуг більш ефективно розподіляти свій час, зосереджуючись на складних випадках і догляді за пацієнтами.

Вибір відповідної моделі ML для вирішення задачі автоматизованої діагностики є одним із ключових навичок інженера по машинному навчанню. Адже неправильний вибір моделі може призвести до хибного діагностування захворювань, що може призвести до некоректного лікування і погіршення стану пацієнтів. Нарешті, розуміння того, коли потрібно використовувати кожну модель, допоможе визначити, які моделі варто навчати в майбутньому.

Невдало обрана модель може призвести до серйозних проблем, які, в свою чергу, можуть призупинити роботу всього приватного медичного закладу. Зрештою, глибоке розуміння, коли слід використовувати кожну конкретну модель, визначає

той шлях, який дозволить оптимізувати процес навчання моделей для досягнення максимально позитивних результатів.

За визначенням Джорджа Бокса всі моделі можуть містити неточності. Це висловлювання відзначає той факт, що всі моделі, які створюються для опису реального світу, завжди будуть спрощеннями та апроксимаціями реальності [18]. Якщо немає ідеальних моделей, немає і ідеального вибору. Іншими словами, всі поточні моделі, за визначенням, є приблизною апроксимацією. Проте це не означає, що вони не можуть бути корисними. У деяких випадках навіть найменші неточності можуть виявитися корисними для отримання цінної інформації та розуміння складних зв'язків.

Для того, щоб отримати об’єктивну оцінку обраної моделі ML в більш конкретних, хоча й не вичерпних термінах, користь моделі зводиться до чотирьох основних характеристик:

– інтерпретованість – здатність надавати інформацію про вирішувану проблему;

– пояснювальність – здатність пояснити отримані результати;

– гнучкість – здатність описувати складні об'єкти та ситуації;

– споживана складність – вартість запуску та навчання моделі.

Ці характеристики визначають, наскільки модель являється придатною до вирішення конкретної задачі, роблячи акцент на розумінні, зрозумілості, гнучкості та ефективності використання.

При виборі моделі ML варто враховувати те, що легко для AI, зазвичай є складним для людини. І навпаки, те, що для людини надзвичайно просто, для AI є складною задачею.

З цього випливає, що при розв'язанні інтуїтивної задачі можна легко перевірити результати алгоритму, просто поглянувши на них. Користувачу не потрібні пояснення чи осмислення. У більшості випадків вимога пояснень від

алгоритму – це, скоріше, інструмент для налагодження, ніж щось, що підходить для використання.

З іншого боку, у складних для людини задачах перевірка логічних конструкцій алгоритму є одним із найпотужніших способів переконатися в надійності його рішень. Більше того, така перевірка часто корисна для вирішення самої проблеми, наприклад, для пояснення пацієнту, чому саме такий діагноз був поставлений та чому варто дотримуватися тих чи інших рекомендацій в лікуванні.

Отже, емпіричне правило можна сформулювати наступним чином: чим потужніша модель, тим менш піддається вона інтерпретації та поясненню. Якщо дерева рішень вважаються одними з найбільш перевірених моделей, то про Random forest це сказати важко. Аналогічний приклад демонструє, що якщо моделі k-найближчих сусідів досить зрозумілі, гаусівські процеси менш прозорі. Ще одним прикладом є лінійна модель SVM порівняно з усіма її ядерними варіантами, які досить складно осмислити [18].

Слід додати, що при порівнянні моделей виявляється, що ефективність часто шкодить їх інтерпретації та поясненню. Наприклад, використання дерева рішень на основі функцій PCA призводить до значної втрати зручності читання, на відміну від підходу із застосуванням чистого дерева рішень. Те ж саме стосується і ансамблів з моделей ML, котрі розглядаються в підрозділі 2.4.

2.4 Дослідження ансамблювання моделей машинного навчання в медичній діагностиці

Ансамблеве навчання – це підхід ML, який намагається покращити ефективність прогнозування шляхом об’єднання прогнозів з декількох моделей ML. Ідея полягає в тому, що різні моделі можуть продемонструвати різні сильні та слабкі

сторони в різних частинах навчального набору даних або у різних умовах [19]. Було доведено, що використання ансамблевого навчання є ключовим аспектом у багатьох практичних застосуваннях [20].

Одним з головних викликів у використанні ансамблевого навчання є вибір правильної конфігурації з окремих моделей ML. Не всі моделі гармонійно співпрацюють, і неправильний вибір може суттєво погіршити продуктивність ансамблю. Хоча окремі моделі можуть бути перспективними для конкретних МІС, їх поєднання може призвести до компромісів у надійності прогнозів діагностики, а також призвести до збільшення витрат та зниження ефективності МІС.

Проблема вибору підходящих моделей ML також породжує необхідність розрахунку обчислювальної складності та ресурсів для впровадження декількох моделей одночасно. Таким чином, дослідження, спрямовані на вирішення цих труднощів, є необхідними як з теоретичного, так і практичного погляду.

Класифікація моделей ансамблювання виділяє два ключових класи: послідовні та паралельні. При використанні послідовних моделей, базові учні генеруються послідовно. Це сприяє утворенню залежності між ними, оскільки продуктивність моделі підвищується шляхом призначення ваг тим учням, які попередньо допускали помилки.

У випадку паралельних методів, таких як Random forest, базові учні генеруються паралельно. Це сприяє створенню незалежності між ними, адже паралельна генерація базових учнів сприяє підвищенню різноманітності та зменшенню загальної помилки завдяки усередненню прогнозів.

Більшість ансамблевих моделей використовують єдиний алгоритм базового навчання, що може призводити до однорідності учнів. Однак існують також моделі, які застосовують неоднорідні базові учні, створюючи ансамблі з різнотипними моделями. Гетерогенні ансамблі використовують учнів різних типів для покращення гнучкості та здатності моделі пристосовуватися до різноманітних умов.

Загалом будь-яку структуру ансамблю можна розглянути та визначити за допомогою трьох характеристик, які впливають на її продуктивність. Перший – це залежність від навчених базових моделей , незалежно від того, є вони послідовними чи паралельними. Другою характеристикою є методи злиття, які передбачають вибір відповідного процесу для об’єднання результатів базових класифікаторів за допомогою голосування з різною вагою або методу метанавчання. Третьою характеристикою є гетерогенність залучених базових класифікаторів, однорідних чи гетерогенних.

У загальному вигляді схему ансамблевого навчання моделей ML може бути представлено у вигляді комбінації кроків, у якій дані можна тренувати на різних базових класифікаторах, а вихідні дані об’єднуються для отримання остаточного прогнозу (рис. 2.2).

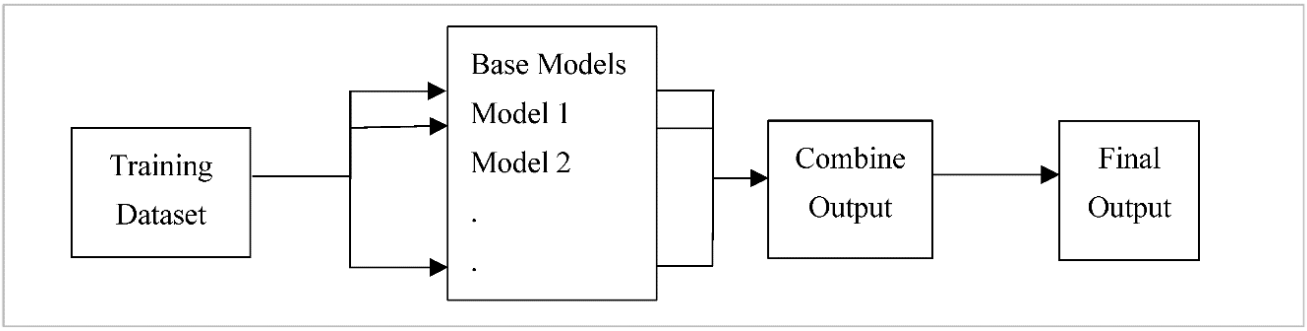

Рисунок 2.2 – Схема ансамблевого навчання моделей ML

Рисунок 2.2 наочно демонструє основну концепцію реалізації ансамблювання моделей ML, де різноманітні моделі можуть навчатися на вхідних даних та об'єднувати свої зусилля для отримання єдиної консолідованої відповіді. Існують різноманітні стратегії для створення ансамблевих моделей. До найбільш популярних з них відносяться техніки, такі як стекінг (stacking), беггінг (bagging), бустінг (boosting) та голосування (voting).

Основні методи ансамблювання моделей ML наведено на рисунку 2.3.

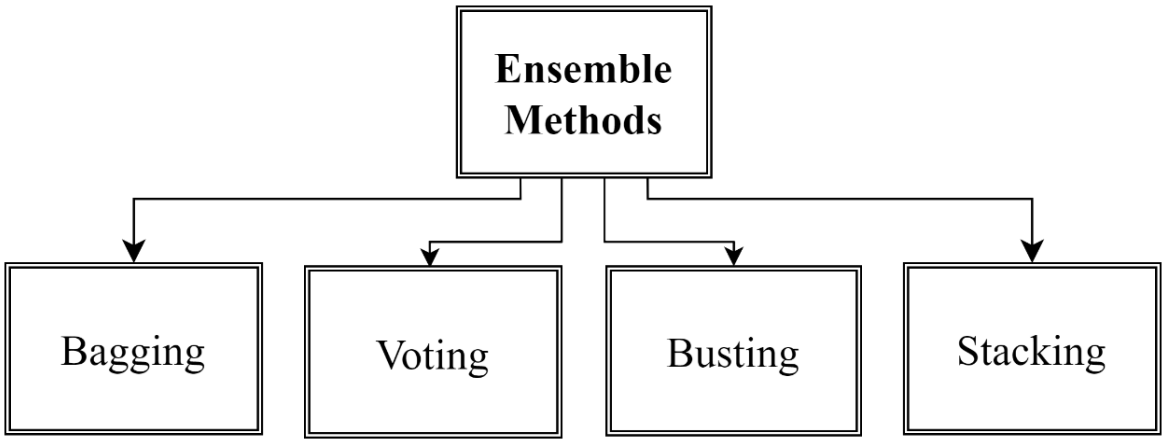

Рисунок 2.3 – Класифікація методів ансамблювання моделей ML

Bagging включає тренування однорідних моделей на різних підмножинах даних, і їхні прогнози об'єднуються шляхом усереднення. Таким чином, прогноз отримується як середнє значення прогнозів кожної окремої моделі.

Voting – це метод, який агрегує прогнози з численних незалежних моделей (базових оцінювачів), щоб зробити остаточний прогноз. Цей підхід дозволяє використовувати множину різноманітних моделей для отримання більш точних та стійких прогнозів.

Boosting – це метод послідовного навчання декількох однорідних моделей, де кожна наступна модель спрямована на корекцію помилок попередніх моделей, що дозволяє системі поступово покращувати свої передбачення.

Stacking – це метод, що включає в себе використання декількох різнорідних слабких моделей, які самостійно навчаються, а потім їхні прогнози об'єднуються для створення остаточного прогнозу, використовуючи результати кожної моделі [21].

Для вибору методу ансамблюваннння виконано порівняльний аналіз, який наведено в таблиці 2.1.

Таблиця 2.1 – Порівняльний аналіз методів ансамблювання

Метод	Опис	Переваги	Недоліки
Bagging	Навчання однорідних моделей на різних підмножинах даних і об'єднання їхніх прогнозів шляхом усереднення	Зменшує ризик перенавчання. Збільшує стійкість та точність	Може бути менш ефективним, якщо моделі взаємозалежні. Не враховує взаємодії між моделями
Voting	Агрегація прогнозів з численних незалежних моделей для отримання остаточного прогнозу	Простий у використанні. Зменшує ризик викидів та шуму у даних	Може бути менш ефективним, якщо моделі сильно відрізняються
Boosting	Послідовне навчання однорідних моделей, де кожна наступна модель виправляє помилки попередніх моделей	Добре працює з невеликими наборами даних. Висока точність, особливо при використанні градієнтного бустінгу	Може бути вразливим до шуму в даних. Може вимагати більше обчислювальних ресурсів
Stacking	Використання декількох різнорідних слабких моделей, які навчаються незалежно та потім об'єднуються	Дозволяє комбінувати сильні сторони різних моделей. Здатний до адаптації до різноманітних завдань	Вимагає більше обчислювальних ресурсів та досвіду

Таблиця 2.1 надає загальний огляд переваг та недоліків кожного методу ансамблювання, вибір конкретного методу може залежати від конкретного завдання та характеристик набору даних. Детальний огляд кожного методу ансамблювання наведено в [22].

Отже, Stacking, або стекінг, є технікою ансамблю, що поєднує в собі використання кількох базових моделей для покращення загальної прогностичної здатності. У контексті медичного аналізу, де точність та надійність є критичними, Stacking може забезпечити значний приріст у визначенні складних взаємозв'язків у медичних даних.

Основною перевагою Stacking є те, що він може ефективно об'єднувати сильні сторони різних моделей, щоб отримати більш точний та стійкий результат. Це особливо важливо в медичних дослідженнях, де невідомі або складні залежності можуть бути ключовими для точного діагнозу.

Також важливо враховувати, що використання Stacking може потребувати певних обчислювальних ресурсів та часу для налагодження та навчання, але це може бути виправданим в областях, де висока точність є вищим пріоритетом. Для того щоб оцінити ефективності обраного методу ансамблювання існує кілька критеріїв оцінки ансамблю, включаючи прогнозні показники. Інші критерії, такі як обчислювальна складність або зрозумілість створеного ансамблю, також можуть бути важливими.

Отже, відповідь на питання, чому варто обрати ансамблювання моделей ML, полягає в тому, що ансамблеві моделі можуть зменшити дисперсію базової моделі, навчаючи кілька моделей на різних підмножинах навчальних даних, що може призвести до підвищення точності.

2.5 Висновки до другого розділу

У другому розділі магістерської роботи було проведено детальне та комплексне дослідження теоретичних аспектів задачі автоматизованої діагностики захворювань. Визначено сутність автоматизованої діагностики як ключового елементу сучасної медичної практики. Аналіз технологій обробки даних дозволив відокремити оптимальні підходи до роботи з медичною інформацією для збільшення точності та швидкості діагностики.

У контексті моделей ML визначено та проаналізовано ключові аспекти, такі як вибір моделей, їхні переваги та недоліки, а також важливість підбору параметрів для досягнення найкращих результатів в медичній діагностиці.

Особлива увага була приділена дослідженню ансамблювання моделей ML як важливого елементу для підвищення ефективності та надійності системи діагностики. Застосування ансамблювання передбачає використання кількох моделей одночасно для прийняття рішення, що сприяє зменшенню впливу випадкових помилок окремих моделей. Цей підхід є важливим у випадках, коли точність та стабільність діагностики є критичними параметрами.

В цілому, зазначені теоретичні аспекти надають необхідний теоретичний фундамент для подальших досліджень та розробки практичних рішень у розділі 3, присвяченому розробці та впровадженню покращених моделей діагностики в приватних МІС.

3 ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ДОСЛІДЖЕННЯ МОДЕЛЕЙ АНСАМБЛЕВОГО НАВЧАННЯ В АВТОМАТИЗОВАНІЙ ДІАГНОСТИЦІ

3.1 Розробка ансамблевої моделі машинного навчання

На сьогодні велика увага приділяється дослідженням комплексних підходів до діагностики захворювань. Незважаючи на це, відзначається недостатньо точні результати, що дають загальновживані комплексні підходи до діагностики п'яти широко досліджених захворювань: діабету, захворювань шкіри, захворювань нирок, печінки та захворювань серця. Однак ці результати є важливим для розуміння патогенезу та розвитку ефективних стратегій лікування та профілактики.

У 2019 році був проведений огляд тенденцій у використанні моделей ML для діагностики [23]. Цей огляд розділив моделі на дві великі категорії: моделі, керовані даними, і моделі, керовані знаннями. Кожна з цих категорій доповнюється контрольованим та неконтрольованим навчанням. Моделі, керовані даними, фокусуються на аналізі та використанні великих обсягів даних для навчання, тоді як моделі, керовані знаннями, використовують збагачені бази експертних знань для розв'язання конкретних завдань. Ці два підходи можуть взаємодіяти. Наприклад, ансамблювання може використовувати моделі, керовані даними, для автоматичного вивчення залежностей у великих обсягах історичних даних, а потім використовувати експертні знання для уточнення та покращення результатів діагностики.

Дослідження, описані в [24], відзначили потенціал ML у медичній діагностиці. Моделі, керовані даними, такі як CNN, проявили вражаючу ефективність в аналізі медичних зображень, зокрема виявленні пухлин на радіологічних зображеннях. Головною перевагою моделі CNN є можливість визначення основних функцій без втручання людини [25]. Використовуючи зображення як вхідні дані, така модель призначає вагові коефіцієнти різним

частинам зображення, дозволяючи виявляти різноманітні аспекти, такі як границі об'єктів, текстури та інші ключові елементи. CNN використовує операцію згортки на базовому рівні замість простого множення матриць [23], відрізняючись від інших моделей.

Контрольована модель навчання, така як SVM, надає користь у багатьох областях прогнозування, зокрема у виявленні захворювань за даними, отриманими з магнітно-резонансної томографії (МРТ). SVM використовується для діагностики нервово-м’язових розладів [26]. NLP застосовується для вилучення цінної інформації з неструктурованих клінічних записів, що допомагає класифікувати захворювання та оцінити ризик перебігу захворювання. Рекурентні нейронні мережі (RNN) здатні розпізнавати послідовності, але їх застосування у випадках зникнення градієнта може ускладнюватися. Однак RNN відкривають можливості для розпізнавання змін у послідовностях [27].

Загальним недоліком моделей ML є їхня складність та висока вартість структур даних навчання. Це може вимагати потужних графічних процесорів та сотень комп'ютерів, збільшуючи вартість для користувачів [28].

Використання декількох моделей у ансамблевому навчанні підвищує обчислювальне навантаження, що може негативно впливати на продуктивність процесу навчання. Щоб подолати ці проблеми, застосовують вдосконалені функції активації, архітектури функцій вартості та методи видалення [29]. ML використовує великі обсяги даних для генерації результатів, що можуть бути використані для персоналізованого прогнозування та клінічних рішень [30]. Це відкриває шлях до розвитку персоналізованої медицини, де беруться до уваги генетичні фактори та спосіб життя людини для профілактики захворювань, лікування та прогнозування перебігу захворювання [31].

Як наслідок, в результаті ансамблювання кілька моделей ML поєднуються для створення точніших прогнозів, ніж ті, що реалізовані за допомогою одного класифікатора. Детальніше кожен тип ансамблювання розглянуто в [32].

Для моделі ансамблю Stacking запропоновано обрати три моделі ML (SVM, CNN і LSTM), які працюють на різних типах даних. Кожна з цих моделей має свої унікальні характеристики та застосування, що дозволяє враховувати різноманітні аспекти вхідних даних. Характеристики ролі кожної моделі наведено в таблиці 3.1.

Таблиця 3.1 – Представлення ролі кожної моделі в ансамблі

Модель	Призначення	Можливості	Вихід
SVM	Використовується для вирішення задач класифікації на основі опорних векторів	Лінійне розділення класів, добре працює з великою кількістю ознак	Прогнози для кожного прикладу в навчальному наборі
CNN	Призначена для обробки зображень, використовується для виявлення шаблонів та особливостей	Виявлення локальних залежностей у вхідних даних, робота з зображеннями	Прогнози для кожного зображення в навчальному наборі
LSTM	Використовується для обробки послідовних даних, таких як текст або часові ряди	Здатна враховувати залежності в послідовних даних, тобто можливість опрацьовувати послідовність в данимих, рекурентні зв'язки	Прогнози для кожного послідовного елемента в навчальному наборі

На першому рівні ансамблю розташовані базові моделі: SVM для класифікації на основі опорних векторів, CNN для обробки зображень та виявлення шаблонів, і LSTM для обробки послідовних даних, таких як текст чи часові ряди. Кожна модель генерує прогнози для свого типу даних.

Другий рівень ансамблю – мета-модель, наприклад, штучна нейронна мережа. Вхідними даними являються прогнози від SVM, прогнози від CNN та прогнози від LSTM. Ця модель комбінує прогнози від базових моделей, навчаючи

їх взаємодію та вагу в кінцевому рішенні. Такий підхід дозволяє враховувати різні аспекти вхідних даних та покращує узагальнюючі можливості ансамблю. Вихід – остаточні прогнози ансамблю.

У більшості ситуацій штучна нейронна мережа зазвичай складається із трьох основних компонентів:

− вхідний шар;

− приховані (обчислювальні) шари;

− вихідний шар.

Типова схема штучної нейронної мережі зображена на рисунку 3.1.

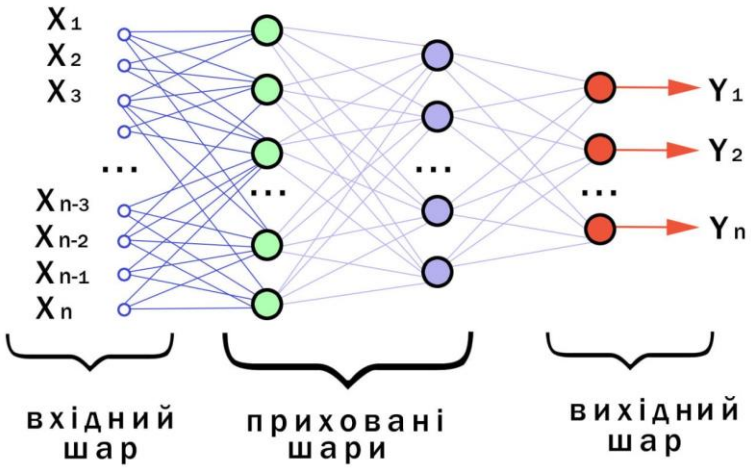

Рисунок 3.1 – Схема основних шарів типової ШНМ

Мета-моделі виступає як вид штучної нейронної мережі ANN, так звана нейронна мережа MLP або Multilayer Perceptron, яка об'єднує прогнози від базових моделей. Мета-модель складається з одного або декількох прихованих шарів MLP, включаючи вхідний і вихідний шари. Кількість нейронів у вхідному шарі дорівнює загальній кількості базових моделей, а кількість нейронів у вихідному шарі дорівнює кількості класів у випадку класифікації або одиничному вихідному значенню у випадку регресії. Кожний вихід базової моделі служить входом для відповідного шару мета-моделі.

Навчання нейронних мереж без учителя – це більш реалістичний підхід до навчання, який враховує біологічні основи штучних нейронних мереж. В основі цього методу лежить використання лише вхідних даних. Під час навчання алгоритм налаштовує ваги мережі так, щоб забезпечити узгоджені відповіді для схожих вхідних даних.

Варто зазначити, що ефективність процесу навчання штучної нейронної мережі залежить від кількох важливих зовнішніх факторів. Зокрема, успішність цього процесу визначається конфігурацією кількох ключових параметрів: Ці параметри, також відомі як гіперпараметри, які визначаються зовнішніми умовами і вимагають уважного налаштування для досягнення оптимальної продуктивності мережі. Важливо зауважити, що ці гіперпараметри не можуть бути автоматично налаштовані самою мережею і вимагають уваги та досвіду розробника. Не маючи функцій активацій, нейромережа втрачає значну частину здатності до навчання.

Для оцінки та налаштування використовуються валідаційні дані. Вагові коефіцієнти для прогнозів кожної базової моделі піддаються налаштуванню для оптимального врахування їхнього внеску. Тестовий набір використовується для оцінки загальної ефективності ансамблю, визначення метрик ефективності, таких як точність, чутливість та специфічність.

Застосування технік нормалізації та регуляризації може підвищити стабільність та уникнення перенавчання. Ця детальна схема дозволяє глибше розуміти взаємодію між базовими моделями та мета-моделлю в ансамблевому навчанні за підходом stacking (рис. 3.2).

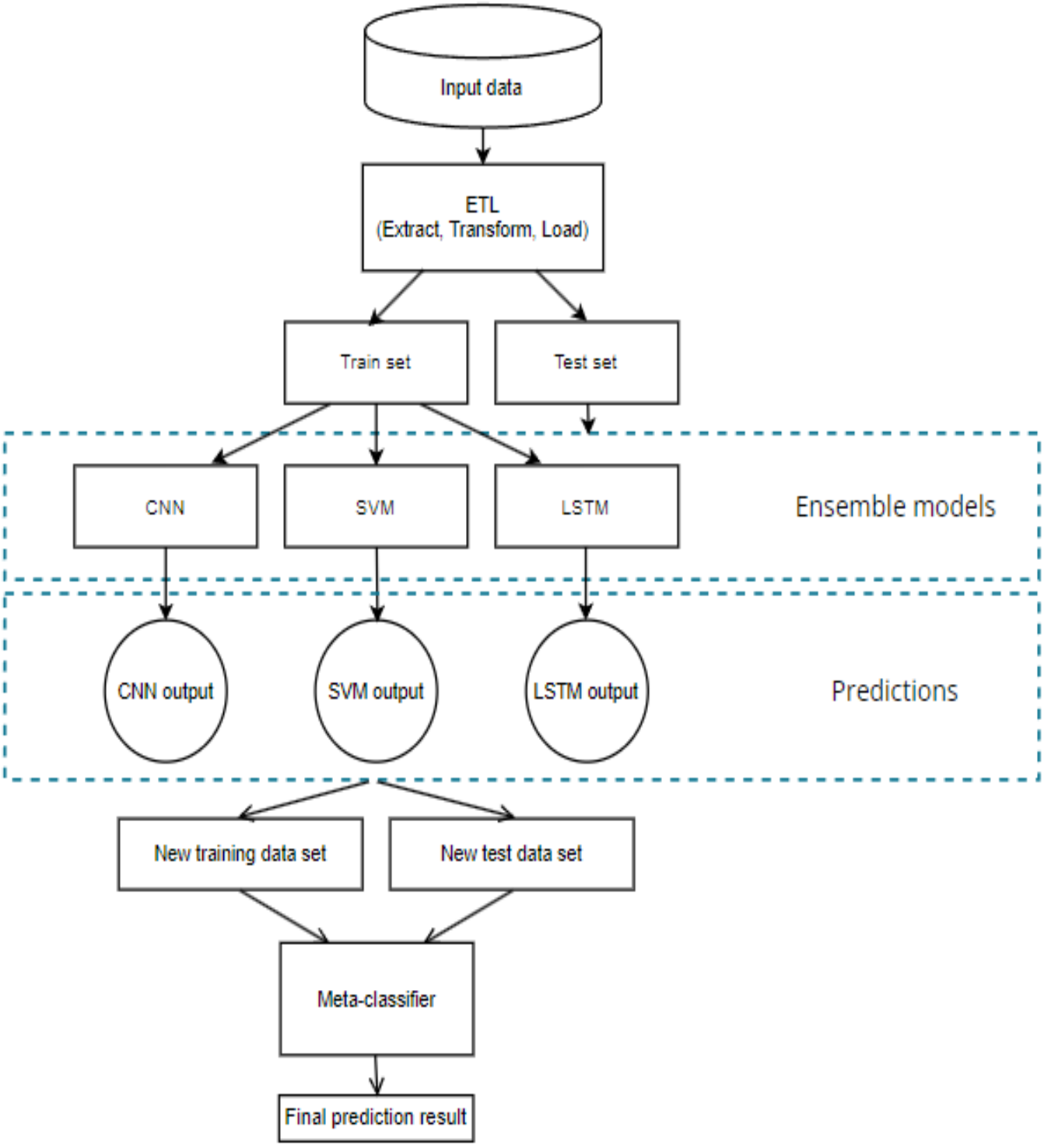

Рисунок 3.2 – Схема архітектури моделі ансамблевого навчання з використанням підходу stacking на основі моделей SVM, CNN та LSTM

3.2 Оцінка можливостей і обмежень впровадження ансамблевої моделі

Розробка ансамблевої моделі ML для діагностики шкірних захворювань є значним кроком вперед у пошуках більш точних і ефективних діагностичних засобів.

Дерматологічні захворювання можуть проявлятися множиною симптомів, що робить точність діагностики складним завданням навіть для досвідчених дерматологів. Помилковий діагноз може призвести до затримки лікування, збільшення витрат на охорону здоров’я та, найголовніше, погіршення самопочуття пацієнтів. ML, що здатне аналізувати величезні обсяги даних і розпізнавати складні закономірності, пропонує вирішення цих проблем. Проте інтеграція моделей ML в МІС охорони здоров'я та МІС клінічних лабораторії породжує численні проблеми і виклики. Однією з них є потенційна декваліфікація медичних та лабораторних працівників через автоматизацію завдань. Моделі ML також мають невизначеності та обмеження продуктивності, що також потребують урахування [33].

Основними характеристиками, якими володіють ансамблеві моделі, являється покращена точність і продуктивність, особливо для завдань з високою складністю та шумом. Ансамблеві моделі також можуть зменшити ризик надмірного та недостатнього оснащення, знаходячи компроміс між зміщенням і дисперсією, а також використовуючи різні підмножини та характеристики даних. Ансамблеві методи можуть обробляти різні типи даних, що є вкрай важливим фактором при діагностуванні шкірних захворювань. Крім того, ансамблеві моделі можуть забезпечити більшу надійність, вимірюючи різноманітність і узгодженість базових моделей, а також надаючи довірчі інтервали та оцінки помилок для прогнозів. Використання ансамблевих моделей дозволить об’єднати завдання класифікації, регресії, кластеризації та виявлення аномалій, використовуючи різні типи базових моделей і методів агрегації.

Однак ансамблеві моделі мають певні недоліки та складнощі, наприклад, вони є вартісні в обчислювальному плані, що вимагає багато часу через потребу в навчанні та зберіганні кількох моделей та об’єднанні їхніх результатів. Це може збільшити складність і вимоги до пам'яті системи. Крім того, результати роботи ансамблевої моделі може бути важко інтерпретувати та пояснити, оскільки вони включають кілька рівнів абстракції та агрегації, що призводить до неоднозначних результатів діагностування.

Для медичного працівника моделі ML є «чорним ящиком» без пояснень, хоча й існують методи для поліпшення інтерпретації, але вони ще не є повністю задовільними [33]. Усунення упереджень медичного працівника, особливо пов'язаних з чутливими показниками, важливе. Зараз наявні методи для пояснення результатів автоматизованої діагностики, котрі ще розвиваються [33-34].

Можливість оцінки продуктивності моделей ML є ключовими аспектами перед впровадженням. Крім того, динамічний характер MIC у сфері охорони здоров'я вимагає постійного розвитку моделей ML для охоплення нових професійних знань і постійного моніторингу.

Недостатня розвиненість інфраструктури MIC, включаючи лабораторні інформаційні системи (LIS) і системи електронних медичних записів (EHR), може ускладнити впровадження та взаємодію з передовими технологіями ML. Покращення інфраструктури MIC має вирішальне значення для підтримки ефективної інтеграції методів ML в MIC.

3.3 Модуль медичної інформаційної системи, що реалізує автоматизовану постановку діагнозу

У МІС може бути ряд модулів, котрі відповідають за вирішення задач, які спрямовані на різні аспекти управління медичними даними та надання зручного та ефективного обслуговування пацієнтів. На базовому рівні можна виділити наступні модулі: модуль реєстрації та прийому пацієнтів, модуль лабораторних досліджень, модуль розкладу та обліку ресурсів, модуль керування процесами, модуль обробки інформації, модуль запису на прийом, модуль аналізу та діагностики, модуль аналітичної та фінансової звітності (рис. 3.3).

Рисунок 3.3 – Модулі МІС

За автоматизацію діагностики відповідає модуль аналізу та діагностики. Всі модулі тісно пов’язані між собою і повинні мати чітко спроєктовану архітектуру, яка відповідає усім вимогам проєктування МІС та забезпечує відмовостійкість, надійність та ефективність.

Модуль аналізу та діагностики складається з трьох основних шарів:

– шар даних;

– обчислювальний шар;

– шар інтерфейсу користувача.

Схематичне зображення модуля аналізу та діагностики представлено на рисунку 3.4.

Рисунок 3.4 – Структура модулю аналізу та діагностики МІС

На рівні шару даних організовано централізоване сховище даних, у якому зберігаються повні дані пацієнтів, включаючи історію хвороби, лабораторні результати, дослідження зображень та іншу відповідну інформацію. На цьому рівні також може бути інтегровано дані із зовнішніх джерел з даними MIC, наприклад, переносних пристроїв або EHR. Необроблені дані пацієнта проходять етапи попередньої обробки для очищення, стандартизації та впорядкування інформації. Це гарантує, що дані мають формат, придатний для аналізу та діагностики.

На наступному рівні будуються ядра обчислювального рівня, які складається з окремих моделей ML, навчених на великих наборах даних. Ці моделі призначені для аналізу даних пацієнтів, діагностики, прогнозування перебігу захворювання та підтримки прийняття рішень медичних працівників.

Дружній інтерфейс користувача дозволяє медичним працівникам використовувати модуль у повсякденній роботі.

Інтерфейс користувача забезпечує інтуїтивно зрозумілу та зручну платформу для медичних працівників для взаємодії з модулем аналізу та діагностики. Інтерфейс містить інформаційні панелі, діаграми та візуалізації для ефективного представлення даних пацієнтів і результатів діагностики. Медичні працівники можуть вводити конкретні симптоми пацієнта, параметри або запити через інтерфейс. Модуль забезпечує зворотний зв’язок у реальному часі, пропонуючи потенційні діагнози або додаткові тести, які можуть знадобитися.

Отже, в цілому алгоритм роботи модуля можна представити в наступному вигляді. Медичний працівник входить у систему через інтерфейс користувача. Відповідні дані пацієнтів витягуються з рівня даних і предстваляються в зрозумілому форматі. Медичний працівник вводить конкретні симптоми або запити. Обчислювальний рівень аналізує вхідні дані за допомогою моделей ML та генерує діагностичні пропозиції. Результати відображаються в інтерфейсі користувача разом із поясненнями та рівнями надійності для кожної пропозиції. Медичний

працівник може провести подальше дослідження або підтвердити діагноз на основі рекомендацій системи.

3.4 Висновки до третього розділу

У розділі «Інформаційна технологія моделей ансамблевого навчання в автоматизованій діагностиці» розроблена ансамблева модель ML виявила себе як ефективний інструмент для підвищення точності та стабільності діагностичних рішень. Запропонована архітектура моделі ансамблевого навчання з використанням підходу stacking на основі моделей SVM, CNN та LSTM.

Модуль аналізу та діагностики, що реалізує автоматизовану постановку діагнозу, реалізує запропоновану ансамблеву модель. Взаємодія з існуючими модулями приватної МІС дозволяє ефективно впроваджувати ансамблеві моделі в реальні умови медичних закладів.

В цілому, розділ визначає ключові кроки та досягнення в розробці та впровадженні інформаційної технології, спрямованої на використання ансамблевих моделей ML в автоматизованій діагностиці захворювань, враховуючи якісний та кількісний аспекти отриманих результатів.

4 ОПИС ОТРИМАНИХ ПРАКТИЧНИХ ТА ТЕОРЕТИЧНИХ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

4.1 Обґрунтування вибору платформи програмного забезпечення

Вибираючи програмну платформу для реалізації моделей ML в медичній діагностиці шкірних захворювань, необхідно враховувати кілька факторів, щоб забезпечити ефективність модуля МІС:

– підтримка відкритого коду;
– простота використання та швидке створення прототипів;
– інтеграція з системами охорони здоров'я;
– масштабованість і продуктивність.

Для розробки архітектури ансамблевого навчання моделей ML необхідно обрати мову програмування та середовище розробки. Зараз більшість об'єктно-орієнтованих мов програмування мають бібліотеки для розробки нейронних мереж. Мережі, написані на мові C++, мають високу швидкість роботи через здатність ефективно управляти пам'яттю та ресурсами системи.

У даній роботі для розробки була використана мова програмування Python, оскільки Python виділяється як переважна мова для ML, оснащена такими бібліотеками та фреймворками, як TensorFlow, NumPy, Keras, PyTorch і scikit-learn.

Python – це інтерпретована, високорівнева, загального призначення мова програмування. Вона надає структури, що дозволяють чітко програмувати на малих і великих масштабах вхідних даних. Python має динамічну систему типів змінних та автоматичне управління пам'яттю. Python підтримує кілька програмних парадигм, включаючи об'єктно-орієнтовану, імперативну, функціональну та процедурну, і має обширний набір стандартних бібліотек [35].

Однією зі стандартних бібліотек є NumPy – це основний пакет для наукового обчислення з використанням Python. Він включає в себе потужний об’єкт N-

вимірного масиву, складні (широкомовні) функції, інструменти для засвоєння коду C/C++ і Fortran, а також цінні функції лінійної алгебри, перетворення Фур'є та генерації випадкових чисел. Крім того, Keras виступає як альтернатива для розробки нейронних мереж, пропонуючи універсальну платформу для цієї мети. Keras – це високорівневий API для нейронних мереж, написаний на Python, який забезпечує просте та швидке створення прототипів, а також повну інтеграцію з іншими популярними бібліотеками глибокого навчання, такими як TensorFlow і Theano. За допомогою Keras є можливість створювати, навчати та розгортати нейронні мережі з мінімальним кодом, що робить його ідеальним вибором як для початківців, так і для досвідчених розробників.

Вибір платформи для розгортання обчислювальних ресурсів для медичної діагностики визначається кількома ключовими факторами. Спочатку, необхідно враховувати важливість забезпечення високого рівня безпеки та конфіденційності медичної інформації. З цієї точки зору, обрана платформа повинна відповідати стандартам безпеки в галузі охорони здоров'я, наприклад, таким як Health Insurance Portability and Accountability Act (HIPAA).

Другим важливим аспектом є необхідність у швидких та продуктивних обчислювальних ресурсах, оскільки медична діагностика може вимагати значної обчислювальної потужності, особливо при використанні глибокого навчання та обробці великих обсягів даних.

Наступним фактором є масштабованість системи, особливо якщо передбачається збільшення обсягу даних чи потреб у обчислювальних ресурсах. З цього погляду, хмарні сервіси, такі як Amazon Web Services (AWS), Azure або Google Cloud Platform (GCP), можуть забезпечити необхідну гнучкість у розширенні обчислювальних ресурсів.

Важливим аспектом є також вартість розгортання та утримання системи, яка може визначатися конкретними умовами використання різних платформ. Наприклад, деякі хмарні платформи можуть пропонувати безкоштовні або доступні

для початку рішення, але вартість може збільшуватися зі зростанням обсягів використання.

З урахуванням цих аспектів, можна зробити висновок, що хмарні платформи, такі як AWS, Azure або GCP, можуть відповідати потребам у розгортанні обчислювальних ресурсів для медичної діагностики, забезпечуючи високу безпеку, продуктивність та масштабованість системи, а також сприяючи спільній роботі та розробці.

У випадку використання Azure, ефективна організація ETL-процесів може бути досягнута за допомогою Databricks кластерів. Databricks – це інтегроване середовище для обробки та аналізу даних, яке надає можливості для розгортання та масштабування кластерів для виконання різноманітних завдань обробки даних.

Крім того, для координації та автоматизації виконання різних завдань в MIC, використання інструменту, який надає моніторинг та планування робочих процесів, може бути надзвичайно корисним. У даному випадку, Apache Airflow може виступати в ролі інструменту для ефективного контролю за виконанням завдань, розподілу ресурсів та планування робочих процесів в системі обробки даних MIC на базі Azure.

Такий підхід дозволяє забезпечити ефективність та автоматизацію обробки медичної інформації, зменшуючи час виконання завдань та забезпечуючи надійність обробки даних у великому обсязі. Враховуючи вищезазначені фактори, вибір такого поєднання програмного забезпечення може виявитися стратегічно доцільним для організації медичної інформації в MIC.

Для забезпечення співпраці та розробки важливо вибрати платформу, яка сприяє легкій спільній роботі та обміну кодом і моделями. Використання GitHub як платформи для спільної роботи не лише спрощує обмін фрагментами коду та повними проєктами, але й допомагає контролю версій, дозволяючи дослідникам та розробникам відстежувати зміни, керувати оновленнями та забезпечувати відтворюваність. Цей спільний синергетичний ефект має глибокий вплив на

розвиток технік медичної діагностики, оскільки практики можуть будувати на існуючій роботі, вдосконалювати моделі та спільно вносити свій внесок у створення точних та ефективних інструментів для діагностики захворювань шкіри.

4.2 Опис вимог до програмного забезпечення

Система діагностики повинна функціонувати як комплекс модулів, кожен з яких відповідає за конкретний аспект аналізу та використовує власну модель даних. Аналітичні модулі мають забезпечити обробку інформації, яка надходить з ІС (джерела даних). Всі компоненти та дані модуля зберігаються на серверах компанії, з можливістю розгортання серверних компонентів на віртуальних серверах. Система гарантує сумісність з актуальними версіями браузерів (Internet Explorer, Mozilla Firefox, Google Chrome, Safari) і забезпечує штатний режим роботи під час робочого часу.

Вимоги до програмного забезпечення користувача включають використання Azure Synapse Analytics для зберігання даних, підтримку веб-браузерів, таких як Internet Explorer, Mozilla Firefox, Google Chrome, Safari, система управління базами даних (СУБД) MS SQL Server версії не нижче 3.23, операційну систему Windows 7/10 і Windows Server 2003, а також можливість розширення моделі даних і структури системи-джерела при необхідності.

Технічне забезпечення сервера передбачає процесор Intel Pentium III 1 ГГц, оперативну пам'ять обсягом 512 МБ Random Access Memory (RAM), та жорсткий диск об'ємом 200 ГБ Hard (magnetic) disk drive (HDD).

Інтерфейс системи реалізовано у вигляді веб-сторінок за допомогою HTML/CSS та плагіну jTable для взаємодії з Asynchronous JavaScript And XML (AJAX) запитами та формування табличного представлення інформації. Взаємодія

клієнтської та серверної частин відбувається через HTTP-запити у форматі JavaScript Object Notation (JSON), що дозволяє оновлювати шаблони документів та використовувати можливості серверної частини для генерації документів, а також перегляду та редагування змісту бази даних.

Система забезпечує зручну навігацію користувачів по доступним ресурсам, використовуючи систему контент-меню та меню з гіперпосиланнями в верхній частині сторінки.

4.3 Експериментальна перевірка отриманих результатів

Першим кроком в виконанні експериментальної перевірки запропонованої моделі ансамблевого навчання є підготовка та обробка вхідних даних. Необхідно ретельно відібрати набір даних, який відображає різноманітність та репрезентативність об'єктів для оптимальної роботи ансамблевого підходу. Після цього важливим етапом є розбиття даних на тренувальний та тестовий набори для оцінки ефективності моделі. Далі необхідно визначити параметри та конфігурації ансамблю, такі як типи базових моделей, їхні гіперпараметри та стратегії об'єднання прогнозів. Після налаштування моделі ансамблю необхідно провести експериментальні обчислення та аналіз результатів, щоб визначити ефективність та потенційні області вдосконалення запропонованого методу.

У сучасному світі існує багато захворювань, які можуть серйозно позначитися на здоров'ї пацієнтів. Серед них виділяють діабет, рак шкіри, захворювання нирок, печінки та серцеві захворювання. Однак у цьому дослідженні особлива увага приділена проблемі діагностики раку шкіри, яка є актуальною та потребує нових підходів.

Експеримент направлено на розробку ансамблю для класифікації меланоми, який буде здатний відрізнити доброякісні (неракові) від злоякісних (ракові) ділянок шкіри.

Для тестування та навчання моделі використовувалися різні зразки зображень, пов'язаних із захворюваннями шкіри. Щоб доповнити цей набір, використовувалася також текстова інформація про атрибути цих зображень, яка була представлена у форматі .csv файлу. Це дозволило покращити якість навчання та точність моделі.

Дані для дослідження було із тестового набору даних з відкритого репозиторію даних Kaggle, що належить корпорації Google [36].

Набір даних характеризується великою кількістю варіацій у вигляді різних типів меланом та їх стадій розвитку. Кожен етап дослідження включав у себе анотацію зображень фахівцями-дерматологами, яка служила основою для визначення правильних класифікацій та навчання моделі. Набір метаданих збережено у файлах CSV, які виглядають наступним чином (рис. 4.1).

image_name	patient_id	sex	age_approx	anatom_site_general_challen	diagnosis	benign_m	target
ISIC_0015719	IP_3075186	female	45	upper extremity	unknown	benign	0
ISIC_0052212	IP_2842074	female	50	lower extremity	nevus	benign	0
ISIC_0068279	IP_6890425	female	45	head/neck	unknown	benign	0
ISIC_0074268	IP_8723313	female	55	upper extremity	unknown	benign	0
ISIC_0074311	IP_2950485	female	40	lower extremity	unknown	benign	0
ISIC_0074542	IP_4698288	male	25	lower extremity	unknown	benign	0
ISIC_0075663	IP_6017204	female	35	torso	unknown	benign	0
ISIC_0075914	IP_7622888	male	30	torso	unknown	benign	0
ISIC_0076262	IP_5075533	female	50	lower extremity	unknown	benign	0
ISIC_0076545	IP_9802602	male	55	upper extremity	unknown	benign	0
ISIC_0076742	IP_2318163	male	75	upper extremity	unknown	benign	0
ISIC_0076995	IP_2235340	female	55	torso	nevus	benign	0
ISIC_0077472	IP_3691360	female	40	torso	unknown	benign	0
ISIC_0077735	IP_1109756	male	70	torso	unknown	benign	0
ISIC_0078703	IP_7279968	male	45	torso	unknown	benign	0
ISIC_0078712	IP_2189124	male	40	lower extremity	unknown	benign	0
ISIC_0079038	IP_5295861	male	70	torso	unknown	benign	0

Рисунок 4.1 – Метадані (інформація про медичні зображення)

Метадані містять наступну інформацію:

– image_name – унікальний ідентифікатор, вказує на ім'я файлу відповідного зображення;

– patient_id – унікальний ідентифікатор пацієнта;
– sex – стать пацієнта (якщо невідомо, буде порожнім);
– age_approx – приблизний вік пацієнта на момент візуалізації;
– anatom_site_general_challenge – розташування ураженої ділянки шкіри;
– diagnosis – детальна інформація про діагностику (тільки тренувальні дані);
– benign_malignant – показник злоякісності зображеного ураження;
– target – двійкова версія цільової змінної.

Зображення надаються в DICOM-форматі, доступ до якого можна отримати за допомогою загальнодоступних бібліотек, наприклад, pydicom. Цей формат включає в себе не лише зображення, але й метадані, що є стандартом для зберігання та обміну медичною інформацією. TFRecord використовується для представлення зображень у розмірі 1024x1024, що сприяє їхньому рівномірному структуруванню та оптимізує обробку даних у контексті дослідження та аналізу.

Ціль аналізу полягає в тому, що модель повинна генерувати ймовірність, в межах від 0.0 до 1.0, що певна область на зображенні є злоякісною (цільовою). У файлах тренування моделі train.csv значення 0 вказують на доброякісну природу, тоді як 1 вказує на злоякісний характер ураження.

Цей підхід до класифікації сприяє створенню точної та чутливої моделі, здатної розрізняти між доброякісними та злоякісними змінами на медичних зображеннях. Використання файлів CSV для навчання надає можливість систематизувати та оптимізувати процес тренування моделі, а також сприяє покращенню її вірогідності в передбаченні класифікацій на нових зображеннях.

Для візуалізації зображень використано наступні інструменти:

– Jupyter notebook;
– Pandas;
– Matplotlib;
– Statistics;
– D3.js.

Ретельний аналіз злоякісних тренувальних зображень дозволяє виявити характеристики, які можуть бути ключовими для правильної класифікації. Різноманітність форм та текстур уражень на цих зображеннях служить основою для навчання моделі розпізнавати варіації у зовнішніх проявах злоякісних уражень.

Аналіз цих зображень також сприяє розумінню можливих викликів та обмежень у завданні класифікації та допомагає вдосконалити навчання моделі для впевнених та точних передбачень. Візуалізація таких зображень представлена на рисунку 4.2.

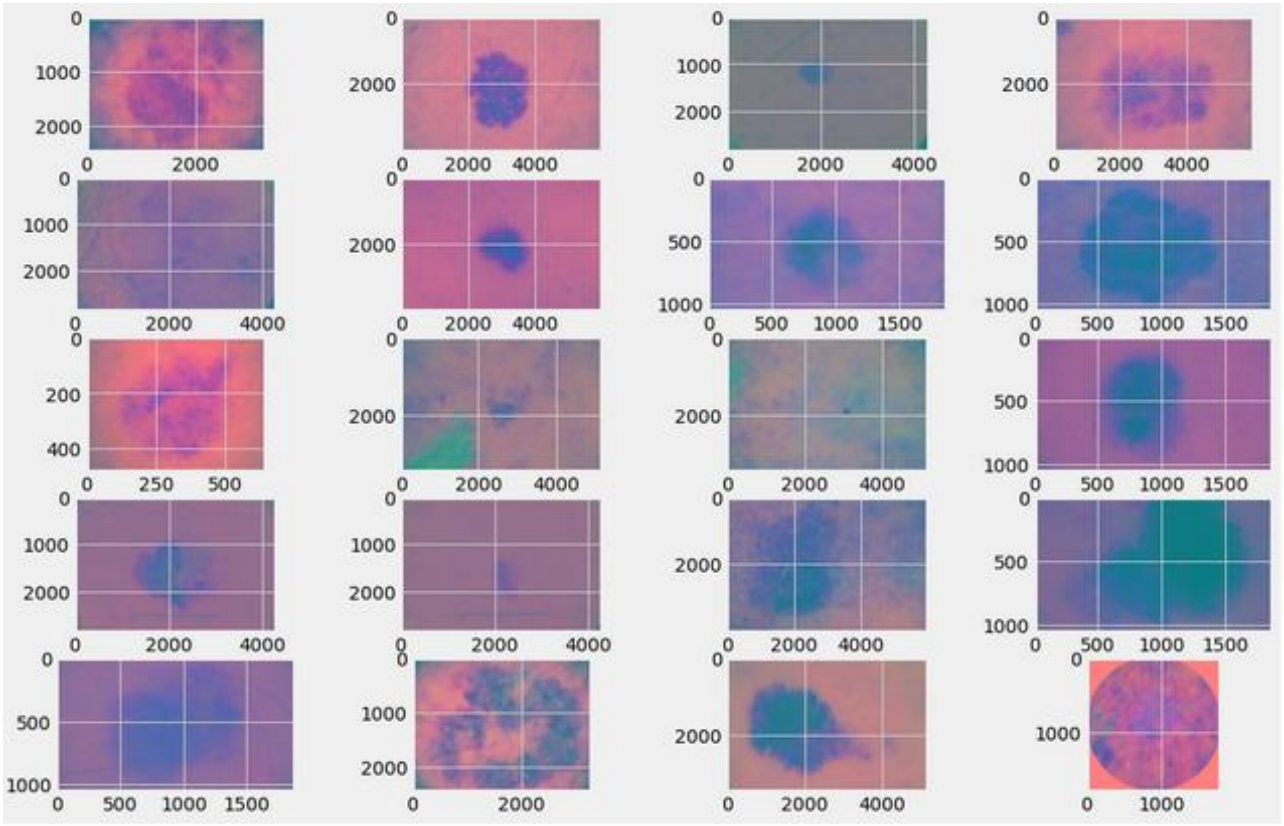

Рисунок 4.2 – Візуалізація злоякісних зображень з тренувального набору даних

Якщо проаналізувати співвідношення доброякісних зображень до злоякісних в тренувальному наборі даних, то можна зробити висновки, що набір даних не є збалансованим (рис.4.3).

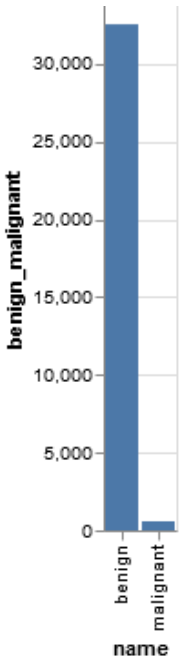

Рисунок 4.3 – Співвідношення доброякісних до злоякісних зображень (train)

Отже, всього є 33126 зображень, з яких 584 є злоякісними, а решта 32542 є доброякісними.

Набір даних містить більшість людей у віці від 40 до 55 років (рис. 4.4).

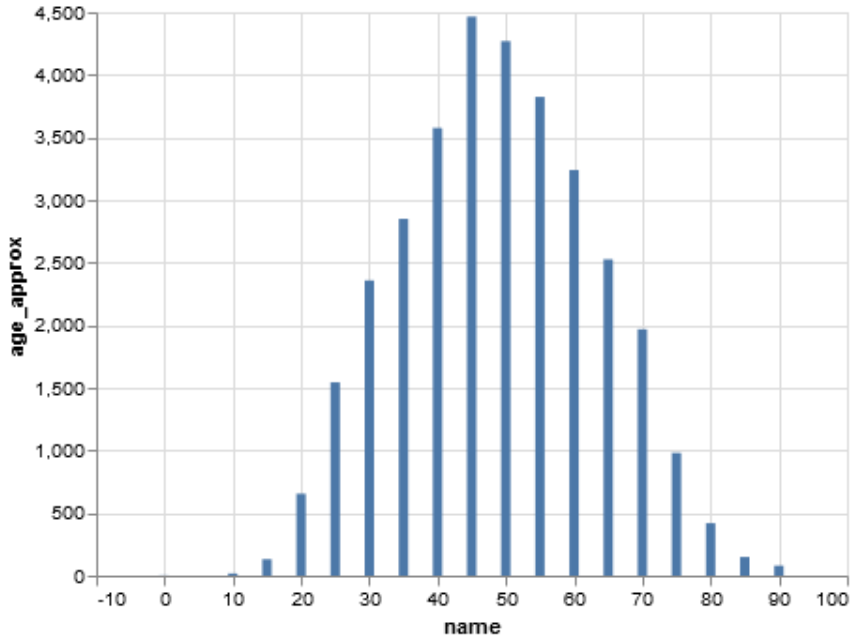

Рисунок 4.4 – Віковий розподіл всіх досліджуваних пацієнтів

Треба звернути увагу на групу осіб з злоякісними ураженнями і вивчти їхню демографічну характеристику в залежності від віку (рис. 4.5).

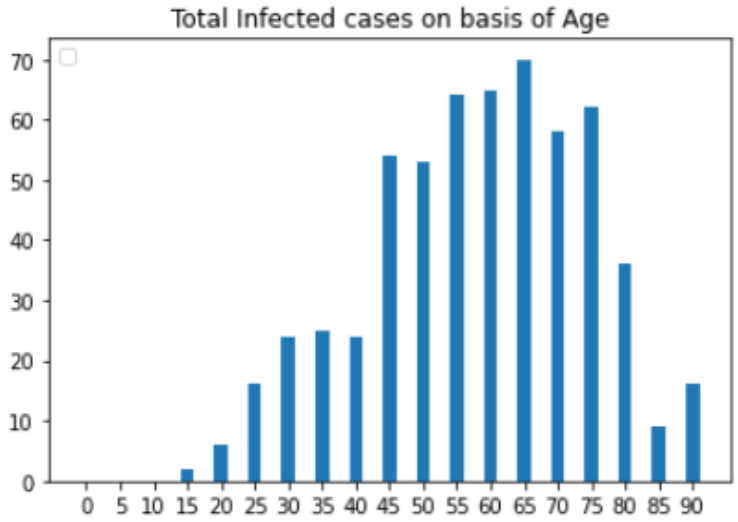

Рисунок 4.5 – Віковий розподіл всіх пацієнтів, у яких виявлено злоякісне ураження шкіри

Більшість випадків злоякісних новоутворень виникає в області тулуба, оскільки саме торс має найбільшу площу поверхні в усьому тілі. Ця особливість анатомії сприяє вищій ймовірності розвитку злоякісних уражень в цій частині організму порівняно з іншими ділянками (рис. 4.6).

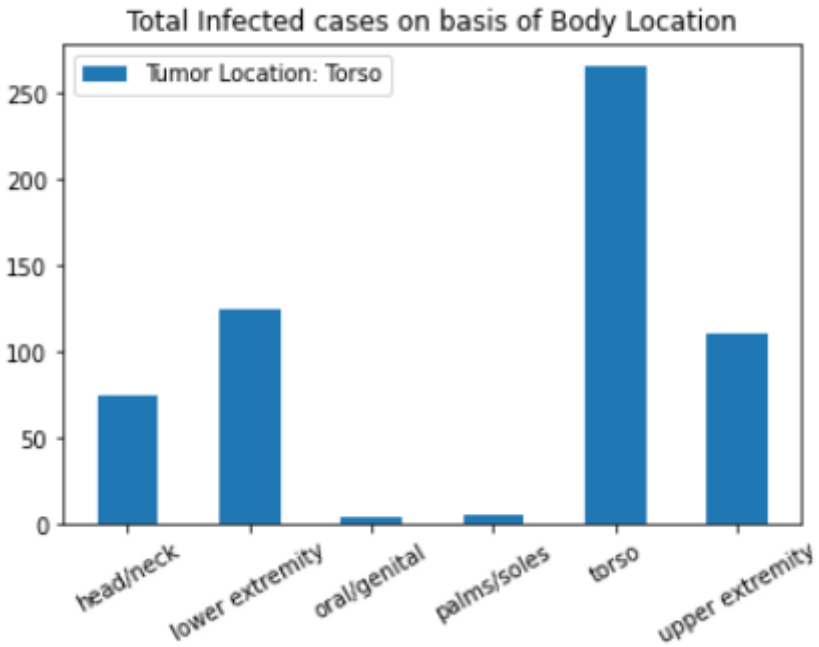

Рисунок 4.6 – Розподіл уражених частин тіла

Попередня обробка даних включає в себе комплекс заходів, спрямованих на очищення та підготовку інформації для подальшого використання в дослідженні. Цей етап включає в себе вирівнювання розширення зображень (розширення змінено до наступного розміру 256 на 256 пікселей), виправлення можливих відхилень, таких як темна віньєтка навколо зображень від мікроскопу (рис. 4.7), агрегацію та стандартизацію даних з метою забезпечення їхньої точності та надійності під час подальшого аналізу.

З цією метою варто структурувати набір даних, групуючи його за пацієнтами, щоб забезпечити детальний та систематичний підхід до спостережень. Це дозволяє зберегти інформацію про динаміку змін та еволюцію ознак на зображеннях пацієнтів, що є важливим для більш точної та інформативної класифікації меланоми.

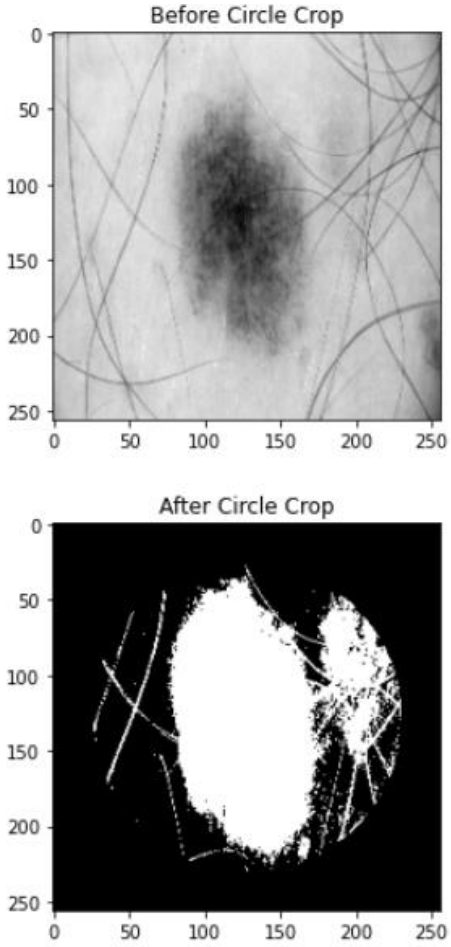

Рисунок 4.7 – Попередня обробка зображень

Аналізуючи зображення пацієнта, лікарі уважно виділяють ознаки меланоми.

Наступним етапом після підготовки даних є тренування моделі, для цього поділимо дані для кожного з методів прогнозування у такому співвідношенні, що 80% даних – тренувальна вибірка, а 20% – тестова.

Налаштування ансамблевих моделей навчання для медичної діагностики шкірних захворювань з використанням SVM, CNN і мереж довгострокової короткочасної пам’яті (LSTM) передбачає визначення гіперпараметрів для кожної базової моделі.

Експериментальний шляхом підібрано та запропоновано гіперпараметри для SVM, CNN та LSTM у контексті класифікації меланоми (табл. 4.1).

Таблиця 4.1 – Налаштування моделей ансамблю

Модель ML	Гіперпараметр	Опис (Впливає)	Можливі Значення	Обране Значення
SVM	Ядро (kernel)	Трансформація вхідних даних	'linear', 'poly', 'rbf', 'sigmoid'	'rbf'
	Регуляризація (C)	Баланс між тренувальною та тестовою помилкою. Контролює ширину маржі	Дійсні позитивні значення	1.0
	Гамма (gamma)	Вплив одного тренувального прикладу. Впливає на форму границі рішення	Дійсні позитивні значення	0.1
	Ваги Класів (class_weight)	Надає різні ваги класам	None, 'balanced'	'balanced'
CNN	Швидкість Навчання	Швидкість оновлення параметрів під час тренування	Дійсні позитивні значення	0.001
	Кількість Фільтрів	Кількість фільтрів в кожному конволюційному шарі	Цілі значення	64
	Розмір Фільтра	Розмір фільтрів в конволюційних шарах	Кортеж цілих чисел	(3, 3)
	Розмір Пулінгу	Розмір вікна пулінгу в шарах пулінгу	Кортеж цілих чисел	(2, 2)
	Відсоток Відкидання	Відсоток вхідних одиниць для видалення під час тренування	Дійсні значення між 0 та 1	0.5

Кінець таблиці 4.1

Модель ML	Гіперпараметр	Опис (Впливає)	Можливі Значення	Обране Значення
LSTM	Кількість Одиниць	Кількість одиниць пам'яті в шарі LSTM	Цілі значення	100
	Відсоток Відкидання	Відсоток вхідних одиниць для видалення під час тренування в шарі LSTM	Дійсні значення між 0 та 1	0.2
	Відсоток Відкидання Рекуренту	Відсоток рекурентних одиниць для видалення під час тренування в шарі LSTM	Дійсні значення між 0 та 1	0.2
	Швидкість Навчання	Швидкість оновлення параметрів під час тренування	Дійсні позитивні значення	0.01

Враховуючи різноманітний набір факторів, які впливають на налаштування моделі, такі як кількість основних алгоритмів у композиції, глибина дерев, умови припинення конструювання дерева, розмір вибірки для кожного основного алгоритму та коефіцієнти регуляризації [14], було вирішено визначити оптимальні значення гіперпараметрів за допомогою методу GridSearchCV. Цей метод реалізує процеси "fitting" та "estimation" з метою вибору найефективніших налаштувань.

Програмна реалізація налаштування моделей ML для ансамблю представлено в додатку А, на рисунках А.1 – А.3 наведено приклад шару метакласифікатора, який приймає прогнози з окремих класифікаторів (SVM, CNN, LSTM) і використовує їх як вхідні функції для навчання метакласифікатора. Використано просту логістичну регресію як метакласифікатор.

Для порівняння результатів моделювання використано наступні параметри. Показник точності (Accuracy) визначає загальну правильність прогнозів, визначаючи співвідношення правильних передбачень до загальної кількості прогнозів. Як розраховуються показники наведено нижче, в формулах (4.1)-(4.3).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}, \quad (4.1)$$

$$Recall = \frac{TP}{(TP+FN)}, \quad (4.2)$$

$$Specificity = \frac{TN}{(TN+FP)}, \quad (4.3)$$

де TP (True Positives) – кількість правильно передбачених позитивних класів;

TN (True Negatives) – кількість правильно передбачених негативних класів;

FP (False Positives) – кількість хибнопозитивних передбачень;

FN (False Negatives) – кількість хибнонегативних передбачень.

Показник точності (Accuracy) визначає загальну правильність передбачень, враховуючи відношення кількості правильних прогнозів до загальної кількості передбачень. Показник чутливості (Recall) відображає частку справді позитивних випадків, які були вірно ідентифіковані, враховуючи відношення справжніх позитивних випадків до суми справжніх позитивних та помилкових негативних випадків. Специфічність (Specificity) відображає частку справді негативних випадків, правильно визначених, як відношення справжніх негативних випадків до суми справжніх негативних та помилкових позитивних випадків.

Результати в таблиці 4.2 було обчислено за допомогою формул (4.1)-(4.3) на основі прогнозів моделі та справжніх міток класу в тестовому наборі даних.

Аналіз даних, представлених у таблиці 4.2, в рамках дослідження моделей ML підтвердив, що певні алгоритми (зокрема, LSTM, SVM, CNN) виявилися ефективними для визначених характеристик, які вказані у таблиці 4.2. Однак інтегрована модель, що об'єднує в собі CNN, SVM і LSTM, продемонструвала загальне покращення точності діагностики. Застосування ансамблювання моделей призвело до досягнення значення AUC-ROC на рівні 0.95, що перевершує результат оцінки, отриманий від окремої моделі діагностики за допомогою лише CNN в ML, що становить 0.89.

Таблиця 4.2 – Результати, отримані в ході дослідження обраних моделей

Модель	Точність	Чутливість	Специфічність	AUC-ROC
LSTM	0.88	0.92	0.86	0.91
SVM	0.82	0.88	0.78	0.86
CNN	0.89	0.93	0.87	0.92
CNN-SVM-LSTM	0.95	0.96	0.93	0.95

Важливо відзначити, що запропонована ансамбльована модель також надає інтерпретовані пояснення своїх рішень, сприяючи клінічному прийняттю результатів моделювання. Це виявляється значущим для медичних фахівців, оскільки це допомагає їм розуміти прогнози та віддавати їм довіру.

4.4 Висновки до четвертого розділу

Отримані результати експерименту вказують на перспективність використання ансамблевих моделей для розпізнавання меланоми та демонструють їх високу точність в порівнянні з іншими підходами. Дані отримані під час дослідження можуть служити основою для подальших робіт у сфері комп'ютерного зору та діагностики захворювань шкіри з використанням ансамблювання моделей ML.

Обираючи платформу для розробки, визначено мову Python як оптимальний вибір для реалізації моделей ML, оскільки ця мова володіє потужними бібліотеками, такими як TensorFlow та PyTorch. Також, використання хмарних платформ, таких як AWS, Azure забезпечує масштабованість та надійність системи.

Визначивши вимоги до програмного забезпечення, розглянуто питання сумісності з різними браузерами та операційними системами.

Аналіз практичного використання моделей ансамблевого навчання в процесі автоматизованої діагностики включає в себе оцінку їхньої точності, ефективності та взаємодії з фахівцями у сфері медицини.

ВИСНОВКИ

У результаті виконання магістерської роботи було проведено комплексне дослідження моделей, методів та технологій автоматизованої діагностики захворювань, зокрема, зосереджено увагу на ансамблюванні моделей ML для реалізації автоматизованої діагностики злоякісних захворювань шкіри, на прикладі меланоми. Але при цьому постало питання поєднання сильних сторін різних моделей ML та мінімізація їх недоліків.

В ході виконання магістерської роботи проведено обґрунтування актуальності дослідження, наведено оцінку сучасного стану дослідження, проаналізовано існуючі моделі, методи та технології в автоматизованій діагностиці захворювань, обґрунтована мета розробки ансамблевої моделі ML, висвітлено проблеми та виклики, пов’язані з реалізацією ансамблю, визначено цілі для досягнення мети магістерської роботи.

Далі визначено поняття автоматизованої діагностики, та діагнозу в цілому, що дозволяє представити діагноз у математичному представлені. Результатом представлення діагнозу у математичній формі є вектор даних, який містить інформацію про стан і симптоми пацієнта. Обґрунтовано доцільність розробки ETL процесу для обробки медичних даних. Визначено та проаналізовано ключові аспекти, такі як вибір моделей, їхні переваги та недоліки, а також важливість підбору параметрів для досягнення найкращих результатів в медичній діагностиці.

В ході даної роботи було розроблену схему архітектури моделі ансамблевого навчання з використанням підходу stacking на основі моделей SVM, CNN та LSTM. Перед проектуванням виконано оцінку можливостей та обмеження впровадження ансамблевої моделі в МІС. Визначено розмір та тип обчислювальних ресурсів для реалізації запропонованої архітектури ансамблевої моделі.

На основі проведених практичних досліджень було виконано порівняльний аналіз отриманих результатів діагностики злоякісних уражень шкіри на окремих моделях ML та комбінації декількох моделей ML, а саме – SVM, CNN та LSTM. Для кожної моделі підібрано оптимальні значення гіперпараметрів. Для апробації результатів дослідження виконано експериментальну перевірку, першим кроком якої стала підготовка та обробка вхідних даних. Наступним кроком після підготовки даних є тренування моделей, для цього виконано налаштування кожної моделі та створення ансамблю та метакласифікатора за допомогою мови програмування Python. Результати перевірки показали досягнення значення AUC-ROC на рівні 95%, що перевершує результат оцінки, отриманий від окремих моделей діагностики.

Результати кваліфікаційної роботи було опубліковано в статті «Дослідження ансамблювання моделей Machine Learning в медичній діагностиці» у збірнику «АСУ і прилади автоматики» [37]. За тематикою атестаційної роботи було проведено публікацію тез доповіді «Аналітика великих даних у службі сховища на платформі Microsoft Azure» на дванадцятій міжнародній науково-технічній конференції «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління» [38].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Методичні вказівки щодо розробки та оформлення кваліфікаційної роботи (для студентів усіх форм навчання другого (магістерського) рівнявищої освіти спеціальності 122 Комп'ютерні науки освітньо-професійної програми «Інформаційні управляючі системи та технології») / Упоряд.: ПетровК.Е., Левикін В.М., Чалий С.Ф., Євланов М.В., Саєнко В.І., Міхнов Д.К., Міхнова А.В., Чала О.В. – Харків: ХНУРЕ, 2021. – 30 с.

2. ДСТУ 3008:2015. Інформація та документація. Звіти у сфері науки і техніки. Структура та правила оформлювання, Чинний від 22.06.2015. Київ: ДП «УкрНДНЦ», 2016, 26 с.

3. ДСТУ 8302:2015 «Інформація та документація. Бібліографічне посилання. Загальні положення та правила складання».

4. Дудка В. В. Переваги приватної медицини та вигоди держави у підтримці розвитку приватного сектора системи охорони здоров'я (до проблеми державного регулювання здоровоохоронної сфери) // Електронне наукове фахове видання «Державне управління: удосконалення та розвиток». 2012. № 1. URL: http://nbuv.gov.ua/UJRN/Duur_2012_1_4 (дата звернення 21.11.2023).

5. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin, Wiley Publishing, Inc., Indianapolis, 2004. 212 c.

6. Електронна система охорони здоров'я eHealth: те, що мають знати лікарі та пацієнти URL: www.umj.com.ua/uk/publikatsia-132874-elektronna-sistema-ohoroni-zdorov-ya-ehealth-te-shho-mayut-znati-likari-ta-patsiyenti (дата звернення 21.09.2023).

7. Медичні інформаційні системи: огляд можливостей та приклади використання. URL: https://evergreens.com.ua/ru/articles/medical-information-systems.html (дата звернення 28.11.2023).

8. Методологія вивчення рівня впровадження інформатизації в систему охорони здоров'я України : метод. рекомендації / Г. О. Слабкий та ін. Київ : МОЗУ, 2014. 20 с.

9. eHealth – що це та як підключитися? URL: https://blog.h24.ua/uk/ehealth-shho-tse-ta-yak-pidklyuchytysya/ (дата звернення 28.11.2023).

10. Класифікація медичних інформаційних систем. URL: https://medic.studio/tehnologiimeditsine-informatsionnyie/klassifikatsiya-meditsinskih-informatsionnyih58978.html (дата звернення 29.11.2023).

11. Прасоленко О. В., Ткаченко І. О. Основи теорії систем і системний аналіз: Харків. нац. ун-т міськ. госп-ва імені О. М. Бекетова. – Харків : ХНУМГ ім. О. М. Бекетова, 2018. – 88 с

12. Luigi B., Sigcha L., Rodriguez Daniel., Olmo G. Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. Artificial Intelligence in Medicine, vol. 135, P. 102459, 2023. DOI:10.1016/j.artmed.2022.102459

13. Zhang Q., Liu Y., Liuz G., Zhao G., Qu Z., Yang W. An automatic diagnostic system based on deep learning, to diagnose hyperlipidemia. 2019. P. 637-645. DOI: 10.2147/DMSO.S198547.

14. Myronova G. Digitalization of healthcare in Ukraine: legal support of public and private interests. Entrepreneurship, Economy and Law, 2023, P. 40–47, DOI: https://doi.org/10.32849/2663-5313/2022.3.05

15. Terrizzano I, Schwarz P, Roth M, Colino JE. Data wrangling: The challenging journey from the wild to the lake. In: CIDR 2015 - 7th Biennial Conference on Innovative Data Systems Research. 2015.

16. Aghajani E, Nagy C, Vega-Marquez OL, Linares-Vasquez M, Moreno L, Bavota G, et al. Software Documentation Issues Unveiled. Proc - Int Conf Softw Eng. 2019. P. 199–210.

17. Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E, et al. Provenance solutions for medical research in heterogeneous IT-infrastructure: An implementation roadmap. Stud Health Technol Inform. 2019. P. 298–302.

18. Box G., Pelham E., Draper R. Empirical Model-Building and Response Surfaces. New York, NY: Wiley, 2007. 119 p.

19. Kluegl P, Toepfer M, Beck P-D, Fette G, Puppe F. UIMA Ruta: rapid development of rule-based information extraction applications. Nat Lang Eng. 2016. P. 22–40.

20. Bell, R.M. and Koren, Y. Lessons from the Netflix prize challenge. SIGKDD Explorations, 2007, P.75-79.

21. Jani R., Shariful Islam Shanto Md., Mohsin Kabir Md., Saifur Rahman Md., Mridha M. F. Heart disease prediction and analysis using ensemble architecture. 2022 International Conference on Decision Aid Sciences and Applications (DASA), 2022. DOI:10.1109/dasa54658.2022.9765237

22. Mahajan P., Uddin S., Hajati F., Moni M. A. Ensemble learning for disease prediction: A Review. Healthcare. vol. 11, no. 12, 2023. P. 1808. doi:10.3390/healthcare11121808

23. Ali R., Hardie R. C., Narayanan Narayanan B., De Silva S. Deep learning ensemble methods for skin lesion analysis towards melanoma detection. 2019 IEEE National Aerospace and Electronics Conference (NAECON), 2019. DOI:10.1109/naecon46414.2019.9058245

24. Jain G., Mittal D., Thakur D., Mittal M. K. A deep learning approach to detect covid-19 coronavirus with X-ray images. Biocybernetics and Biomedical Engineering. vol. 40, no. 4, 2020. P. 1391–1405. doi:10.1016/j.bbe.2020.08.008

25. Noor, M.B. et al. Application of deep learning in detecting neurological disorders from Magnetic Resonance Images: A survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia. 2020. 11 p. doi:10.1186/s40708-020-00112-2.

26. Jeena, R.S, Kumar, S. Stroke prediction using SVM. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) 2016. doi:10.1109/iccicct.2016.7988020.

27. Apostolopoulos I. D., Mpesiana T. A. COVID-19: Automatic detection from X-ray images utilizing transfer learning with Convolutional Neural Networks. Physical and Engineering Sciences in Medicine, vol. 43, no. 2, 2020. P. 635–640. doi:10.1007/s13246-020-00865-4

28. Ballin A, Karlinsky L, Alpert S, Hasoul S, Ari R, Barkan E. A region based convolutional network for tumor detection and classification in Breast Mammography. Deep Learning and Data Labeling for Medical Applications, 2016. P. 197–205. doi:10.1007/978-3-319-46976-8_21

29. Anavi Y., Kogan I., Gelbart E., Geva O., Greenspan H. Visualizing and enhancing a deep learning framework using patients age and gender for chest X-ray image retrieval. Medical Imaging 2016: Computer-Aided Diagnosis, 2016. doi:10.1117/12.2217587

30. Hassan M., Ali S., Alquhayz H., Safdar K. Developing Intelligent Medical Image Modality Classification system using Deep Transfer Learning and LDA. Scientific Reports, vol. 10, no.1, 2020. doi:10.1038/s41598-020-69813-2.

31. Abbas A., Abdelsamea M. M., Gaber M. M. Classification of covid-19 in chest x-ray images using DeTraC deep convolutional neural network. Applied Intelligence, vol. 51, no. 2, 2020. P. 854–864. doi:10.1007/s10489-020-01829-7

32. Cabitza F., Rasoini R., Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 P. 418–517. https://doi.org/10.1001/jama.2017.7797

33. Ghassemi M., Oakden-Rayner L., Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021. P. 745–50. https://doi.org/10.1016/s2589-7500(21)00208-9.

34. Markus A. F., Kors J.A., Rijnbeek P.R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the

terminology, design choices, and evaluation strategies. 2021. https://doi.org/10.1016/j.jbi.2020.103655.

35. Beazley D., Python Essential Reference. Pearson Education, Limited, 2021. P. 717–724.

36. Siim-ISIC melanoma classification. Kaggle, URL: https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/162486 (дата звернення 28.11.2023).

37. Буцька А. С., Панфьрова І. Ю., Дослідження ансамблювання моделей machine learning в медичній діагностиці // «АСУ і прилади автоматики», №179, 2023. С. 50-57. DOI: 10.20837/0135-1710.2023.179.050

38. Панфьорова І.Ю., Левченко А.С. Аналітика великих даних у службі сховища на платформі Microsoft Azure // Дванадцята міжн. наук.-техн конф. «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління». Збірник наукових праць. Баку – Харків – Жиліна: ХНУРЕ. 2022. С.94.