

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ
(повна назва інституту/факультету)
кафедра БІОМЕДИЧНОЇ КІБЕРНЕТИКИ
(повна назва кафедри)

«На правах рукопису»
УДК _____

«До захисту допущено»
В.о.завідувача кафедри БМК

_____ Світлана АЛХІМОВА
(підпис) (ініціали, ПРІЗВИЩЕ)
“ ____ ” _____ грудня 2025р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Комп'ютерні технології в біології та
медицині»
зі спеціальності 122 «Комп'ютерні науки»
на тему: _____ Система диференціальної діагностики захворювань легень
методами машинного навчання

Виконала: студентка II курсу, групи ЗК-41мп

ЛУТЧЕНКО ВІКТОРІЯ ГЕННАДІЇВНА
(прізвище, ім'я, по батькові)


(підпис)

Науковий керівник: *професор каф. біомедичної кібернетики*
(БМК), д.б.н., к.т.н., проф. Настенко Євген Арнольдович
(посада, науковий ступінь, вчене звання, прізвище, ім'я, по ініціали)


(підпис)

Консультант з розділів магістерської дисертації:

(назва розділу) (посада, вчене звання, науковий ступінь, прізвище, ініціали)

(підпис)

Рецензент: *доцент каф. біомедичної інженерії, доцент, д.т.н.*
Шликов Владислав Валентинович
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали)

(підпис)

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних посилань.

Студентка _____
(підпис)

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет біомедичної інженерії
Кафедра біомедичної кібернетики

Рівень вищої освіти – другий (магістерський)

Спеціальність – 122 «Комп'ютерні науки»

Освітньо-професійна програма «Комп'ютерні технології в біології та медицині»

ЗАТВЕРДЖУЮ

В.о.завідувача кафедри БМК

_____ Світлана АЛХІМОВА

« ____ » листопада ____ 2025 р.

ЗАВДАННЯ
на магістерську дисертацію студентці
ЛУТЧЕНКО ВІКТОРІІ ГЕННАДІЇВНІ

(прізвище, ім'я, по батькові)

1. Тема дисертації **Система диференціальної діагностики захворювань**
легень методами машинного навчання

науковий керівник дисертації

Настенко Євген Арнольдович, д.б.н., к.т.н., професор

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «03» листопада 2025 р. №4751-с_

2. Термін подання студентом дисертації **20-22 листопада 2025 року**

3. Об'єкт дослідження *Процес диференціальної діагностики захворювань легень за даними комп'ютерної томографії.*

4. Вихідні дані *Набір зрізів комп'ютерної томографії.*

5. Перелік завдань, які потрібно розробити *1) аналіз сучасних підходів та аналогів, 2) розробка методики радіомічного аналізу, 3) виконання міжсайтової гармонізації ознак, 4) реалізація моделей машинного навчання, 5) розробка програмного застосунку.*

6. Орієнтовний перелік графічного (ілюстративного) матеріалу *28 рисунків, 4 таблиці, презентація з захисту МД на 16 слайдах.*

7. Орієнтовний перелік публікацій *1 стаття у фаховому журналі категорії Б, 1 стаття у фаховому виданні категорії А ("Scopus").*

8. Консультанти розділів дисертації^{1*} не передбачено.

^{1*} Якщо визначені консультанти. Консультантом не може бути зазначено наукового керівника магістерської дисертації.

9. Дата видачі завдання **28 серпня 2025р.**

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів МД	Примітка
1	Отримати завдання на МД	До 28.08.2025	виконано
2	Завершення виконання практичної частина МД	До 25.10.2025	виконано
3	Завершення оформлення розділів МД (Вступ, Основні розділи та висновки до них, Загальні висновки, Список використаних джерел)	До 24.11.2025	виконано
4	Апробація та публікація результатів дослідження МД (отримати підтвердження про прийняття до друку)	До 01.12.2025	виконано
5	Перевірка МД науковим керівником	24.11.2025	виконано
6	Подання в електронному вигляді МД на перевірку нормоконтролера	До 01.12.2025	
7	Подання в електронному вигляді МД на перевірку подібності Strike Plagiarism com. Отримати позитивний звіт подібності.	До 05.12.2025	
8	Підготовка Експертної оцінки до звіту подібності.	До 10.12.2025	
9	Отримати відгук наукового керівника	10.12.2025	
10	Надати на кафедру пакет документів в паперовому та електронному вигляді (МД, відгук керівника, звіт подібності, Експертної оцінки до звіту подібності)	10-11.12.2025	
11	Отримати допуск до захисту МД в ЕК та направлення до рецензента (засідання кафедри)	Засідання кафедри	
12	Подання МД рецензенту. Отримати на надання рецензію до ЕК / на кафедру	До 19.12.2025	
13	Подання супровідного пакету документів по МД до захисту в ЕК ²	До 19.12.2025	
14	Захист МД в ЕК	22 – 26.12.2025	

Студент



(підпис)

Вікторія ЛУТЧЕНКО

(ім'я, ПРІЗВИЩЕ)

Науковий керівник



(підпис)

Євген НАСТЕНКО

(ім'я, ПРІЗВИЩЕ)

Нормоконтролер

(підпис)

Галина КОРНІЄНКО

(ім'я, ПРІЗВИЩЕ)

² не пізніше ніж за один тиждень до затвердженої дати захисту МД в ЕК

РЕФЕРАТ

Магістерська дисертація за темою «Система диференціальної діагностики захворювань легень методами машинного навчання» виконана студенткою кафедри біомедичної кібернетики ФБМІ Лутченко Вікторією Геннадіївною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині» та складається з: вступу; 5 розділів (*Аналіз літературних джерел та програмних аналогів систем диференціальної діагностики захворювань легень, Засоби реалізації системи диференціальної діагностики, Обґрунтування вибору методів та інструментів, Програмна реалізація та методика роботи програмного застосунку, Обговорення отриманих результатів*), розділу стартап проєкту, висновків до кожного з цих розділів; загальних висновків та списку використаних джерел, який налічує 80 джерел. Загальний обсяг роботи 135 сторінок.

Актуальність теми. Своєчасна та коректна диференціальна діагностика захворювань легень за даними комп'ютерної томографії органів грудної клітини є однією з ключових задач сучасної променевої діагностики. Її ускладнюють значне перекриття візуальних патернів ураження між різними нозологіями, а також міжсайтова варіабельність КТ-даних, зумовлена відмінностями між томографами та протоколами сканування. Остання, в свою чергу, призводить до зміни розподілів інтенсивностей і текстурних характеристик зображень, що ускладнює побудову узагальнюваних моделей та коректне перенесення напрацьованих рішень між різними клініками без додаткової адаптації даних.

Мета і задачі дослідження.

Метою роботи є підвищення об'єктивності та відтворюваності процесу диференціальної діагностики захворювань легень за даними комп'ютерної томографії шляхом використання методів гармонізації міжсайтової

варіабельності радіомічних ознак та побудови моделей машинного навчання у складі локальної програмної системи підтримки прийняття рішень.

Сформульовано наступні **задачі**:

1. Проаналізувати сучасні підходи та програмні засоби діагностики захворювань легень для виявлення їхніх обмежень.
2. Розробити методику попередньої обробки КТ-зображень та виокремлення інтерпретованих текстурних ознак.
3. Програмно реалізувати метод гармонізації для усунення міжсайтової варіабельності даних.
4. Побудувати та валідувати моделі машинного навчання для диференціальної діагностики та виявлення синдрому зникаючої легені.
5. Розробити локальну програмну систему (MVP) для автоматизованого аналізу знімків та формування звітів.

Об'єкт дослідження. Процес диференціальної діагностики захворювань легень за даними комп'ютерної томографії.

Предмет дослідження. Моделі машинного навчання та алгоритмічні засоби гармонізації радіомічних ознак для класифікації захворювань легень на зображеннях комп'ютерної томографії.

Методи дослідження. Цифрова обробка зображень, текстурний (радіомічний) аналіз, машинне навчання, міжсайтова гармонізація текстурних ознак.

Практичне значення одержаних результатів. Магістерська дисертація виконана на реальних даних Національного інституту фтизіатрії і пульмонології ім. Ф.Г. Яновського, що були надані згідно договору про співпрацю від 5 січня 2023 року, і куди були впроваджені попередні результати актом впровадження від 26 грудня 2024 року. Отримані в даній роботі результати і програмний застосунок також впроваджені в роботу Інституту (акт впровадження від ...).

Також робота виконується в рамках ініціативної наукової роботи кафедри НДР д/р № 0123U100866 «Методи та моделі ідентифікації станів об'єктів в задачах прийняття медичних рішень» (06.02.2023р.) [80].

Робота містить окремі результати власних наукових публікацій у співавторстві [28; 74].

Апробація результатів дослідження. Не заплановано.

Публікації. За результатами виконаної роботи була опублікована 1 наукова стаття у фаховому журналі категорії Б:

1. Лутченко В. Г. Ефективність алгоритмів машинного навчання для класифікації змін структури легень у постковідних та гострих стадіях COVID-19 / В. Г. Лутченко, В. О. Бабенко, Є. А. Настенко, М. І. Линник // Біомедична інженерія і технологія. – Київ, 2024. - №15/2024. – С. 27–35 (стаття). – DOI: <https://doi.org/10.20535/.2024.15.306763> (дата звернення 15.11.2025) [28].

Була опублікована 1 наукова стаття у фаховому виданні категорії А (індексована в Scopus):

1. Ie.A. Nastenko, Diagnosis of COVID-19-Associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: A Review of Methods and Future Research Directions / Ie.A. Nastenko, M.I. Linnik, M.O. Honcharuk, I.V. Davydovych, V.H. Lutchenko, V.O. Babenko // Innovative Biosystems & Bioengineering. – 2025. – Vol. 9, № 3. – P. 16-27 (стаття). DOI: <https://doi.org/10.20535/ibb.2025.9.4.325335> (дата звернення 27.11.2025) [74].

Ключові слова. Комп'ютерна томографія, радіоміка, текстурні ознаки, міжсайтова гармонізація, машинне навчання, диференціальна діагностика, COVID-19, пневмонія, long-COVID, синдром зникаючої легені.

Бібліографічний опис МД

Лутченко В. Г. Система диференціальної діагностики захворювань легень методами машинного навчання : магістерська дис. : 122 Комп'ютерні науки / Лутченко Вікторія Геннадіївна. – Київ, 2025. – 135 с

ABSTRACT

The master's thesis entitled “ System for Differential Diagnosis of Pulmonary Diseases using Machine Learning” was prepared by *Viktoriia Lutchenko*, a student of the Department of Biomedical Cybernetics, Faculty of Biomedical Engineering, *specialty 122 «Computer Science» within the educational and professional program «Computer Technologies in Biology and Medicine»*. The thesis comprises an introduction; 5 chapters (*Analysis of literary sources and software analogues of systems for differential diagnosis of lung diseases, Tools for implementing the differential diagnosis system, Justification of the choice of methods and tools, Software implementation and operating methodology of the application, Discussion of the obtained results*); a chapter devoted to the startup project; conclusions to each chapter; general conclusions; a list of references comprising 80 sources; and appendices. The total length of the thesis is 135 pages.

Relevance of the Topic. Timely and accurate differential diagnosis of pulmonary diseases based on chest computed tomography (CT) data is one of the key tasks of modern radiology. It is complicated by substantial overlap of visual patterns of lung involvement across different nosological entities, as well as by inter-site variability of CT data caused by differences between scanners and acquisition protocols. The latter, in turn, leads to shifts in the distributions of image intensities and textural characteristics, which hinders the development of generalizable models and the reliable transfer of established solutions between clinical sites without additional data adaptation.

Research Aim and Objectives.

The aim of this work is to enhance the objectivity and reproducibility of the process of differential diagnosis of pulmonary diseases based on computed tomography data by employing methods for harmonizing inter-site variability of radiomic features and by developing machine learning models within a local decision-support software system.

The following **tasks** were defined:

1. To analyse contemporary approaches and software tools for the diagnosis of pulmonary diseases in order to identify their limitations.
2. To develop a methodology for the preprocessing of CT images and extraction of interpretable textural features.
3. To implement a harmonization method in software to mitigate inter-site variability of the data.
4. To construct and validate machine learning models for differential diagnosis and detection of vanishing lung syndrome.
5. To develop a local software system (MVP) for automated image analysis and report generation.

Research Object. The process of differential diagnosis of pulmonary diseases based on computed tomography data.

Research Subject. Machine learning models and algorithmic tools for harmonizing radiomic features for the classification of pulmonary diseases on computed tomography images.

Research Methods. Digital image processing, textural (radiomic) analysis, machine learning, inter-site harmonization of textural features.

Practical Significance of the Obtained Results. The Master's thesis was carried out using real-world data from the Yanovsky National Institute of Phthisiology and Pulmonology, provided under a cooperation agreement dated 5 January 2023, to which preliminary results were introduced by an implementation act dated 26 December 2024. The results obtained in this work, as well as the developed software application, have also been implemented in the Institute's practice (implementation act dated ...).

The study is additionally conducted within the framework of the departmental initiative research project R&D № 0123U100866 «Methods and models for identifying object states in medical decision-making tasks» (06.02.2023) [80].

The work contains separate results of the author's own scientific publications co-authored [28; 74].

Approbation of the research results. Not planned.

Publications. Based on the results of the present work, one research article has been published in a category B peer-reviewed journal:

1. V. H. Lutchenko. Efficiency of machine learning algorithms for classifying lung structural changes in post-COVID-19 and acute stages of COVID-19 / V. H. Lutchenko, V. O. Babenko, Ie. A. Nastenko, M. I Lynnyk // Biomedical Engineering and Technology. – Kyiv, 2024. – No. 15/2024. – P. 27–35 (article). DOI: <https://doi.org/10.20535/.2024.15.306763> (accessed 15 November 2025) [28].

One scientific article was published in a category A peer-reviewed journal (indexed in Scopus):

1. Ie.A. Nastenko. Diagnosis of COVID-19-Associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: A Review of Methods and Future Research Directions / Ie.A. Nastenko, M.I. Linnik, M.O. Honcharuk, I.V. Davydovych, V.H. Lutchenko, V.O. Babenko // Innovative Biosystems & Bioengineering. – 2025. – Vol. 9, № 3. – P. 16-27 (article). DOI: <https://doi.org/10.20535/ibb.2025.9.4.325335> (accessed 27 November 2025) [74].

Keywords. Computed tomography, radiomics, textural features, inter-site harmonization, machine learning, differential diagnosis, COVID-19, pneumonia, long COVID, vanishing lung syndrome.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ	14
ВСТУП.....	17
РОЗДІЛ 1 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ ТА ПРОГРАМНИХ АНАЛОГІВ СИСТЕМ ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЛЕГЕНЬ	21
1.1 Клінічний контекст задачі та постановка проблеми	21
1.2 Огляд програмних аналогів на ринку	22
1.2.1 <i>LungQ та CAD4COVID-CT (Thirona)</i>	22
1.2.2 <i>AI-Rad Companion Chest CT (Siemens Healthineers)</i>	23
1.2.3 <i>LDA+, LTA (Imbio)</i>	24
1.2.4 <i>Aidoc</i>	26
1.2.5 <i>contextflow ADVANCE Chest CT (INSIGHTS, SEARCH)</i>	26
1.2.6 <i>dragonfly</i>	28
1.3 Методологічний контур на основі аналізу аналогів.....	28
Висновок з розділу 1	29
РОЗДІЛ 2 ЗАСОБИ РЕАЛІЗАЦІЇ СИСТЕМИ ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ.....	31
2.1 Дані та структура датасету.....	31
2.2 Теоретичні основи текстурних ознак.....	33
2.2.1 <i>Gray Level Co-occurrence Matrix</i>	34
2.2.2 <i>Gray Level Difference Statistics</i>	36
2.2.3 <i>Gray Level Run Length Matrix</i>	38
2.2.4 <i>Local Binary Pattern</i>	41
2.3 RELIEF як метод гармонізації	42
2.4 Моделі машинного навчання	45

	11
2.4.1	<i>Мультиноміальна логістична регресія</i> 45
2.4.2	<i>Випадковий ліс</i> 47
2.4.3	<i>Метод опорних векторів</i> 49
2.4.4	<i>Метод k-найближчих сусідів</i> 51
2.4.5	<i>XGBoost</i> 52
2.4.6	<i>LightGBM</i> 54
2.5	Метрики оцінювання 55
	Висновок з розділу 2 58

РОЗДІЛ 3 ОБҐРУНТУВАННЯ ВИБОРУ МЕТОДІВ ТА ІНСТРУМЕНТІВ 59

3.1.	Обчислювальне середовище та програмний стек..... 59
3.2	Архітектура конвеєра обробки даних 60
3.3	Вибір та параметризація текстурних ознак 61
3.3.1	<i>Gray Level Co-occurence Matrix</i> 61
3.3.2	<i>Gray Level Difference Statistics</i> 62
3.3.3	<i>Gray Level Run Length Matrix</i> 63
3.3.4	<i>Local Binary Pattern</i> 64
3.4	Обґрунтування вибору та конфігурації методу RELIEF для гармонізації ознак 65
3.5	Вибір моделей класифікації 66
3.5.1	<i>Мультиноміальна логістична регресія</i> 66
3.5.2	<i>Випадковий ліс</i> 67
3.5.3	<i>Метод опорних векторів</i> 68
3.5.4	<i>Метод k-найближчих сусідів</i> 68
3.5.5	<i>XGBoost</i> 69
3.5.6	<i>LightGBM</i> 69
3.6	Запобігання витоку даних і схеми підвибірок 70
3.7	Захист даних і приватність..... 72

РОЗДІЛ 4 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА МЕТОДИКА РОБОТИ ПРОГРАМНОГО ЗАСТОСУНКУ	75
4.1 Попередня обробка даних	75
4.1.1 Сегментація масок легень.....	75
4.1.2 Обчислення текстурних ознак.....	76
4.1.3 Гармонізація між сайтами.....	78
4.1.4 Стратегія формування навчальної, валідаційної та тестової підвбірок.....	79
4.2 Архітектура програмного застосунку	80
4.3 Демонстрація функціоналу програмного застосунку	81
4.3.1 Мультикласова класифікація.....	82
4.3.2 Визначення ознак СЗЛ	85
4.3.3 Генерація звіту.....	87
РОЗДІЛ 5 ОБГОВОРЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ.....	91
5.1 Виконання вимог технічного завдання.....	91
5.2 Аналіз отриманих результатів моделей.....	93
5.2.1 Мультикласифікація.....	93
5.2.2 Бінарна класифікація	95
5.2.3 Пропозиції щодо покращення результатів	96
5.3 Порівняння з аналогами та практична цінність	98
5.4 Обмеження і зовнішня валідність	99
РОЗДІЛ 6 СТАРТАП-ПРОЄКТ З ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ ЛЕГЕНЬ «LUNGSCAN».....	103
6.2 Бізнес-модель та цільові споживачі проєкту.....	104
6.2.1 Сегменти споживачів та ключові користувачі	104
6.2.2 Ціннісні пропозиції для цільових сегментів.....	105
6.2.3 Канали збуту та взаємодія з клієнтами	106
6.3 Організаційне та ресурсне забезпечення стартап-проєкту	107

6.3.1	<i>Ключові види діяльності.....</i>	107
6.3.2	<i>Ключові ресурси.....</i>	108
6.3.3	<i>Ключові партнери</i>	110
6.4	Економічне обґрунтування та фінансова модель	110
6.4.1	<i>Структура витрат на розробку та впровадження MVP....</i> <i>.....</i>	110
6.4.2	<i>Формування роздрібної ціни ліцензії</i>	111
6.4.3	<i>Оцінка окупності та фінансових перспектив</i>	112
6.5	Аналіз ринку та конкурентного середовища.....	113
6.6	Перспективи розвитку та масштабування стартап-проєкту ...	115
ЗАГАЛЬНІ ВИСНОВКИ		118
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ		122

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI – Artificial Intelligence – штучний інтелект.

CT – Computed Tomography – комп'ютерна томографія.

CV – Cross-Validation – крос-валідація.

DL – Deep Learning – глибинне навчання.

EFB – Exclusive Feature Bundling – метод об'єднання взаємовиключних ознак, що використовується в LightGBM.

FN – False Negative – хибно-негативний результат.

FOS – First-Order Statistics – ознаки першого порядку.

FP – False Positive – хибно-позитивний результат.

GGO – Ground-Glass Opacity – «симптом матового скла» на зображеннях легень.

GLCM – Gray Level Co-occurrence Matrix – матриця співзустрічі рівнів сірого.

GLDS – Gray Level Difference Statistics – статистики різниці рівнів сірого.

GLNU – Gray-Level Non-Uniformity – неоднорідність рівнів сірого.

GLRLM – Gray Level Run Length Matrix – матриця довжин послідовностей рівнів сірого.

GOSS – Gradient-based One-Side Sampling – метод вибірки об'єктів за градієнтом у LightGBM.

HGRE – High Gray-Level Run Emphasis – підкреслення пробігів із високими рівнями сірого.

HU – Hounsfield Units – одиниці Хаунсфілда.

IDE – Integrated Development Environment – інтегроване середовище розробки.

KNN – k-Nearest Neighbors – метод k найближчих сусідів.

LBP – Local Binary Patterns – локальні бінарні патерни.

LDA – Lung Density Analysis – комерційний програмний модуль (Imbio).

LGRE – Low Gray-Level Run Emphasis – підкреслення пробігів із низькими рівнями сірого.

LRE – Long Run Emphasis – підкреслення довгих пробігів.

LTA – Lung Texture Analysis – комерційний програмний модуль (Imbio).

MAD – Median Absolute Deviation – медіанне абсолютне відхилення.

ML – Machine Learning – машинне навчання.

MVP – Minimum Viable Product – мінімально життєздатний продукт.

PACS – Picture Archiving and Communication System – система архівування та передавання медичних зображень.

RAM – Random Access Memory – оперативна пам'ять.

RIS – Radiology Information System – радіологічна інформаційна система.

RLNU – Run Length Non-Uniformity – неоднорідність довжин пробігів.

ROI – Region of Interest – область інтересу на зображенні.

RP – Run Percentage – відсоток пробігів.

SRE – Short Run Emphasis – підкреслення коротких пробігів.

SSD – Solid-State Drive – твердотільний накопичувач.

SVD – Singular Value Decomposition – сингулярний розклад матриці.

SVM – Support Vector Machine – метод опорних векторів.

TN – True Negative – істинно-негативний результат.

ТР – True Positive – істинно-позитивний результат.

VLM – Vision-Language Model – візуально-мовна модель.

ГБ – гігабайт.

КТ – Комп’ютерна томографія.

МІС – медична інформаційна система.

МРТ – Магнітно-резонансна томографія.

ПЗ – Програмний застосунок.

СЗЛ – Синдром зникаючої легені.

ВСТУП

Своєчасна та коректна диференціальна діагностика захворювань легень за даними комп'ютерної томографії органів грудної клітини є однією з ключових задач сучасної променевої діагностики [1]. Візуальні патерни ураження легеневої паренхіми, такі як ділянки «матового скла», консолідації, емфізематозні зміни чи булли, можуть спостерігатися при різних нозологіях і часто істотно перекриваються між собою. Це, у свою чергу, може суттєво ускладнювати відмежування, наприклад, вірусної (COVID-19) та бактеріальної пневмонії, віддалених наслідків перенесеної інфекції (long-COVID), а також рідкісних станів на кшталт синдрому зникаючої легені [2; 3]. За умов високого навантаження на лікарів-радіологів та обмеженого часу на опис досліджень така клінічна невизначеність підвищує ризик пізньої або надто суб'єктивної інтерпретації результатів КТ.

Додатковим викликом є міжсайтова варіабельність КТ-даних, що зумовлена відмінностями в моделях томографів, протоколах сканування, реконструкційних ядрах та параметрах дозування. Навіть за однакових клінічних сценаріїв зображення, отримані на різних сканерах, можуть мати різні контрастно-шумові характеристики та текстуру, що ускладнює безпосереднє використання накопичених даних для побудови узагальнених моделей машинного навчання. Без спеціальних методів гармонізації така варіабельність призводить до зміщення оцінок, зниження точності алгоритмів та навіть обмежує можливість їх перенесення між клініками [4].

На цьому тлі зростає інтерес до радіоміки – кількісного аналізу медичних зображень шляхом виокремлення текстурних ознак із визначених зон інтересу [5]. Такі ознаки, побудовані на основі матриці співзустрічей рівнів сірого, градієнтних та лінійних структур чи локальних бінарних шаблонів, дають змогу перетворити візуально складні патерни у структуровані числові вектори. Завдяки своїй інтерпретованості текстурні характеристики

можуть узгоджуватися з відомими радіологічними феноменами та доповнювати методи глибинного навчання, підвищуючи прозорість та обґрунтованість автоматизованої підтримки прийняття рішень [6].

Використання методів машинного навчання для класифікації легеневих захворювань на основі текстурних ознак відкриває можливість створення локальних, ресурсно ощадних та інтерпретованих систем допомоги лікарю. На відміну від комерційних рішень, орієнтованих переважно на великі мультицентрові дані та хмарну інфраструктуру, запропонований підхід забезпечує прозоре, інтерпретоване представлення результатів для лікаря, а також додатково дає змогу виконувати обробку КТ-знімків локально, без передачі персоніфікованої інформації за межі медичного закладу.

Таким чином, реалізація системи диференціальної діагностики захворювань легень є актуальною задачею, яка поєднує текстурний аналіз КТ-зображень, методи міжсайтової гармонізації, класичні алгоритми машинного навчання, а також реалізована у вигляді програмного прототипу, придатного для інтеграції у робочий процес лікаря-радіолога. Розв'язання цієї задачі сприятиме підвищенню об'єктивності та відтворюваності інтерпретації КТ-досліджень та оптимізації використання ресурсів закладів охорони здоров'я.

Мета і завдання роботи

Метою роботи є підвищення об'єктивності та відтворюваності процесу диференціальної діагностики захворювань легень за даними комп'ютерної томографії шляхом використання методів гармонізації міжсайтової варіабельності радіомічних ознак та побудови моделей машинного навчання у складі локальної програмної системи підтримки прийняття рішень.

Для досягнення даної мети було сформульовано наступні *завдання*:

1. Проаналізувати сучасні підходи та програмні засоби діагностики захворювань легень для виявлення їхніх обмежень.
2. Розробити методику попередньої обробки КТ-зображень та виокремлення інтерпретованих текстурних ознак.
3. Програмно реалізувати метод гармонізації для усунення

міжсайтової варіабельності даних.

4. Побудувати та валідувати моделі машинного навчання для диференціальної діагностики та виявлення синдрому зникаючої легені.

5. Розробити локальну програмну систему (MVP) для автоматизованого аналізу знімків та формування звітів.

Використані методи. Застосовано комплексний підхід, що поєднує методи цифрової обробки та сегментації КТ-зрізів (моделі U-net), а також методи текстурного аналізу (GLCM, GLDS, GLRLM, LBP) для формування інформативних радіомічних ознак. Особливістю дослідження є реалізація власної Python-адаптації методу гармонізації RELIEF для усунення міжсайтової варіабельності. Для задач класифікації використано алгоритми машинного навчання, валідація яких здійснюється з урахуванням специфіки незбалансованих вибірок.

Отримані результати. У результаті роботи реалізовано повний радіомічний конвеєр аналізу КТ-зображень легень, інтегрований у десктопний офлайн-застосунок «LungScan». У задачі мультикласифікації найкращі результати показала мультиноміальна логістична регресія з асигуру до 0.6 та F1-зваженою до 0.61, що є помірними, але відтворюваними показниками в умовах дисбалансу вибірки. У бінарній задачі виявлення синдрому зникаючої легені найвищі метрики досягнуто для SVM (асигуру та F1-зважена до 0.81), а гармонізація ознак загалом зберегла прийнятний рівень якості, особливо для стійких ансамблевих моделей. Це підтверджує, що запропонований радіомічний підхід є практично придатним для скринінгового виявлення СЗЛ та попередньої диференціації станів легень за умови подальшого розширення й балансування вибірки.

Публікації. За результатами виконаної роботи була опублікована 1 наукова стаття у фаховому журналі категорії Б:

1. Лутченко В. Г. Ефективність алгоритмів машинного навчання для класифікації змін структури легень у постковідних та гострих стадіях COVID-

19 / В. Г. Лутченко, В. О. Бабенко, Є. А. Настенко, М. І. Линник // Біомедична інженерія і технологія. – Київ, 2024. - №15/2024. – С. 27–35 (стаття). – DOI: <https://doi.org/10.20535/.2024.15.306763> (дата звернення 15.11.2025) [28].

Була опублікована 1 наукова стаття у фаховому виданні категорії А (індексована в Scopus):

1. Іє.А. Nastenko, Diagnosis of COVID-19-Associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: A Review of Methods and Future Research Directions / Іє.А. Nastenko, М.І. Linnik, М.О. Honcharuk, І.В. Davydovych, V.H. Lutchenko, V.O. Babenko // Innovative Biosystems & Bioengineering. – 2025. – Vol. 9, № 3. – Р. 16-27 (стаття). DOI: <https://doi.org/10.20535/ibb.2025.9.4.325335> (дата звернення 27.11.2025) [74].

Структура роботи

Магістерська дисертація за темою «Система диференціальної діагностики захворювань легень методами машинного навчання» виконана студенткою Лутченко Вікторією Геннадіївною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині», побудована за класичним типом та викладена на 135 сторінках машинописного тексту. Вона складається з: вступу; 6 розділів (*Аналіз літературних джерел та програмних аналогів систем диференціальної діагностики захворювань легень, Засоби реалізації системи диференціальної діагностики, Обґрунтування вибору методів та інструментів, Програмна реалізація та методика роботи програмного застосунку, Обговорення отриманих результатів, Стартап-проект з диференціальної діагностики легень “LungScan”*), висновків до кожного з цих розділів; загальних висновків та списку використаних джерел, який налічує 80 джерел (7 – на кирилиці, 73 – на латиниці). В роботі представлено 28 рисунків і 4 таблиці.

РОЗДІЛ 1

АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ ТА ПРОГРАМНИХ АНАЛОГІВ СИСТЕМ ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЛЕГЕНЬ

У даному розділі окреслено клінічний контекст поставленої задачі із зазначення типових труднощів у інтерпретації знімків. Крім того, подано огляд наявних аналогів із визначенням їх технічних аспектів, на основі аналізу яких – сформовано методичний контур роботи з акцентом на відтворюваність, інтерпретованість та врахування міжсайтової варіабельності.

1.1 Клінічний контекст задачі та постановка проблеми

На сьогодні комп'ютерна томографія є не тільки базовим, а і провідним інструментом для диференціальної діагностики, завдяки високій просторовій роздільній здатності та чутливості [7]. Зокрема, це стосується і уражень легень, адже у клінічній практиці часто доводиться відмежовувати типові стани з подібними або перекривними патернами – це можуть бути як вірусні пневмонії (зокрема COVID-19), так і постковідні зміни, бактеріальні пневмонії та емфізематозні ураження. Таке перекриття зумовлене спільними патобіологічними механізмами запалення та ремоделювання паренхіми (набряк, інтерстиціальна інфільтрація, фіброз тощо), які на КТ проявляються як матове скло, консолідації чи ретикулярність, відмінні переважно за вираженістю й розповсюдженням [8; 9].

Крім вищезгаданого, проблематика також зумовлена між- та внутрішньоспостерігачовою варіабельністю інтерпретації, потребою у відтворюваних кількісних показниках і нерівномірним навантаженням на діагностичні відділення. Саме тому, останні роки досить активно впроваджуються в роботу так звані «AI-as-assist» – це алгоритмічні

інструменти, що здатні надавати певні підказки, тим самим пришвидшуючи процес діагностики та зменшуючи рутинне навантаження, але зберігаючи за лікарем остаточне клінічне рішення [10].

1.2 Огляд програмних аналогів на ринку

1.2.1 LungQ та CAD4COVID-CT (Thirona)

Першими з таких аналогів є продукти від нідерландської компанії Thirona, що спеціалізується на AI-рішеннях для автоматичного аналізу КТ легень. Її програмна платформа LungQ (рис. 1.1) отримала перше схвалення FDA у 2018 році і використовує алгоритми глибокого навчання для високоточної сегментації легеневих структур та кількісної оцінки патологічних змін у легеневій паренхімі. Зокрема, система автоматично вимірює об'єм і частку легеневої тканини з ознаками емфіземи та інших змін, спираючись на HU-орієнтовані порогові правила з подальшою лобарною агрегацією [11].

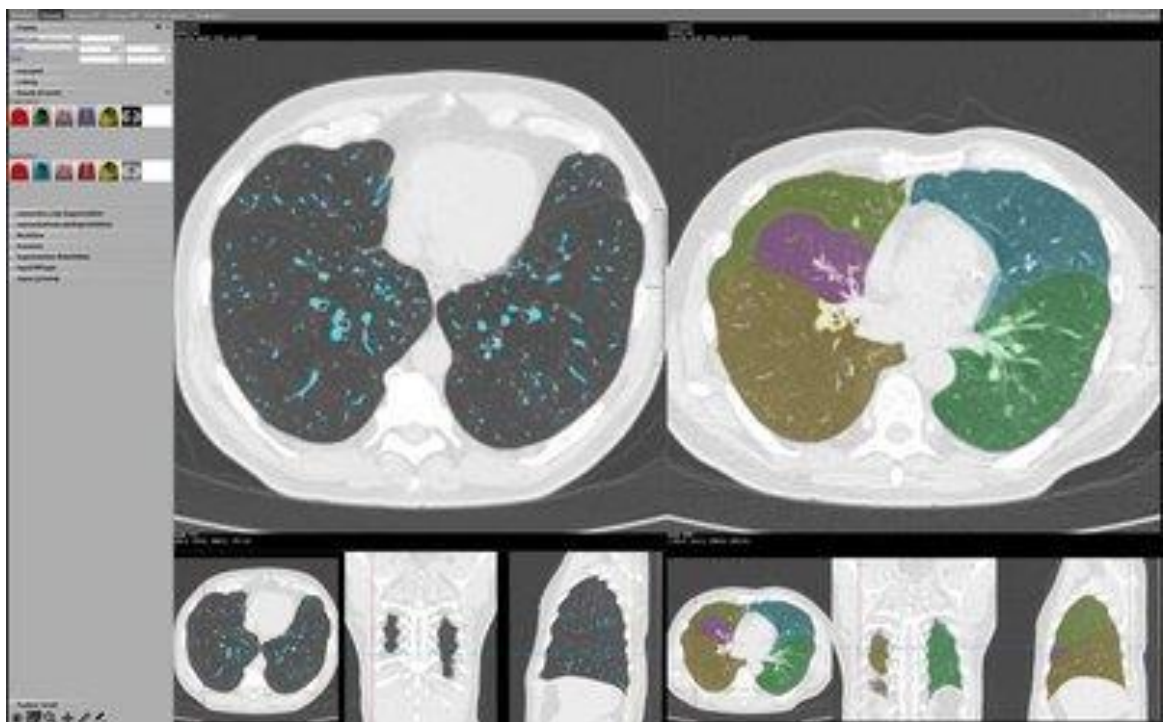


Рисунок 1.1 – Приклад інтерфейсу платформи LungQ [12]

Окрім LungQ, у відповідь на пандемію COVID-19 компанія досить оперативно розробила новий модуль CAD4COVID-CT для кількісної оцінки ураження легень саме при коронавірусній пневмонії. Цей AI-інструмент на базі DL автоматично виконує сегментацію та детекцію КТ грудної клітини пацієнта з COVID-19, після чого обчислює відсоток ураженої легеневої тканини. На основі цих даних формується звіт з деталізацією по частках легень і накладеним кольоровим «heatmap»-зображенням, що демонструє локалізацію патологічних змін. У Медичному центрі Лейденського університету (LUMC, Нідерланди) програму CAD4COVID-CT було інтегровано в робочий процес, де лікарі відзначали, що завдяки даному інструменту вдалось прискорити діагностику та підвищити її точність, доповнюючи «людську» здатність розпізнавати патерни з екстремально точними кількісними вимірюваннями асиметричних уражень [13].

1.2.2 AI-Rad Companion Chest CT (Siemens Healthineers)

Іншим схожим рішенням є AI-Rad Companion Chest CT – це комерційний продукт від компанії медичних технологій Siemens Healthineers, який був представлений у 2019 році як багатофункціональний AI-помічник для аналізу КТ органів грудної клітини. Подібно до попередніх, він працює на основі підходів глибокого навчання та HU-орієнтованих порогових правил. Даний програмний застосунок здійснює автоматизовану оцінку відразу декількох систем: дихальної, серцево-судинної та кістково-м'язової. Проте, в межах цієї роботи для аналізу нас цікавить лише модуль Pulmonary, що виконує сегментацію легень на частки і може виявляти та виділяти патологічні утворення в легенях [14].

Суттєво зросла актуальність впровадження AI-Rad Companion Chest CT вже під час пандемії COVID-19, оскільки цей інструмент має здатність автоматично визначати та кількісно оцінювати ділянки підвищеної щільності в легенях на неконтрастних КТ, які відповідають матовим затемненням та консолідації. Алгоритм підсвічує подібні області (рис. 1.2), після чого обчислює «opacity score» (індекс непрозорості легеневої паренхіми), що дає

змогу об'єктивно оцінити ступінь ураження легень пневмонією. Згідно з оцінками радіологів, такий «постійний експерт» на робочому місці допомагає стандартизувати опис патологій і економить час, не замінюючи лікаря, а лише доповнюючи його роботу [14].

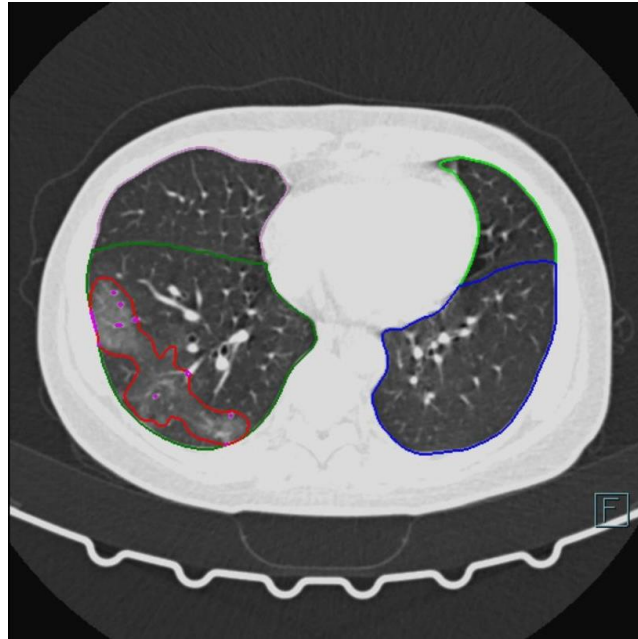


Рисунок 1.2 – Виділення патологічних областей на КТ зрізах у AI-Rad Companion Chest CT [14]

1.2.3 LDA+, LTA (Imbio)

Також вдалими прикладами є продукти американської компанії Imbio, що з 2012 року займається розробкою AI-рішень для кількісного аналізу знімків легень та серця. Її перший продукт під назвою Lung Density Analysis (LDA) був запущений на ринок ще у 2015 році, як інструмент автоматичної оцінки об'ємів легеневої тканини зниженої щільності [15]. Здебільшого його використовували у програмах скринінгу раку легень і моніторингу курців, допомагаючи об'єктивно відстежувати накопичення емфіземи з часом [16].

Але, у період пандемії, на основі базового LDA Imbio розробила розширену версію – LDA+, яка є більш актуальною саме для оцінки дифузних уражень легень, таких як пневмонія чи вірусні інфекції. Цей новий продукт, що наділений здатністю автоматично кількісно оцінювати вираженість і

розповсюдженість запальних змін (на рис. 1.3) допомагає лікарю отримати об'єктивний відсоток ураження легень COVID-пневмонією та оцінити, чи є у пацієнта супутня патологія легень, що може впливати на прогноз [17].

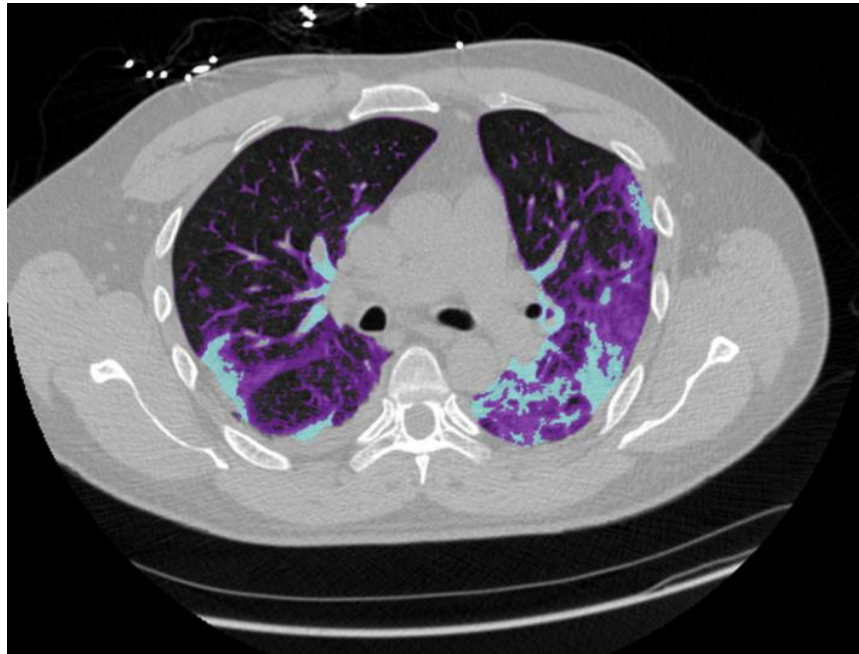


Рисунок 1.3 – Приклад роботи LDA+ з позначенням класів щільності легеневої тканини [17]

Окремо варто виділити Lung Texture Analysis (LTA) – алгоритм для аналізу текстур легеневої паренхіми, розроблений на основі технології CALIPER (Mayo Clinic), що автоматично класифікує легеневу тканину на декілька типових патернів на основі локальних DCT-ознак та класичних ML. Зазвичай виділяються щонайменше 5 класів – нормальна тканина, ground-glass помутніння, ретикулярні зміни, стільниковий легеневий рисунок (honeycombing) та гіперлюцентні зони, і саме тому, LTA особливо доречний для постінфекційних і хронічних станів, як, наприклад, long-COVID [18].

Як результат, система надає детальну кольорову карту легень, де різними кольорами позначені області з тим чи іншим типом змін, а також табличний звіт із відсотковим співвідношенням виявлених патернів в кожній частці та сегменті легень [18].

1.2.4 Aidoc

Не менш цікавим аналогом є Aidoc – один зі світових лідерів у галузі медичного AI, відомий насамперед своїми алгоритмами для автоматичного виявлення гострих патологій на КТ та МРТ і системою тріажу невідкладних випадків. Під час пандемії компанія швидко адаптувала свої напрацювання для боротьби з COVID-19, а саме: реалізувала алгоритм, що здатний аналізувати будь-які КТ грудної клітини (навіть ті, що виконані з іншою метою) на предмет характерних ознак COVID-пневмонії. Технічно, цей інструмент діє як «другий читач»: DL-детектор працює як «always-on» сервіс у зв'язці з PACS/RIS, і автоматично позначає випадки з підозрілими змінами в легенях – передусім, наявність множинних матових затемнень (GGO) та консолидацій [19].

Хоч формально він не пройшов повноцінного клінічного схвалення, уповноважений орган (FDA) дозволив його запуск і впровадження через важливість задачі. Вона полягає в тому, що якщо пацієнту роблять КТ (наприклад, з приводу болю в животі або травми), а на зображеннях присутня типова картина двобічної вірусної пневмонії, AI відразу сповістить рентгенолога про таку знахідку. Це, в свою чергу, допоможе своєчасно запідозрити прихований перебіг COVID-19, ізолювати пацієнта та розпочати необхідне лікування. Зокрема, за повідомленнями окремих центрів, до 8–10% пацієнтів, яким КТ виконували з іншого приводу, мали на знімках зміни, сумісні з безсимптомним COVID-19 [20].

1.2.5 contextflow ADVANCE Chest CT (INSIGHTS, SEARCH)

Стартапом серед аналогів є contextflow ADVANCE Chest CT – інноваційне рішення від розробників з Відня, яке надає всеосяжну AI-підтримку при читанні КТ грудної клітини. Це продукт, що поєднує DL-детекцію ознак із контент-орієнтованим пошуком подібних випадків (content-based retrieval) [21]. Він налічує досить багато модулів, проте найбільше нас цікавлять INSIGHTS та SEARCH, оскільки саме вони зосереджені на аналізі легеневих тканин і патернів їх змін.

1.2.6 dragonfly

Dragonfly – це універсальне програмне середовище для візуалізації, сегментації та кількісного аналізу багатовимірних зображень (2D/3D), яке підтримує медичні дані у форматі DICOM і дає змогу працювати як із цілими серіями КТ, так і з локальними регіонами інтересу [24]. Зокрема, платформа поєднує ручні та напівавтоматичні інструменти розмітки, має засоби постобробки масок і побудови вимірювань у вибраних ROI: від базових гістограм HU-розподілу з налаштуванням кількості бінів до розрахунку текстурних характеристик на кшталт GLCM/GLRLM. Наявні модулі дозволяють отримувати таблиці показників для кількох зон, порівнювати їх між собою та вивантажувати необхідні результати [25].

У контексті легеневих КТ Dragonfly зазвичай застосовують як інструмент підготовки та кількісної оцінки: лікар вручну виділяє зони патології, що вже є досить трудомістким процесом. Потім він аналізує їхній інтенсивнісний профіль (гістограми, описові статистики тощо) і може обчислювати текстурні матриці для оцінки однорідності чи гетерогенності уражень. Надалі всі отримані кількісні метрики потребують подальшої інтерпретації спеціалістом або можуть бути передані у зовнішні моделі [25].

1.3 Методологічний контур на основі аналізу аналогів

Виконаний аналіз комерційних аналогів показав, що серед них абсолютною більшістю переважають підходи на основі глибинного навчання, у вигляді певних модулів для сегментації легеневої тканини та визначення патологічних змін в ній, з подальшою HU-орієнтованою кількісною агрегацією метрик. Щодо нюансів гармонізації міжсайтових відмінностей, то вони здебільшого описуються у вигляді загальних підходів, водночас конкретні статистичні методи зазвичай не декларуються, принаймні публічно.

Таким чином, враховуючи все вищеописане та продовжуючи напрацювання бакалаврської роботи, було прийнято рішення використовувати

тандем текстурних ознак (GLCM, GLDS, GLRLM, LBP) із класичними моделями машинного навчання. У межах магістерського дослідження, цей підхід було розгорнуто до методологічного конвеєра, здатного вирішувати дві окремі задачі: диференціальну діагностику таких нозологій, як COVID-19, здорові легені, бактеріальна пневмонія, long-COVID та бінарну класифікацію для виявлення ознак синдрому зникаючої легені у пацієнтів з гострим COVID-19. Крім того, окрему увагу приділено міжсайтовій гармонізації текстурних ознак за допомогою методу RELIEF (Removal of Latent Inter-scanner Effects through Factorization), раніше відомого як UNIFAC harmonization.

Такий вибір мотивовано вимогами інтерпретованості й відтворюваності у клінічному середовищі, адже для лікаря критично важливо розуміти, чому система схилиється до того чи іншого рішення, які саме патерни зображення це зумовили та чи узгоджується це з клінічною картиною. Додатковою практичною перевагою такого підходу є краща стійкість до обмежених вибірок і можливість контролювати кожний крок підготовки даних і навчання моделей, що важливо для уникнення методологічних похибок і «витоку даних».

Проте є і деякі слабкі сторони обраної методології, на які варто звернути увагу під час реалізації. Сюди можна віднести чутливість до якості сегментації та налаштувань, обмежену здатність захоплювати дуже складні просторові патерни порівняно з DL, потреба в ретельній гармонізації між сайтами та ризик втрати інформації через ручне конструювання ознак.

Висновок з розділу 1

У даному розділі було виконано огляд літератури та окреслено клінічну постановку задачі, де зазначається, що інтерпретація КТ уражень легень переважно ускладнюється перекриттям патернів для різних нозологій і нерівномірним навантаженням на діагностичні центри. Усе це підсилює

актуальність підходу «AI-as-assist» саме як інструменту підтримки, а не заміни спеціаліста.

Огляд наявних аналогів свідчить про переважне застосування DL-модулів із сегментацією та HU-орієнтованими кількісними індикаторами. Проте питання явної статистичної гармонізації міжсайтових відмінностей у цих рішеннях здебільшого або взагалі не висвітлюється, або не містить детальної методики.

Таким чином, виходячи з усієї отриманої інформації, було визначено методологічний конвеєр з використанням тандему текстурних ознак та класичних методів машинного навчання, здатний вирішувати дві окремі задачі: диференціальну діагностику та бінарну класифікацію. Крім того, окрему увагу приділено міжсайтовій гармонізації текстурних ознак.

РОЗДІЛ 2

ЗАСОБИ РЕАЛІЗАЦІЇ СИСТЕМИ ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ

У даному розділі викладено методологічні засади та інструментальні засоби, використані при розробці системи для диференціальної діагностики захворювань легень. Зокрема, тут описується структура та особливості вихідного датасету, теоретичні основи текстурних ознак, що використовуються для радіомічного опису легеневої паренхіми, а також підхід до їх гармонізації. Окрему увагу приділено вибору та математичним засадам моделей машинного навчання, метрикам оцінювання їх якості й обчислювальному середовищу, в якому реалізовано запропонований MVP.

2.1 Дані та структура датасету

Магістерське дослідження базувалося на спеціалізованому датасеті зрізів КТ легень, сформованому в рамках співпраці НТУУ «КПІ ім. Ігоря Сікорського» та Національного інституту фтизіатрії і пульмонології ім. Ф. Г. Яновського НАМН України [75; 78; 79]. До бази включено пацієнтів із підтвердженими нозологіями, що забезпечує репрезентативну варіативність як патологічних, так і непатологічних станів.

Досліджувана вибірка складається з 97 пацієнтів. Вона включає 48 випадків COVID-19 (із них 27 із СЗЛ), 13 здорових, 5 із бактеріальною пневмонією та 31 з long-COVID. Сукупний обсяг становить 26 623 зрізи, розподілені за класами: COVID-19 – 14 478 (з них 9 346 із ознаками СЗЛ), здорові – 3 092, бактеріальна пневмонія – 1 619, long-COVID – 7 434. Усі ці розподіли відображені на рисунках 2.1 та 2.2 у вигляді стовпчикових гістограм.

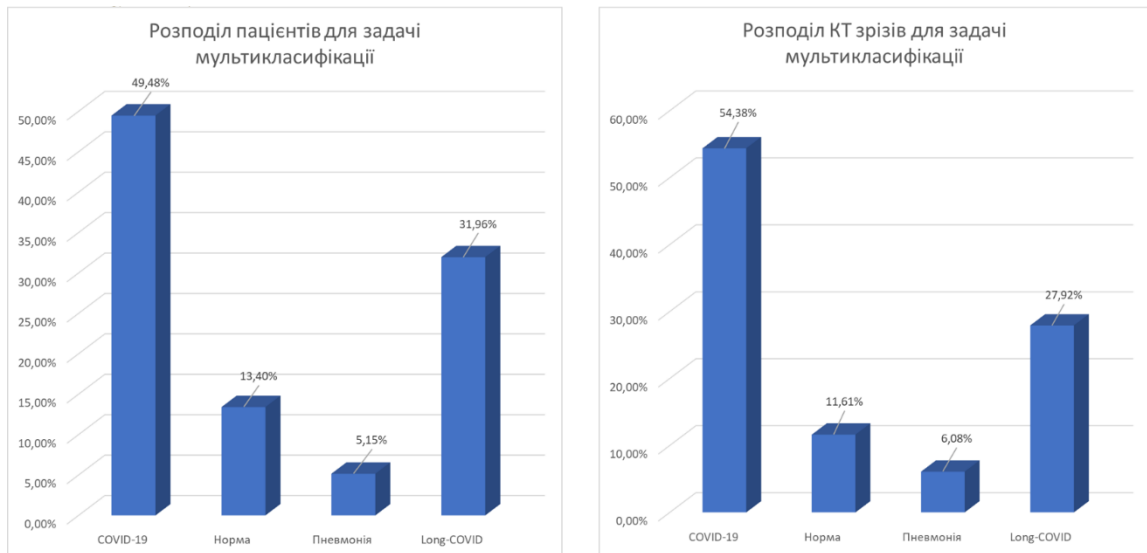


Рисунок 2.1 – Гістограмне відображення розподілів пацієнтів (ліворуч) та зрізів (праворуч) для задачі мультикласифікації

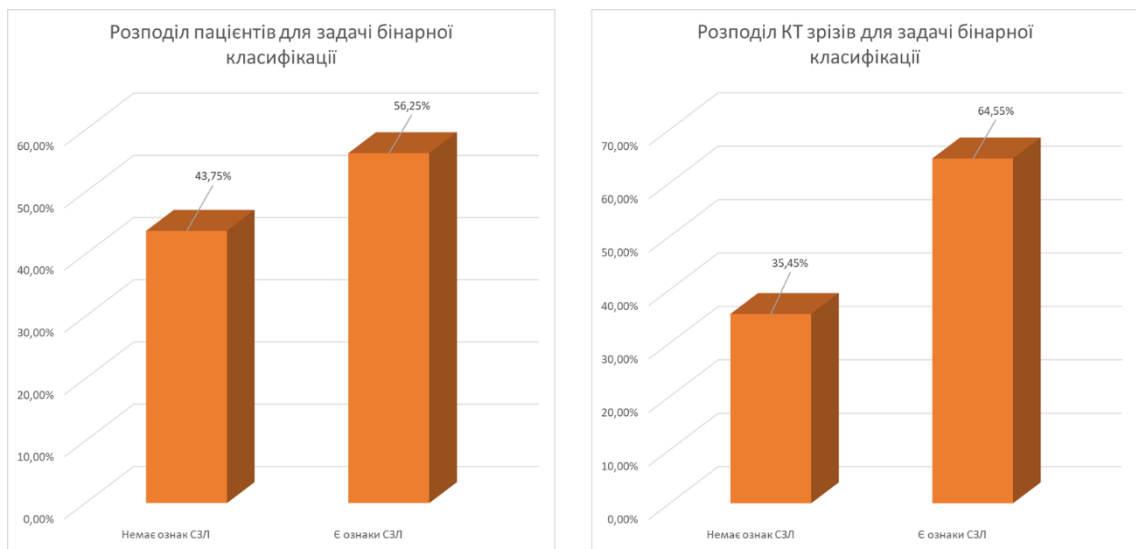


Рисунок 2.2 – Гістограмне відображення розподілів пацієнтів (ліворуч) та зрізів (праворуч) для задачі бінарної класифікації

Таким чином, попри доволі велику кількість зрізів, розподіл класів є нерівномірним із суттєвим переважанням COVID-19. Особливо це стосується класу «бактеріальна пневмонія», що налічує лише 5 пацієнтів, оскільки це суттєве обмеження для даного дослідження, усі результати, що отримані для даного класу варто сприймати радше як попередні, а не як статистично достовірні. Крім того, присутня внутрішньокласова неоднорідність, оскільки

ми маємо велику підмножину СЗЛ у межах COVID-19, що також ускладнює коректне навчання моделей.

Маючи такий вихідний набір даних, формулювання задач класифікації лишається незмінним – це мультикласова або ж диференціальна з визначенням 4 класів (COVID-19, здорова легеня, бактеріальна пневмонія, long-COVID) та бінарна, яка орієнтована на виявлення СЗЛ серед пацієнтів із COVID-19.

Первинні зображення були надані у форматі DICOM, після чого для кожного з них було створено відповідну бінарну маску області інтересу у форматі PNG. Назви сегментованих масок відповідають найменуванню оригінального зрізу, для того, аби спростити візуальний контроль якості та автоматизацію подальших кроків конвеєра. Щодо формату, то використання PNG для сегментованих масок є цілком доцільним і виправданим, оскільки він забезпечує стиснення без втрати якості, точно зберігає піксельні значення (що критично для бінарних і багатокласових міток), має широку підтримку інструментами, і крім цього всього дає менший розмір файлів порівняно з форматами без стиснення. У підсумку контури сегментації не деградують під час збереження, а інтеграція масок у конвеєр обробки залишається простою та відтворюваною [31].

2.2 Теоретичні основи текстурних ознак

В цілому, текстура описується як певна сукупність статистичних характеристик просторового розподілу інтенсивностей, яка дозволяє переходити від візуального сприйняття «зернистості» або ж «гетерогенності», безпосередньо до кількісних показників. Серед класичних підходів для виокремлення подібних ознак виділяють, матрицю суміжності рівнів сірого (Gray Level Co-occurrence Matrix), статистику різниць рівнів сірого (Gray Level Difference Statistics), матрицю довжин пробігів рівнів сірого (Gray Level Run Length Matrix) та локальні бінарні шаблони (Local Binary Pattern) [26].

У наступних підпунктах ми розглянемо кожен із вищезгаданих методів більш детально.

2.2.1 Gray Level Co-occurrence Matrix

І почнемо з матриці суміжності рівнів сірого, що має абревіатуру GLCM. Це класичний інструмент другого порядку, який описує структуру текстури через статистику пар пікселів (рис. 2.3) та доволі широко використовується для аналізу медичних зображень [27; 28 с. 3-4; 74].

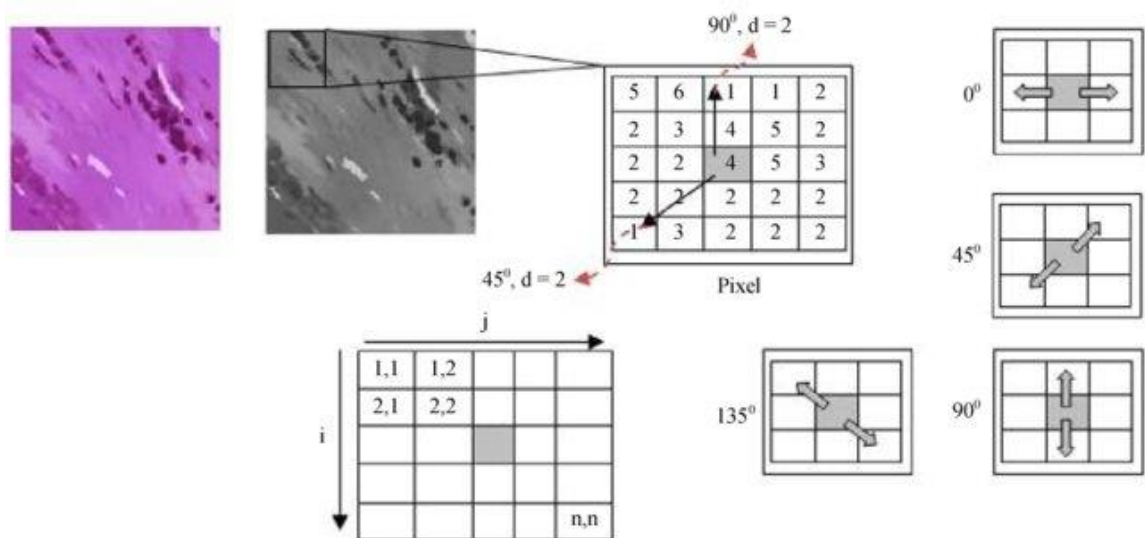


Рисунок 2.3 – Схематична ілюстрація побудови GLCM [29]

Як видно зі схематичного зображення на рис. 2.3, значення записуються у квадратну матрицю, де кількість рядків і стовпців визначається градацією сірого в зображенні. Надалі відбувається розрахунок, з якою частотою певні пари пікселів зустрічаються один з одним – це і дає розуміння текстури зображення, що аналізується: його контраст, кореляцію, енергію, однорідність тощо [28, с. 3-4; 30; 31, с. 34; 74].

Вищеописані ознаки є основними і кожна з них розраховується за допомогою чітко визначених формул, де $P(i, j)$ – значення матриці спільної зустрічальності для інтенсивностей i та j , μ_i та μ_j – середні значення, σ_i та σ_j – стандартні відхилення відповідних інтенсивностей [31, с. 35].

Для контрасту формула має наступний вигляд:

$$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2 P(i,j) \quad (2.1)$$

Якщо дана метрика рівна нулю, то можна зробити логічний висновок, що сусідні пари пікселів не мають різниці у яскравості, а отже зображення є однорідним. Таким чином, чим більшим буде значення – тим більшими є розбіжності між значеннями інтенсивності пікселів, а отже і тим менш однорідним є така область зображення [32].

Кореляція відображає лінійну залежність між значеннями сірих рівнів у парі сусідніх пікселів, і її розрахунок виконується за формулою (2.2):

$$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i \sigma_j} \quad (2.2)$$

Значення даної ознаки може бути в діапазоні від -1 до 1, де значення, що наближені до меж даного інтервалу свідчать про сильну пряму чи обернену залежність інтенсивностей сусідніх пікселів, а 0 – відсутність лінійного зв'язку [32].

Енергія розраховується за формулою (2.3):

$$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i,j)^2 \quad (2.3)$$

Енергія, що також відома як другий кутовий момент, ж мірою текстурної рівномірності. Вона демонструє високі значення, у тому разі, якщо розподіл співзустрічальностей зосереджений в кількох домінантних комірках GLCM – це означає, що багато пар пікселів мають однакові або подібні рівні сірого, а отже зображення є здебільшого або повністю однорідним. У протилежному випадку, коли значення низькі – зріз має складну неоднорідну структуру [32].

Останньою з базових ознак є однорідність, яка обчислюється за формулою (2.4):

$$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{P(i,j)}{1+|i-j|} \quad (2.4)$$

Результат розрахунку однорідності має бути в межах від 0 до 1. Тут високі значення (близькі до 1) вказують на те, що розподіл концентрується вздовж головної діагоналі матриці, тож сусідні пари пікселів мають приблизно однаковий рівень сірого, а чим нижчі значення – тим більше зростають локальні градієнти яскравості і переважають контрастні відмінності [32].

2.2.2 Gray Level Difference Statistics

Як основа для статистики різниць рівнів сірого використовуються статистичні властивості різниць між значеннями сірого у сусідніх пікселях [28; 31; 33; 74]. До основних ознак GLDS відносять середнє значення різниць, стандартне відхилення різниць, ентропію, енергію та контраст [28, с. 4; 74]. Розглянемо кожну з них детальніше разом з їх формулами.

Середнє значення різниць (mean) характеризує середню абсолютну різницю інтенсивності між значеннями сусідніх пікселів та обчислюється за формулою (2.5) [34].

$$\text{Mean} = \frac{1}{N} \sum_{i=1}^N |I(x_i) - I(x_{i+1})| \quad (2.5)$$

де $I(x_i)$ та $I(x_{i+1})$ – значення інтенсивності пікселя i та його сусіда; N – загальна кількість пар пікселів [31, с.38].

Якщо значення mean низьке, то це може свідчити про дуже подібну інтенсивність сусідніх пікселів, і навпаки: високе значення – помітна відмінність яскравості у парі [34].

Наступним є стандартне відхилення різниць, що розраховується за допомогою використання формули (2.6):

$$\text{Std Dev} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (|I(x_i) - I(x_{i+1})| - \text{Mean})^2} \quad (2.6)$$

За допомогою даної метрики ми можемо аналізувати ступінь варіативності локальних контрастів у зображенні. Високі значення свідчать про те, що розподіл різниць є досить широким, тож є як дуже малі, так і великі коливання значень яскравості між суміжними пікселями. Якщо ж значення невелике, то логічно, що різниці інтенсивностей є більш скупченими близько до середнього значення і перепади менші [34].

Ентропія розподілу різниць – це міра про те, наскільки невпорядкованими або ж непередбачуваними є значення між сусідніми пікселями [34]. Для її розрахунку існує відповідна формула (2.7):

$$\text{Entropy} = - \sum_{k=1}^K p_k \log(p_k) \quad (2.7)$$

де p_k – ймовірність різниці інтенсивності k серед усіх пар пікселів [31, с. 38].

Розподіл різниць є рівномірним у тому випадку, якщо значення ентропії високе – це свідчить про те, що жодна конкретна величина різниці не домінує, і всі вони зустрічаються з приблизно однаковою частотою. У випадку, коли ентропія низька – розподіл, відповідно, є неоднорідним, і певні різниці з’являються в ньому значно частіше, ніж решта [34].

Подібно до GLCM, тут також є енергія, що обчислюється за дещо подібною формулою (2.8).

$$\text{Energy} = \sum_{k=1}^K p_k^2 \quad (2.8)$$

Якщо деякі інтенсивності з’являються частіше за інші, то такий розподіл є сконцентрованим і показники energy високі. Таким чином, логічно припустити, що текстури, де переважають деякі фіксовані різниці яскравості між сусідніми пікселями, матимуть високу енергію, тоді як текстури без явного домінування (рівномірні) – низьку [34].

Останньою з ознак в GLDS є також вже знайомий нам з попереднього методу контраст, що знаходиться за формулою (2.9):

$$\text{Contrast} = \sum_{k=1}^K k^2 p_k \quad (2.9)$$

Аналогічно до пояснення для GLCM: високий контраст – різні рівні сірого та висока гетерогенність інтенсивностей у локальних областях, низький контраст - значення яскравості сусідніх пікселів близькі і різких локальних перепадів немає [34].

2.2.3 Gray Level Run Length Matrix

Матриця довжин пробігу рівнів сірого або ж GLRLM використовується в текстурному аналізі для опису протяжності однорідних фрагментів зображення вздовж певного фіксованого напрямку, де кожен елемент $P(i, r | \theta)$ відповідає числу серій довжини r з інтенсивністю i [28; 31; 35; 74]. Схематично її роботу відображено на рис. 2.4:

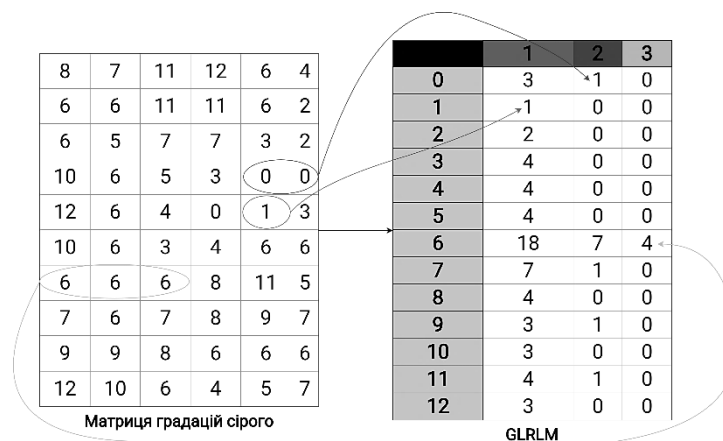


Рисунок 2.4 – Схематична ілюстрація побудови GLRLM-матриці [36]

До основних ознак довжин пробігу рівнів сірого відносять короткий пробіг (Short Run Emphasis), довгий пробіг (Long Run Emphasis), нерівномірність рівнів сірого (Gray Level Non-Uniformity), нерівномірність довжин пробігу (Run Length Non-Uniformity), відсоток пробігів (Run Percentage), підкреслення пробігів з низьким рівнем сірого (Low Gray Level

Run Emphasis) та підкреслення пробігів з високим рівнем сірого (High Gray Level Run Emphasis) [31; 37].

Аналогічно, як у всіх попередніх методах, розглянемо кожну з даних ознак окремо і більш детально. Також, у всіх наступних формулах $P(i, j)$ – значення в GLRLM-матриці для рівнів сірого i та довжини пробігу j ; N_g – кількість рівнів сірого; N_r – максимальна довжина пробігу; N – загальна кількість пробігів у матриці [31, с. 38].

І почнемо з Short Run Emphasis – він є показником розподілу коротких однорідних пробігів у зображенні [38]. Тобто, високі значення даної метрики вказують на те, що у текстурі переважають короткі однорідні сегменти, якими є послідовності пікселів однакового рівня сірого, а відповідно, структура зображення є більш дрібнозернистою і детальною [37]. Розрахувати короткі пробіги можна за формулою (2.10).

$$\frac{1}{N} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i, j)}{j^2} \quad (2.10)$$

Long Run Emphasis характеризує розподіл довгих однорідних пробігів у зображенні, і з цього логічно випливає, що високий показник довгого пробігу вказує на текстуру з переважно великими однорідними ділянками. Такі зображення за своєю структурою є більш грубозернистими [37].

Формула (2.11) слугує для обчислення даної ознаки:

$$\frac{1}{N} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j^2 P(i, j) \quad (2.11)$$

Нерівномірність рівнів сірого (GLNU), як можна зрозуміти з назви, характеризує неоднорідність розподілу пробігів за інтенсивностями [38]. Для її обчислення використовується формула (2.12):

$$\frac{1}{N} \sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_r} P(i, j) \right)^2 \quad (2.12)$$

Якщо значення GLNU низьке, то це є ознакою рівномірного розподілення пробігів між різними значеннями інтенсивності, тобто жоден з рівнів сірого не є домінуючим і текстура більш однорідна. У разі, коли нерівномірність рівнів сірого висока – більшість пробігів зосереджена лише на кількох рівнях сірого і в текстурні є певні домінуючі інтенсивності [37].

Коли ж потрібно розглянути неоднорідність розподілу пробігів за їх довжиною, то використовується ознака нерівномірності довжин пробігу (RLNU) [38], що обчислюється за формулою (2.13).

$$\frac{1}{N} \sum_{j=1}^{N_r} \left(\sum_{i=1}^{N_g} P(i, j) \right)^2 \quad (2.13)$$

Якщо розподіл однорідних сегментів рівної довжини рівномірний (присутні як короткі, так і довгі пробіги у зрівноважених пропорціях). Тоді значення даної ознаки будуть низькими. І навпаки, високе значення Run Length Non-Uniformity означатиме, що в зображенні домінують пробіги певної довжини, а отже текстура має характерний масштаб однорідних ділянок і є менш різноманітною за розміром структур [37].

Наступною ознакою є відсоток пробігів (RP) – це відносний показник кількості пробігів, що характеризує загальну «грубість» або «деталізованість» текстури [37]. Для його розрахунку користуються формулою (2.14):

$$\frac{N}{N_p} \quad (2.14)$$

Високі значення RP відповідають дрібнозернистим текстурам, оскільки вони мають багато коротких пробігів на визначеній площі, тоді як низькі

значення – показник однорідної структури, де на тій самій площі пробігів менше, оскільки вони довгі [37].

Останніми двома ознаками є підкреслення пробігів з низьким рівнем сірого (LGRE, формула (2.15)) та підкреслення пробігів з високим рівнем сірого (HGRE, формула (2.16)):

$$\frac{1}{N} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j)}{i^2} \quad (2.15)$$

$$\frac{1}{N} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} i^2 P(i,j) \quad (2.16)$$

LGRE використовують для підкреслення наявних темних однорідних фрагментів у текстурі, а HGRE, відповідно, для виокремлення світлих однорідних фрагментів [38]. З цього логічно зробити висновок, що чим вищим буде значення пробігів з низьким рівнем сірого – тим більше темних однорідних областей на зображенні, і чим більше підкреслення пробігів з високим рівнем сірого – тим більше яскравих однорідних ділянок є на зображенні [37].

2.2.4 Local Binary Pattern

Щоб кількісно описати мікротекстуру, часто використовують локальні бінарні шаблони (LBP). Уперше цей метод було представлено у 1994 році і його сутність полягає в тому, що він перетворює мікрооколиці пікселя на двійкові коди, де для кожного центрального пікселя зі значенням g_c порівнюють інтенсивності P сусідів на радіусі R . Якщо інтенсивність сусіда не менша за g_c , записують одиницю, в іншому випадку – нуль [31; 39].

Таким чином формується певний двійковий візерунок, який інтерпретують вже як ціле число, а вся область інтересу подається у вигляді гістограми частот таких кодів [39]. Увесь вищеописаний процес відображено нижче на рис. 2.5.

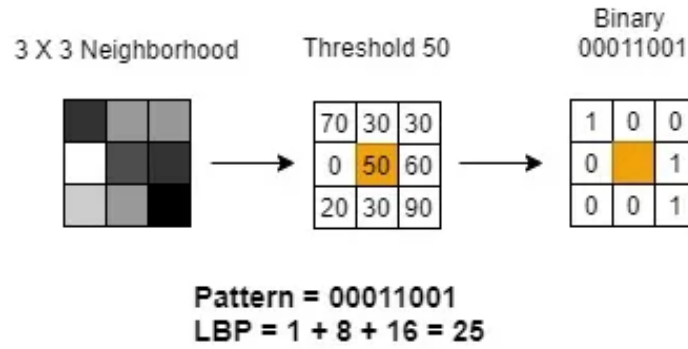


Рисунок 2.5 – Схематична ілюстрація формування LBP [40]

Для розрахунку локальних бінарних шаблонів використовується формула (2.17):

$$\text{LBP}(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c) \cdot 2^p \quad (2.17)$$

де $s(x) = 1$, якщо $x \geq 0$, і $s(x) = 0$, якщо $x < 0$; i_c – інтенсивність центрального пікселя; i_p – інтенсивність сусіднього пікселя; P – кількість сусідніх пікселів; (x_c, y_c) – координати центрального пікселя [31, с. 40].

Оскільки порівнюються відносні, а не абсолютні значення, базовий оператор інваріантний до монотонних змін яскравості й чутливо фіксує мікротекстури без залежності від рівнів сірого [39].

2.3 RELIEF як метод гармонізації

RELIEF (Removal of Latent Inter-scanner Effects through Factorization) – це новий мультіваріативний підхід у гармонізації даних, розроблений для усунення міжсайтової варіабельності та приведення радіомічних ознак до порівнянного виду у мультицентрових дослідженнях. Вперше його було запропоновано групою дослідників із Торонто у 2022-2023 роках під назвою UNIFAC harmonization. Проте в останніх статтях він вже згадується саме як акронім RELIEF [41].

На відміну від попередніх методів гармонізації (ComBat, CovBat тощо), більшість з яких застосовували одноваріативний підхід окремо до кожної ознаки, RELIEF є мультिवаріативним методом, який дозволяє оцінити та видалити як явні відмінності між сканерами, так і приховані латентні ефекти, що більш притаманні даним окремих центрів [42; 43].

Ключовою ідеєю даного методу є те, що він виконує одночасне зменшення розмірності та факторизацію даних з різних сканерів – так звану факторизацію «взаємопов'язаних матриць». Інакше кажучи, багатовимірні дані, що були отримані з декількох різних сканерів, моделюються через набір прихованих параметрів, які в свою чергу поділяються на спільні (common/shared) та сканер-специфічні фактори [42].

Спільні латентні фактори вважаються носіями біологічно значущого сигналу, оскільки вони описують варіативність, спільну для всіх наборів даних (усіх сканерів), і таким чином відображають закономірності, присутні незалежно від місця збору даних. Натомість сканер-специфічні фактори інтерпретуються як технічні відмінності або артефакти конкретного сайту, бо відповідають за ту варіабельність ознак, яка властива лише окремим сканерам [42].

У моделі RELIEF вихідні багатовимірні дані кожного сканера декомпонуються на кілька складових (рис. 2.6) [42]:

- i. ефект коваріат – вплив таких пояснювальних змінних, як, наприклад, клінічні або демографічні фактори (якщо вони враховані в модель) [42];
- ii. сканер-специфічні зсуви середнього – відмінності в середніх значеннях ознак між сканерами, що зводяться ефектами локації для кожного сайту [42];
- iii. латентні спільні варіації – приховані фактори, що є спільними для всіх сканерів, тож відображають варіації, що присутні у даних незалежно від сканера [42];

- iv. латентні сканер-специфічні варіації – приховані фактори, що проявляються лише в даних одного сканера, тобто є унікальними для сайту варіації [42];
- v. шум – це випадкова залишкова варіація, яка може відрізнитися між сканерами [42].

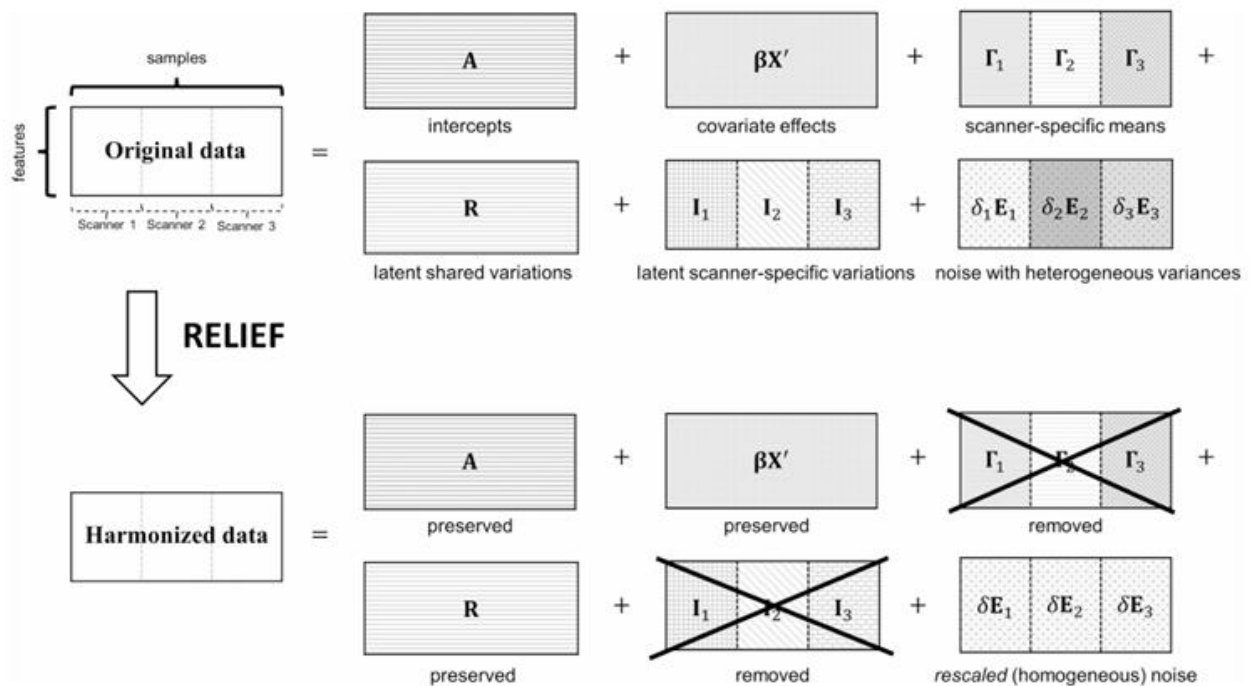


Рисунок 2.6 – Схематичне відображення роботи методу RELIEF [42]

Для досягнення бажаного ефекту в гармонізації, RELIEF видаляє сканер-специфічні зсуви середнього (ii) та латентні сканер-специфічні варіації (iv). Крім того, вирівнюється гетерогенний шум шляхом гомогенізації дисперсій між наявними наборами даних. У той самий час, фактори ефекту коваріат (i) та латентні спільні варіації (iii) лишаються незмінними, тим самим зберігаючи після гармонізації корисну біологічну інформацію про об'єкт дослідження [42].

Із проведених симуляцій та експериментів, що більш детально описані в оригінальній статті «RELIEF: A structured multivariate approach for removal of latent inter-scanner effects» [42] дійсно спостерігається як RELIEF досить ефективно знижує міжсканерні зміщення, не пригнічуючи при цьому значущі

асоціації ознак з біологічними факторами, тобто підтримує статистичну силу виявлення істинних ефектів. Таким чином, дані стають більш однорідними і їх вже можна застосовувати для подальшого аналізу, не ризикуючи отримати похибки, що викликані використанням різних сканерів («ефект сканера») [42].

Варто також зазначити, що хоча першочергово RELIEF було розроблено для усунення міжсайтової варіабельності на МРТ-даних, його також можна застосовувати і для КТ зрізів, оскільки різні комп'ютерні томографи, протоколи сканування та параметри реконструкції зображень також можуть вносити значну варіабельність у кількісні радіомічні показники між медичними центрами [42].

Таким чином, RELIEF – новий метод гармонізації, що здатен врахувати складні багатовимірні зсуви в даних, які не охоплюються простими покомпонентними підходами і спрямований на максимально повне усунення міжсканерних ефектів без втрати справжніх сигналів. Зокрема порівняно з популярним методом ComBat, який коригує лише середні та дисперсії окремих ознак, RELIEF усуває також латентні (scanner-specific) патерни варіацій, властиві конкретним сканерам, що в підсумку підвищує якість даних після гармонізації [42; 43].

2.4 Моделі машинного навчання

У наступних підпунктах наведено теоретичні відомості про моделі машинного навчання – як лінійні, так і нелінійні алгоритми різної природи.

2.4.1 Мультиноміальна логістична регресія

Мультиноміальна логістична регресія – це узагальнення бінарної логістичної регресії, яке застосовується у випадку, коли кількість класів перевищує два, і коли цільова змінна є категоріальною без природного порядку [44]. Таким чином, модель задає для кожного з існуючих класів лінійний предиктор (2.18), після чого обчислюється ймовірність класу через

softmax-функцію (2.19) – це дозволяє гарантувати невід’ємність ймовірностей [45].

$$\eta_{ik} = \beta_{0k} + \beta_k^T x_i \quad (2.18)$$

де β_{0k} – вільний член, β_k – вектор коефіцієнтів для класу k , а x_i – вектор ознак i -го спостереження [45].

$$P(Y_i = k|x_i) = \frac{\exp(\beta_{0k} + \beta_k^T x_i)}{\sum_{j=1}^K \exp(\beta_{0j} + \beta_j^T x_i)}, \quad k = 1, \dots, K \quad (2.19)$$

Таким чином, подібно до бінарної логістичної регресії, тут softmax-функція також забезпечує гладку залежність ймовірностей від ознак. Наступним кроком є об’єднання цих індивідуальних ймовірностей в єдину метрику, що описує якість моделі на всій вибірці. Для цього вводиться логарифмічна функція правдоподібності (log-likelihood) [45].

Функція правдоподібності моделей (2.20) має наступний вигляд:

$$\mathcal{L}(\beta) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \ln P(Y_i = k|x_i) \quad (2.20)$$

де y_{ik} – змінна для індикаторного кодування міток [45], і вона визначається як (2.21):

$$y_{ik} = \begin{cases} 1, & \text{якщо } Y_i = k \\ 0, & \text{інакше} \end{cases} \quad (2.21)$$

На практиці, оптимізаційна задача фокусується на мінімізації негативної log-likelihood функції за формулою (2.22), що інтерпретується як сума крос-ентропій між емпіричним розподілом міток та модельними ймовірностями по класах [45].

$$-\log L = -\sum_{i=1}^n \sum_{j=1}^K \delta_{j,y_i} \log(P(Y_i = j)) = -\sum_{j=1}^K \sum_{y_i=j} \log(P(Y_i = j)) \quad (2.22)$$

Для зменшення ризику перенавчання моделей мультиноміальної логістичної регресії та підвищення стабільності її оцінок досить розповсюдженим є використання регуляризації. Одним з варіантів є L2-регуляризація (2.23), де до цільової функції додається штраф, у разі великої норми векторів коефіцієнтів [45]:

$$-\mathcal{L}(\beta) + \frac{\lambda}{2} \sum_{k=1}^K \|\beta_k\|_2^2 \quad (2.23)$$

де $\lambda > 0$ – це параметр регуляризації. Це дозволяє обмежити величину коефіцієнтів і збільшити стійкість оцінок, якщо є проблема мультиколінеарності [45].

Таким чином, мультиноміальна логістична регресія поєднує в собі імовірнісний підхід до прогнозу, гнучкість для кількох класів і відносно просте навчання – усе це робить її цілком придатною для аналізу та класифікації комплексних радіомічних даних [45].

2.4.2 Випадковий ліс

Одним із найбільш популярних алгоритмів машинного навчання є Випадковий ліс (Random Forest). Його основна ідея полягає у формуванні ансамблю дерев рішень, які у свою чергу, навчаються на певній випадковій підвибірці об'єктів і ознак. У задачах класифікації фінальний прогноз отримується шляхом агрегації передбачень кожного з дерев, і є найбільш частою категоріальною змінною, що обрана через більшість «голосів» (рис. 2.7) [31, с. 28; 46; 47; 74].

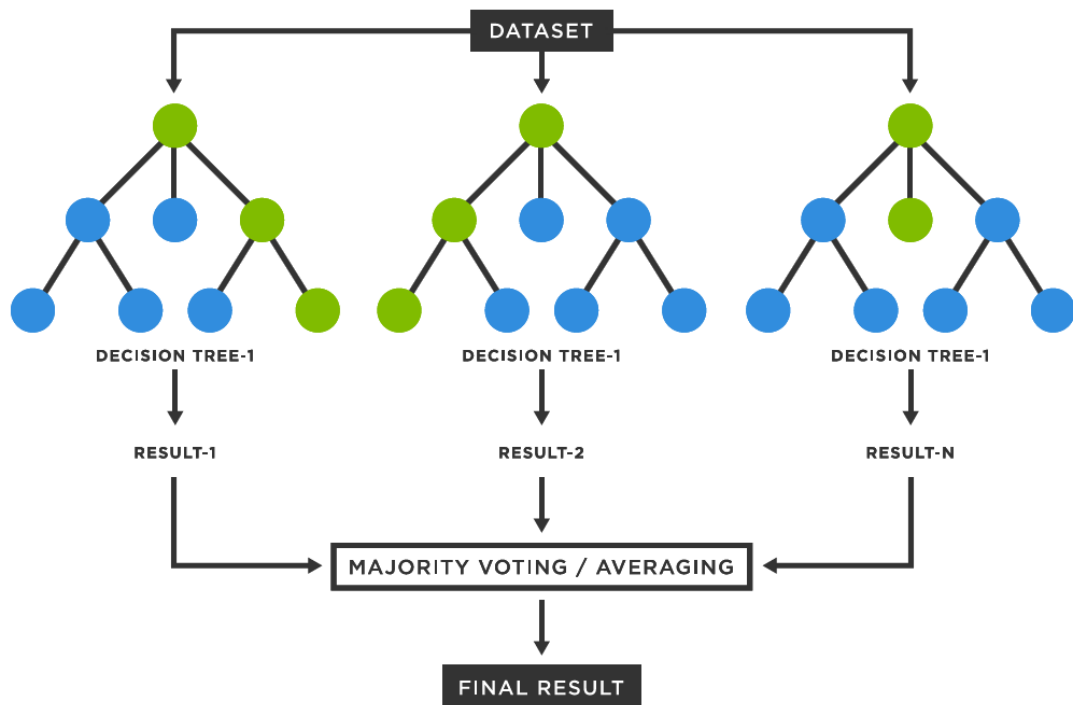


Рисунок 2.7 – Схематична ілюстрація роботи алгоритму Випадкового лісу [48]

Для керування складністю окремих дерев та ступенем «випадковості» сформованого ансамблю використовують певні гіперпараметри. Основними є такі параметри, як:

- `n_estimators` (кількість дерев у лісі) – впливає на стабільність та час навчання, адже чим більше буде дерев в моделі, тим меншою стане варіативність, проте при цьому збільшаться обчислювальні витрати [48];
- `max_depth` (глибина дерев) – обмежує складність моделі шляхом регулювання кількості рівнів, яке може мати дерево рішень. Важливо бути обережним, адже занадто велике значення у цьому параметрі може викликати *overfitting* (перенавчання), а замале – недонавчання [48];
- `min_samples_split` (мінімальна кількість прикладів для розбиття вузла) – визначає скільки об'єктів має містити вузол, щоб його ще можна було розбити надалі. Чим більшим є цей параметр, тим

«менш» дрібне дерево, гладкіші розбиття і менша схильність до перенавчання, і навпаки [48];

- `min_samples_leaf` (мінімальна кількість прикладів у листі) – задається яким має бути кінцевий розмір листового вузла, що у свою чергу, впливає на чутливість моделі до шуму [48];
- `max_features` (кількість ознак, що розглядаються при кожному розбитті) – контролює скільки ознак має випадково відбиратися на кожному кроці для побудови вузла [48].

Також одним з ключових параметрів є і `bootstrap`, що слугує для навчання окремих дереві на бутстреп-вибірках вихідних даних – це зменшує кореляцію між деревами та дисперсію моделі, покращуючи її здатність до узагальнення. Крім того, можна за допомогою параметру `criterion` визначати міру якості розиття вузла дерева: це може бути індекс Джині (`gini`) або ентропія (`entropy`) [48].

Загалом, Random Forest можна розглядати як компроміс між потужністю й стійкістю моделі, оскільки завдяки ансамблю випадково побудованих дерев він досить добре узагальнює дані та зменшує ризик перенавчання порівняно з одним деревом [47].

2.4.3 Метод опорних векторів

Метод опорних векторів (SVM) є алгоритмом, що займається створенням гіперплощини, яка максимізує відстань між класами у навчальній підвибірці. Якою саме і де буде ця межа визначають приклади, які є найближчими до неї – це і є опорні вектори, що фігурують у назві. Крім того, використання ядрових функцій дозволяє SVM неявно відображати дані в просторі вищої розмірності, де складні нелінійні патерни стають лінійно відокремлюваними від решти [49]. Схематично даний процес відображено на рис. 2.8.

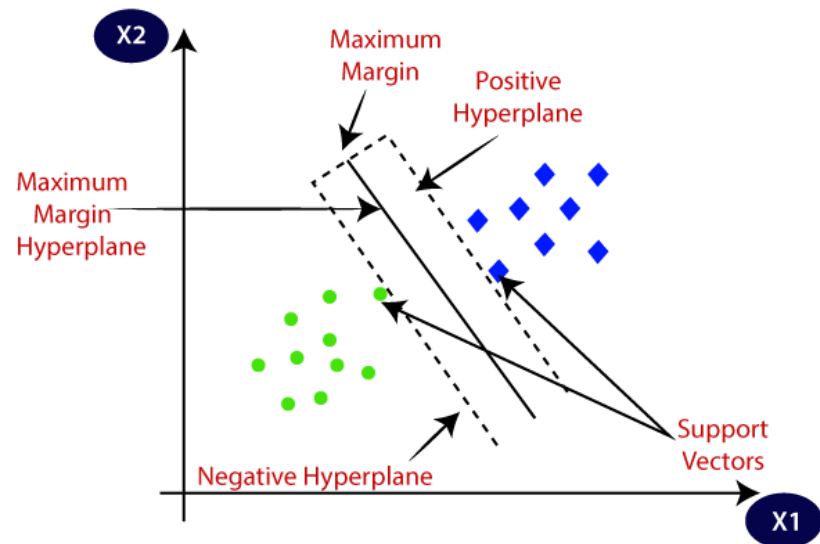


Рисунок 2.8 – Схематична ілюстрація роботи Методу опорних векторів [50]

До найбільш поширених ядер відносять радіально-базисне (rbf), поліноміальне (poly) і сигмоїдальне (sigmoid) – саме обраний тип ядра та його налаштування будуть визначати гнучкість моделі та її чутливість до локальної структури даних [49].

Існує два типи Support Vector Machine: лінійний та нелінійний. У випадку з лінійним, тут використовується лінійний тип ядра, що реалізує прямолінійну межу між класами. Він є найбільш ефективним в обчисленнях та легко інтерпретується, проте його доцільно використовувати, коли дані є (або майже є) лінійно роздільними, оскільки він створює просту прямолінійну межу рішення [31, с. 29-30; 51].

У протилежному випадку використовується нелінійний тип SVM, шляхом застосування таких функцій ядра, що надають змогу неявно відобразити дані і знайти межу рішення. Такий підхід дуже добре працює з більш складними ознаками, і відповідно, може мати більшу точність на класифікації [31, с. 30; 51].

Загалом SVM добре працює у високорозмірних просторах і часто демонструє високу якість на невеликих і середніх за розміром наборах, проте він є досить чутливим до масштабування ознак та вибору гіперпараметрів. Насамперед це стосується параметра регуляризації C та параметрів ядра, тож

у даному випадку варто виконувати стандартизацію ознак та крос-валідаційний добір параметрів. Ще однією особливістю є те, що базовий вихід моделі (відстань до межі) не є ймовірністю, а тому для використання у клінічних сценаріях його калібрують (наприклад, за допомогою Platt scaling), щоб отримати ймовірнісні оцінки, узгоджені з емпіричними частотами [49].

2.4.4 Метод k-найближчих сусідів

Метод k-найближчих сусідів можна назвати таким собі «ледачим» класифікатором, який не будує явної моделі. Його основна ідея полягає в тому, аби знайти k найближчих навчальних зразків за допомогою обраної метрики та привласнити клас за правилом більшості (рис. 2.9) [52].

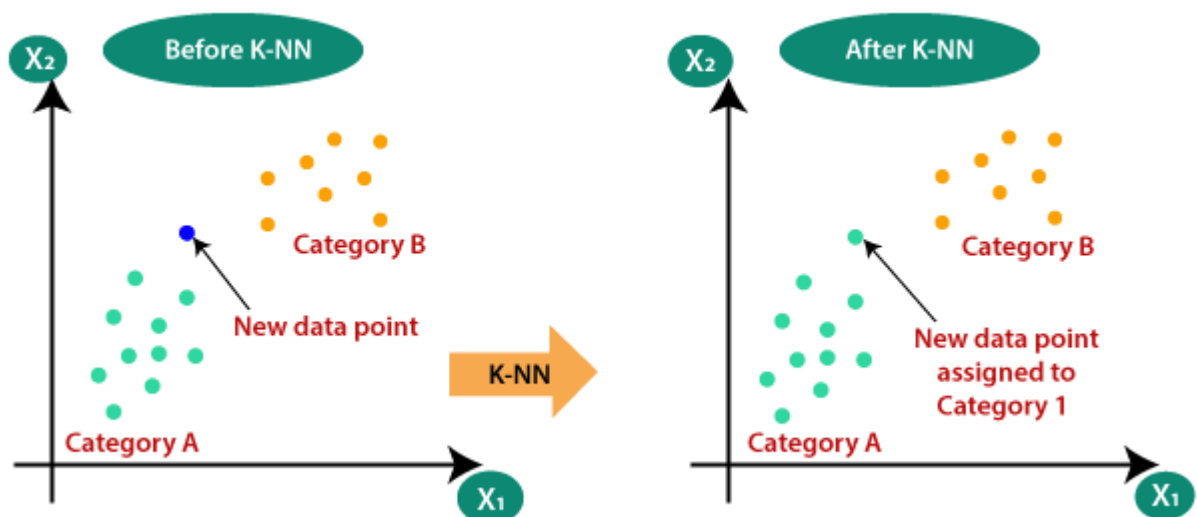


Рисунок 2.9 – Схематична ілюстрація роботи Методу k-найближчих сусідів [53]

До вищезгаданих метрик належать Евклідова відстань (2.24), Манхеттенська відстань (2.25), відстань Мінковського (2.26) та відстань Чебишева (2.27) [31, с. 30; 54].

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.24)$$

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (2.25)$$

$$d(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (2.26)$$

$$d(x, y) = \max_i |x_i - y_i| \quad (2.27)$$

Загалом, даному алгоритму не потрібне навчання моделі до того моменту, доки не буде потреби виконати прогноз. Крім того, він зберігає тренувальні дані й використовує їх уже безпосередньо під час процесу класифікації [31, с. 31-32].

Все це робить KNN досить простим у використанні й реалізації, проте зі зростанням обсягу навчальної вибірки зростає й обчислювальна вартість передбачення. Крім того, чутливість до високої розмірності простору ознак (зокрема до концентрації метричних відстаней) зумовлює потребу в нормалізації, відборі ознак або ж редукції вимірності [31, с. 32; 54].

2.4.5 XGBoost

Переходячи від базових моделей до більш гнучких ансамблів, розглянемо eXtreme Gradient Boosting, або більш звична аббревіатура – XGBoost. Даний алгоритм є покращеним варіантом градієнтного бустингу, що у свою чергу, посилює його продуктивність і ефективність створення моделей [31, с. 32].

Градієнтний бустинг – це ансамблевий підхід, за якого окремі базові моделі навчаються послідовно, і кожна наступна модель будується так, щоб скоригувати помилки в прогнозах попередньої. Подібний принцип роботи дозволяє поетапно зменшувати залишкову похибку та забезпечувати досить високу точність фінального передбачення [31, с. 32; 55]. Схематично цей процес зображено на рис. 2.10.

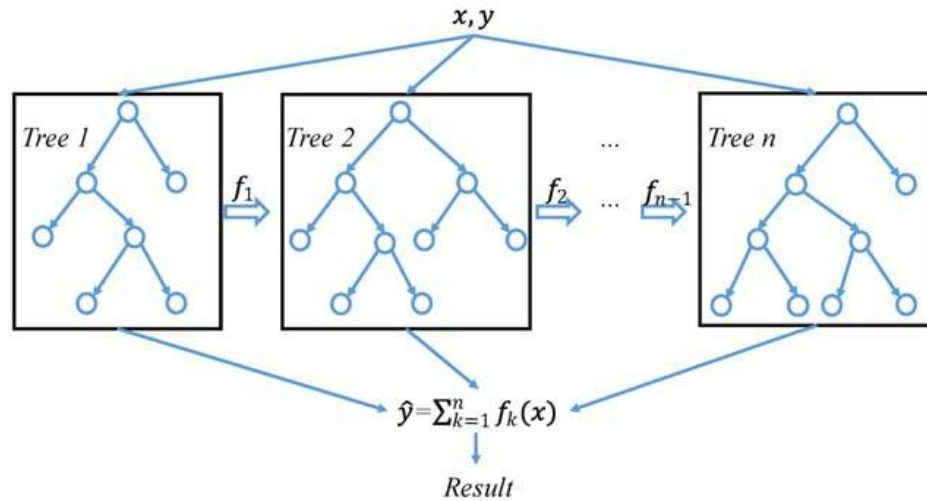


Рисунок 2.10 – Схематична ілюстрація роботи алгоритму XGBoost [56]

Загальна цільова функція даного методу має наступний вигляд (2.28) [57]:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2.28)$$

де $\sum_i l(\hat{y}_i, y_i)$ – функція втрат (loss function), яка вимірює те, наскільки передбачення $\hat{y}_i$ відрізняються від істинних значень y_i ; $\sum_k \Omega(f_k)$ – доданок для регуляризації, який штрафує надто складні дерева f_k [57].

Крім того, модель має вбудовані оцінки важливості ознак, що забезпечують прозорий внесок предикторів у підсумкове рішення, тим самим полегшуючи інтерпретацію моделі в прикладних сценаріях [55].

Володіючи високою продуктивністю та точністю класифікації, цей алгоритм входить до числа найпопулярніших, серед усіх методів машинного навчання у медичних задачах класифікації та ризик-стратифікації [31, с. 32]. Також завдяки використанню зовнішньої пам'яті він пристосований до роботи з великими даними, тож може працювати навіть якщо ресурси обмежено [31, с. 32-33]. Проте варто пам'ятати, що він є досить чутливим до налаштування гіперпараметрів і враховувати це під час впровадження [55].

2.4.6 LightGBM

LightGBM (Light Gradient Boosting Machine) – ще один метод, який реалізовано на основі градієнтного бустингу, розроблений компанією Microsoft. Проте, його особливістю є наявність певних архітектурних оптимізацій для забезпечення роботи з великими обсягами даних.

Зокрема, на відміну від типової стратегії XGBoost (level-wise), LightGBM зростає листово (leaf-wise) – дерево росте асиметрично, обираючи для розбиття той листок (на будь-якому рівні), який забезпечує найбільше зменшення функції втрат (рис. 2.11). Разом із цим підхід доповнюють за допомогою GOSS (Gradient-based One-Side Sampling) для відбору найбільш інформативних спостережень та EFB (Exclusive Feature Bundling) для об'єднання взаємовиключних ознак – це дозволяє мінімізувати обчислювальні витрати й вимоги до пам'яті, що в свою чергу пришвидшує навчання [58].

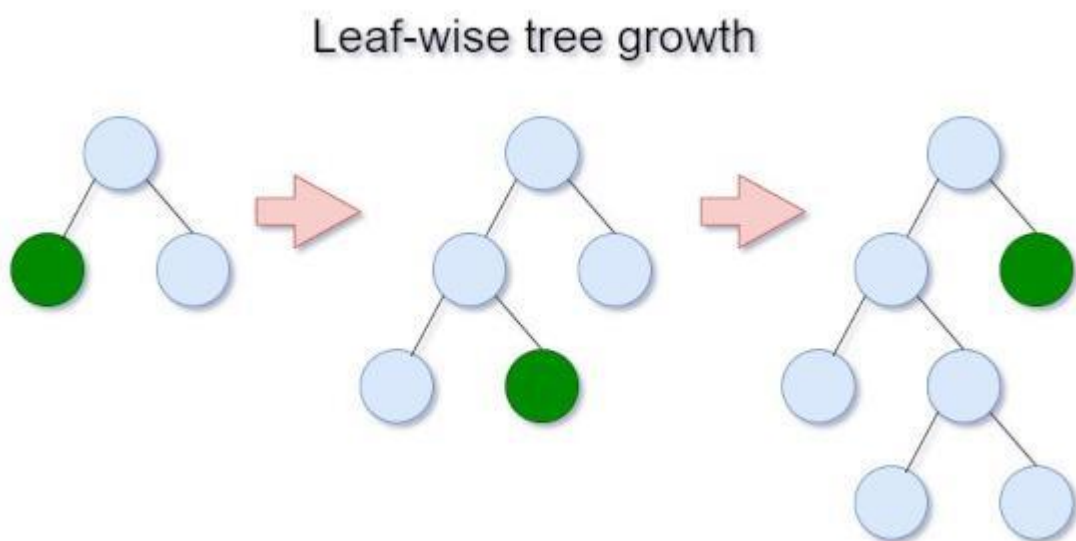


Рисунок 2.11 – Схематична ілюстрація роботи алгоритму LightGBM з leaf-wise стратегією [59]

Крім того, бінінг дозволяє алгоритму перетворювати неперервні ознаки в дискретні значення, що також робить алгоритм ще швидшим. Також, подібно до попередніх алгоритмів, LightGBM має можливість автоматично оброблювати пропущені значення [31, с. 33].

Як і XGBoost, LightGBM є чутливим до гіперпараметрів, тож у разі їх некоректного підбору можливе перенавчання моделі або нестабільні оцінки. Якщо ж налаштування виконано правильно, то LightGBM поєднуватиме високу швидкість із конкурентною якістю, що робить його дуже привабливим для виробничих конвеєрів і сценаріїв з ітеративною оптимізацією моделей [58].

2.5 Метрики оцінювання

Оцінювання якості моделей є невід’ємним етапом побудови будь-якої системи машинного навчання, оскільки саме воно визначає, наскільки отримані результати є надійними та придатними для практичного використання.

Якість моделей класифікації зазвичай оцінюють на основі зіставлення спрогнозованих міток із фактичними класами та аналізу похибок. Для бінарної задачі результати узагальнюються за допомогою матриці похибок (confusion matrix), елементи якої відображають кількість істинно-позитивних (TP), істинно-негативних (TN), хибно-позитивних (FP) та хибно-негативних (FN) рішень. На основі цих чотирьох величин визначаються основні метрики ефективності класифікатора. Для задачі мультикласифікації також використовується confusion matrix, тільки її розмірність $K \times K$, де K – кількість класів, а TP/FP/FN/TN визначаються окремо для кожного класу за схемою «один проти всіх» [60].

На основі вищезазначених величин визначаються основні метрики, а саме: загальна точність (accuracy), повнота (recall), прецизійність (precision), специфічність (specificity) та f1-міра (f1-score).

Загальна точність моделі (2.29) є найбільш простою інтегральною характеристикою, яка визначає частку правильно класифікованих спостережень серед усіх наявних прикладів [60].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.29)$$

Ця метрика дає більш узагальнене уявлення про якість моделі, тому в разі значного дисбалансу класів може створювати оманливе враження про високу якість класифікації, через вирішальний вплив домінуючого класу [60].

Для більш детального аналізу роботи моделі на позитивному класі використовують прецизійність (2.30) та повноту (2.31):

$$Precision = \frac{TP}{TP + FP} \quad (2.30)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.31)$$

Прецизійність відображає частку позитивних прикладів, що визначені коректно, серед усіх спостережень для яких модель надала позитивний прогноз. Натомість, повнота демонструє частку фактично позитивних результатів, що були правильно передбачені моделлю. Це критично важливо в задачах, де метою є виявлення або ж класифікація захворювання, оскільки пропуск позитивних результатів є недопустимим [60].

Щодо специфічності, то вона характеризує здатність моделі правильно виділяти негативний клас і вона визначається як частка істинно-негативних прикладів серед усіх фактично негативних (2.32) [60]:

$$Specificity = \frac{TN}{TN + FP} \quad (2.32)$$

За допомогою цього ми можемо аналізувати наскільки добре модель «не спрацьовує» на прикладах, що не належать до позитивного класу. У медичному контексті висока специфічність означає малу кількість необґрунтованих підозр на патологію [60].

Повертаючись до прецизійності та повноти, варто зазначити, що вони перебувають у взаємозалежному співвідношенні, коли поліпшення однієї метрики зазвичай погіршує іншу. Тому використовується F1-міра, яка об'єднує їх у єдиний показник як гармонійне середнє та розраховується за формулою (2.33) [60].

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (2.33)$$

F1-міра набуває високих значень лише тоді, коли і точність, і повнота є досить високими, тому її часто використовують як основний інтегральний критерій у задачах із дисбалансом класів, що є частим явищем для медичної інформатики та біоінформатики [60].

У багатокласовій класифікації F1-міру можна узагальнити, обчислюючи її окремо для кожного класу (за схемою «один проти всіх»), а потім агрегуючи отримані значення. Макро F1-міра (F1-macro) визначається як звичайне арифметичне середнє значень F1-міри по всіх класах і однаково зважує кожен з них, незалежно від розміру [60]. Її формула (2.34) має наступний вигляд:

$$F1\ macro = \frac{1}{K} \sum_{k=1}^K F1_k \quad (2.34)$$

Існує також зважена F1-міра (F1-weighted), яка враховує підтримку класу – кількість об'єктів у ньому. Вона обчислюється як зважене середнє F1-міри для кожного класу, де ваги пропорційні кількості об'єктів у цьому класі [60]. Це відображено у формулі (2.35):

$$F1\ weighted = \frac{1}{N} \sum_{k=1}^K n_k F1\ score_k \quad (2.35)$$

де n_k – кількість об'єктів k -го класу, а $N = \sum_{k=1}^K n_k$ – загальне число спостережень [60].

У сукупності всі ці метрики дозволяють оцінювати не тільки загальну частку правильних рішень та баланс між різними типами помилок, а і коректність роботи моделі за наявності дисбалансу класів.

Висновок з розділу 2

У даному розділі було розглянуто теоретичну основу для подальшої побудови системи диференціальної діагностики захворювань легень. Насамперед було детально охарактеризовано вихідний датасет КТ-зрізів, де показано наявність класового дисбалансу та внутрішньокласової неоднорідності, що в свою чергу ускладнює навчання моделей та коректну оцінку їхньої якості. Надалі подано структурований опис методів виокремлення текстурних ознак: GLCM, GLDS, GLRLM та LBP, які забезпечують інтерпретований радіомічний опис структури легеневої паренхіми. Оскільки вони є чутливими до варіацій сканерів, окрему увагу приділено методу RELIEF, як сучасному підходу до гармонізації ознак між різними сканерами.

На завершення розділу було систематизовано класичні моделі машинного навчання (мультиноміальна логістична регресія, Random Forest, SVM, KNN, XGBoost, LightGBM) та метрики їх оцінювання з урахуванням дисбалансу класів, що створює цілісну методологічну базу для аналізу й порівняння результатів у наступних розділах.

РОЗДІЛ 3

ОБҐРУНТУВАННЯ ВИБОРУ МЕТОДІВ ТА ІНСТРУМЕНТІВ

У цьому розділі МД наведено аргументацію щодо вибору методів та інструментів, на яких базується програмна реалізація системи диференціальної діагностики захворювань легень. Беручи до уваги вимоги до системи та узагальнюючи теоретичні положення попередніх розділів тут узгоджуються підходи до організації обчислювального середовища, побудови конвеєра обробки даних, формування ознак та вибору моделей класифікації. Крім того, окремо окреслюються принципи роботи з медичними даними, включно з мінімізацією ризиків витоку інформації та дотриманням вимог до конфіденційності.

3.1. Обчислювальне середовище та програмний стек

Розробка MVP та проведення експериментів виконувалися на персональному комп'ютері з 64-бітною операційною системою сімейства Windows 10, 8 ГБ оперативної пам'яті (RAM), процесором Intel Core i5-8265U та SSD-накопичувачем обсягом близько 500 ГБ. Такий об'єм RAM виявився на межі, тож рекомендується обсяг оперативної пам'яті не менше 16 ГБ, щоб це було достатньо для завантаження таблиць радіомічних ознак, виконання процедур гармонізації, крос-валідації та якісного навчання ансамблевих моделей класифікації без суттєвих обмежень за ресурсами.

Основним інструментом реалізації є мова програмування Python, що завдяки своїй розвиненій екосистемі пакетів орієнтована на наукові обчислення, обробку даних та машинне навчання [61]. Розробка здійснювалася у середовищі інтегрованої розробки (IDE) з використанням віртуального оточення, що забезпечує ізоляцію залежностей, фіксацію версій пакетів та відтворюваність експериментів [62].

Програмний стек проєкту, у свою чергу, включає засоби для:

- роботи з числовими масивами та табличними даними;
- виокремлення радіомічних ознак;
- проведення статистичної гармонізації даних;
- реалізації класичних алгоритмів машинного навчання для класифікації;
- обробки та аналізу медичних зображень у форматі DICOM;
- побудови настільного графічного інтерфейсу користувача та формування підсумкових звітів у форматі PDF.

3.2 Архітектура конвеєра обробки даних

Архітектура методологічного конвеєра для обробки даних у розробленій системі побудована таким чином, щоб чітко розділити етапи підготовки даних, побудови моделей та їх подальшого використання в клінічному сценарії, але водночас забезпечити єдиний наскрізний маршрут проходження даних.

На першому кроці здійснюється завантаження DICOM-знімків, які надалі нормалізуються до певного узгодженого формату, включно з базовим контролем коректності метаданих. Після чого, застосовуючи сегментовані маски легень (виділяючи ROI), виконується інтенсивнісна нормалізація та виконання зображення в режимі «легеневе вікно». Як результат, з кожного зрізу формується набір інтерпретованих текстурних ознак методами GLCM, GLDS, GLRLM, LBP, що становлять собою єдиний числовий опис структури легеневої тканини.

Окремим блоком у конвеєрі виділено гармонізацію міжсайтової варіабельності ознак. На цьому етапі вихідні вектори радіомічних параметрів, що доповнені інформацією про сайт, перетворюються таким чином, щоб зменшити залежність розподілів ознак від технічних відмінностей сканерів і протоколів. Уже готові гармонізовані ознаки використовуються для навчання набору класифікаційних моделей та оцінювання їх якості.

На етапі інференсу в складі програмного застосунку використовується той самий набір блоків попередньої обробки: завантажений користувачем DICOM-знімок послідовно проходить через сегментацію, обчислення текстурних ознак та гармонізацію, після чого отриманий вектор ознак подається на вхід раніше навчених класифікаційних моделей. Архітектура конвеєра таким чином забезпечує відтворюваність результатів між етапами навчання та використання, можливість подальшого масштабування та можливість відстеження кожного кроку обробки даних.

3.3 Вибір та параметризація текстурних ознак

Для подальшого аналізу було обрано саме ці чотири родини ознак (GLCM, GLDS, GLRLM та LBP), оскільки вони зосереджені на кількісній оцінці просторової організації та мікроструктури легеневої паренхіми.

Інші родини, такі як статистики першого порядку (FOS), були виключені, оскільки вони описують лише загальний розподіл інтенсивностей (гістограму) і не містять жодної інформації про просторові взаємозв'язки між пікселями, що є ключовим для аналізу текстури. Ознаки форми (Shape) також не було включено, оскільки вони характеризують геометрію всієї маски легень, а не внутрішню структуру ураженої тканини, що є предметом даного дослідження.

3.3.1 Gray Level Co-occurrence Matrix

Почнемо з GLCM-ознак: вони фіксують те, наскільки виразними є переходи інтенсивностей у тканині, наскільки вона зерниста чи однорідна, а також наскільки закономірно повторюються певні локальні патерни. Усе це дозволяє вийти за межі простих характеристик розподілу інтенсивностей та, натомість, кількісно описати просторову організацію легеневої паренхіми [27].

У практичній реалізації для КТ легень зазвичай використовують кілька кутів і подальше усереднення відповідних матриць, що зменшує анізотропні

ефекти, а також ретельно узгоджують діапазон інтенсивностей та кількість градацій сірого, інакше значення ознак стають нестійкими до змін протоколів. Це дає змогу кількісно оцінювати регулярність альвеолярних структур, виявляти ділянки з підвищеною зернистістю та кластеризацією інфільтратів, тим самим доповнюючи базові статистики інтенсивностей текстурними патернами, які вони самі по собі не відображають [37].

Щодо використання в моєму дослідженні, то GLCM обчислюється всередині бінарної маски легень за схемою: пікселі поза маскою кодуються як фон зі значенням 0, тоді як усі пікселі в межах маски зсуваються на один рівень (із подальшим обмеженням до верхньої межі діапазону). Надалі фоновий рівень нуля ігнорується, що в свою чергу дозволяє уникнути домінування фону в матриці суміжності рівнів сірого й зосередити аналіз саме на легеневій тканині.

3.3.2 Gray Level Difference Statistics

На відміну від матриці суміжності рівнів сірого, що враховує спільні появи пар рівнів, статистика різниць рівнів сірого концентрується саме на величині локальних змін. що робить даний метод природним індикатором локального контрасту та структурної неоднорідності зображення. Це може пояснювати той факт, що в класичних порівняльних роботах метрики на основі різниць часто демонстрували досить конкурентну ефективність поряд із другопорядковими підходами (яким є GLCM) та в окремих режимах знімання виявляли меншу чутливість до вибору параметрів [63].

У контексті медичної візуалізації GLDS є придатним інструментом для кількісної характеристики локального контрасту легеневої паренхіми, що полегшує розрізнення дрібнозернистих інфільтратів і більш гомогенних структур [64]. Крім того, він методично доповнює GLCM і GLRLM, підвищуючи стійкість класифікацій, коли міжкласові відмінності зумовлені інтенсивністю переходів, а не орієнтаційними патернами [65].

Під час практичної реалізації використовувався набір векторів зсуву $(D_x, D_y) = \{(0,1), (1,1), (1,0), (1,-1)\}$, які відповідають чотирьом основним

напрямкам у площині зрізу: горизонтальний, вертикальний та дві діагоналі, на одиничній відстані. Для кожного з цих зсувів у межах маски відбираються всі допустимі пари пікселів «центральный–сусід», для яких обидві точки належать легеневій області, і обчислюється абсолютна різниця їхніх квантованих рівнів сірого. Таким чином, на основі отриманих значень формується емпіричний розподіл локальних різниць, який потім узагальнюється в набір статистик першого порядку.

3.3.3 Gray Level Run Length Matrix

Для застосування GLRLM в ході роботи застосовується той самий препроцесинг зображення, що і для решти методів текстурного аналізу. У результаті цього формується двовимірне зображення типу uint8 з діапазоном значень від 0 до 63. Визначення області інтересу також аналогічне: пікселі, що належать легеневій паренхімі, позначаються як True, решта – як False. Крім того, перед обчисленням ознак додатково застосовується фільтр за мінімальним розміром ROI, встановленим на рівні 64 пікселів: якщо кількість істинних пікселів у масці менша за цей поріг, зріз пропускається і ознаки для нього не обчислюються, для того аби уникнути нестабільних значень текстурних характеристик на надто малих областях.

Також, безпосередньо перед розрахунком матриці довжин пробігу рівнів сірого перевіряється ситуація з константними зрізами. Мається на увазі, що у разі, коли всі значення двовимірних зображень однакові, виконується «деконстантизація» одного пікселя – копія зображення модифікується так, що в позиції (0, 0) рівень сірого змінюється на сусіднє значення (по модулю NG_LEVELS). Такий підхід дозволяє уникнути числових збоїв при побудові матриць пробігів на повністю однорідних зображеннях, не змінюючи при цьому помітно структуру даних у межах маски.

У КТ легень GLRLM доречна, коли потрібно кількісно охарактеризувати смугасті або протяжні однорідні структури. До прикладу, довгі серії можуть відповідати ділянкам консолідації чи фіброзним змінам, тоді як переважання коротких серій відображає високу неоднорідність і дрібнозернисті

інфільтрати [66]. Окрема його перевага полягає у здатності відділяти ефекти «серійної» однорідності від загальної мінливості інтенсивностей. Це досить важливо для легеневої паренхіми, адже тут можуть співіснувати великі зони відносної однорідності та дрібні вогнища неоднорідності. До того ж, поєднання GLRLM із GLCM зазвичай підвищує дискримінативність радіомічних панелей, оскільки перша група фіксує протяжність однотипних фрагментів, а друга кодує локальні парні залежності [35; 66].

3.3.4 Local Binary Pattern

Наостанок, розглянемо метод локальних бінарних шаблонів. Як вже згадувалось в теоретичному описі, він досить добре відгукується на дрібні коливання мікроструктури, релевантні для ранніх інфільтративних змін і дифузних процесів, що може бути корисним для діагностування ранніх стадій захворювання.

Крім того, для того щоб охопити різні просторові масштаби, на практиці застосовуються багатомасштабні LBP, які здатні поєднувати кілька пар (P, R) і зважено конкатенують відповідні гістограми. Перевагою такої реалізації є те, що покривається спектр мікропатернів від дуже дрібних «зернин» до більших локальних неоднорідностей. У складі радіомічних панелей це допомагає LBP доповнювати другопорядкові статистики, додаючи орієнтаційно-локальну мікроструктурну інформацію, яку GLCM та GLRLM можуть фіксувати дещо опосередковано – це, в свою чергу, підсилює дискримінативність і відтворюваність моделей [39].

У реалізованій системі було вирішено використати багатомасштабний підхід, у якому використовується набір конфігурацій (P, R) з кількістю сусідів $P = \{8, 16, 24\}$ та відповідними радіусами $R = \{1, 2, 3\}$. Такий вибір дозволяє при малому радіусі фіксувати тонкі інтерстиціальні потовщення та «матове скло», тоді як більші радіуси краще захоплюють грубішу ретикулярність, ознаки «бруківки» (crazy paving) або стільникові зміни. Отримані для кожного з цих трьох масштабів гістограми ознак об'єднуються в єдиний вектор, що

забезпечує комплексний опис текстури легеневої паренхіми одночасно на різних просторових рівнях.

3.4 Обґрунтування вибору та конфігурації методу RELIEF для гармонізації ознак

Вихідна вибірка має мультицентрову природу, оскільки дані отримані з різних комп'ютерних томографів, що відрізняються виробником, типами реконструкційних фільтрів та параметрами протоколу сканування. Як показано у розділі 1, такі технічні відмінності здатні суттєво впливати на розподіли радіомічних ознак і створювати виражений «ефект сканера», який маскує справжні біологічні відмінності між нозологіями.

Одноваріативні методи гармонізації, подібні до ComBat або CovBat, коригують середні та дисперсії окремих ознак, але не моделюють складні багатовимірні патерни варіацій, властиві конкретним сканерам [42; 43]. Тож з огляду на обмежений обсяг вибірки та значну кількість текстурних ознак, у даній роботі було прийнято рішення використовувати RELIEF як основний метод структурованої гармонізації, здатний одночасно враховувати спільні та сканер-специфічні латентні фактори і тим самим зменшувати міжсайтову варіабельність без пригнічення корисних сигналів.

Додатковим аргументом на користь RELIEF є його орієнтація на збереження біологічно значущої варіабельності. Адже у контексті поставленої задачі диференціальної діагностики критично важливо не лише «вирівняти» дані між центрами, але й забезпечити здатність моделей і надалі розрізняти COVID-19, здорові легені, бактеріальну пневмонію та постковідні зміни (long-COVID).

У даній роботі RELIEF застосовується до матриці радіомічних ознак, попередньо нормалізованих та очищених від некоректних значень, з урахуванням сайту сканування.

3.5 Вибір моделей класифікації

Для вирішення задач мультикласової класифікації та бінарного виявлення ознак СЗЛ було обрано набір із шести алгоритмів машинного навчання різної природи. Такий різнорідний підхід дозволяє порівняти ефективність лінійних, метричних, ансамблевих та бустингових методів на радіомічних даних.

Оскільки вибірка характеризується обмеженим обсягом (97 пацієнтів) та дисбалансом класів, при формуванні набору моделей враховувалася необхідність застосування різних стратегій боротьби з перенавчанням. Такий підхід дозволяє порівняти, яка саме стратегія (лінійна, метрична чи ансамблева) виявиться найбільш робастною в умовах малої вибірки.

3.5.1 Мультиноміальна логістична регресія

У практичних застосуваннях регресії, у тому числі мультиноміальної, підтримується регуляризація L1, L2 або їх комбінація (elastic net). Це дає змогу зменшити варіативність оцінок і впоратися з надлишковими чи корельованими предикторами [67].

Щоб гарантувати максимальну ефективність методу на специфічних радіомічних даних, у даній роботі стратегія налаштування передбачає емпіричний вибір типу та сили регуляризації шляхом перебору по сітці (Grid Search). Такий підхід дозволить моделі адаптивно визначати, що є пріоритетнішим: відбір ознак (L1) чи боротьба з мультиколінеарністю (L2).

Критично важливою для даної роботи є здатність методу повертати узгоджені ймовірності приналежності до кожного з класів завдяки Softmax-перетворенню. Це дозволяє використовувати отримані оцінки не лише для жорсткої класифікації, а й як вхідні дані для ансамблю моделей та як міру впевненості системи для інформування лікаря [67].

Враховуючи все вищеписане, ключовими перевагами мультиноміальної логістичної регресії є швидке навчання на табличних даних, калібровані ймовірності та відносно низька схильність до перенавчання за

коректно підібраних штрафів. Водночас модель має низку обмежень: вона задає лінійну межу рішень, чутлива до мультиколінеарності та викидів, потребує ретельної інженерії ознак або нелінійних розширень для фіксації взаємодій і криволінійних ефектів [67].

Крім того, її якість може падати за сильного дисбалансу класів, тому для нівелювання цього ефекту в роботі застосовується автоматичне зважування класів (class weights). Як наслідок сукупності всіх факторів, цілком доцільно використовувати мультиноміальну логістичну регресію як базову модель для еталонного порівняння зі складнішими алгоритмами, а також як інтерпретовану складову в ансамблях або метамоделях, де нестачу гнучкості компенсують інші методи.

3.5.2 Випадковий ліс

Як вже зазначалось у теоретичному описі, Випадковий ліс – це ансамблевий алгоритм, що завдяки особливостям своєї реалізації здатний природно моделювати нелінійні взаємодії та порогові ефекти без явної інженерії ознак, а також надає оцінки важливості, що допомагає відбирати інформативні предиктори й підвищувати пояснюваність результатів [46].

У контексті даної роботи особливо важливими є властивості Випадкового лісу щодо роботи з високорозмірними табличними даними та складними кореляційними структурами між ознаками. Радіомічні показники, що використовуються в моделі, формують простір із кількома десятками текстурних ознак, причому частина з них потенційно корельована між собою. На відміну від методів, які вимагають жорсткого попереднього відбору ознак, Random Forest здатний ефективно працювати в такому просторі «як є», автоматично знижуючи вагу нерелевантних або надлишкових предикторів через механізм випадкового вибору ознак на рівні вузлів [68].

Враховуючи обмежений обсяг вибірки, стратегія налаштування моделі в даній роботі передбачає жорстке обмеження глибини дерев (max_depth) та налаштування мінімальної кількості зразків у листі (min_samples_leaf) для запобігання запам'ятовуванню окремих пацієнтів.

З урахуванням цих міркувань Випадковий ліс у даній роботі розглядається як базовий нелінійний ансамблевий метод для табличних радіомічних ознак, який доповнює логістичну регресію, забезпечуючи вищу гнучкість у фіксації складних взаємодій між ознаками та зберігаючи прийнятний рівень інтерпретованості моделі завдяки аналізу важливостей предикторів та стабільності оцінок на мультицентровому датасеті.

3.5.3 Метод опорних векторів

SVM є одним із найпотужніших алгоритмів для роботи з малими наборами даних у багатовимірному просторі. Завдяки різним варіантам налаштування ядер є можливість гнучко моделювати як лінійні, так і складні нелінійні межі рішень у просторі ознак. Головною перевагою методу опорних векторів вважається саме той максимальний відступ між класами, про який згадувалось у теоретичному описі роботи алгоритму. Завдяки цьому він демонструє високу точність, а також це забезпечує його стійкість до перенавчання. Окрім того, робота з нелінійними даними робить його універсальним інструментом для вирішення задач класифікації [31, с. 30].

Проте варто пам'ятати його чутливість до вибору гіперпараметрів, і насамперед це стосується параметра регуляризації C та параметрів ядра. Враховуючи це важливим є виконання стандартизації ознак та застосування крос-валідаційного методу підбору параметрів [51]. Ще однією особливістю є те, що базовий вихід моделі (відстань до межі) не є ймовірністю, а тому для того, щоб використовувати Метод опорних векторів у клінічних сценаріях було застосовано калібрування методом Платта (`probability=True`) для того, аби отримати ймовірнісні оцінки, узгоджені з емпіричними частотами [69].

3.5.4 Метод k-найближчих сусідів

Незважаючи на недоліки, що були описані в теоретичному розділі, метод k-найближчих сусідів включено до дослідження як представника класу метричних алгоритмів. Його вибір базується на гіпотезі локальної подібності. Її суть полягає в припущенні, що пацієнти зі схожими текстурними патернами матимуть однаковий діагноз. Також перевагами KNN у даному контексті є

відсутність фази явного навчання та здатність моделювати складні нелінійні межі рішень без припущень про розподіл даних [54].

Щодо практичної реалізації, критично важливою умовою коректної роботи методу k -найближчих сусідів є попередня стандартизація ознак, оскільки він чутливий до різниці в масштабах. Стратегія налаштування передбачає використання крос-валідації з GridSearch для пошуку не лише оптимальної кількості сусідів, але й метрики відстані та стратегії зважування, що дозволяє надати більшу вагу найближчим до тестового зразка прикладам [53].

3.5.5 XGBoost

Володіючи високою продуктивністю та точністю класифікації, цей алгоритм входить до числа найпопулярніших, серед усіх методів машинного навчання [31, с. 32]. Крім того, вбудована обробка пропущених даних та наявність механізмів регуляризації (L1, L2) роблять його гнучким інструментом, здатним контролювати складність моделі навіть на обмежених наборах даних [70].

Проте, повертаючись до особливостей всіх попередніх алгоритмів, XGBoost також є досить чутливим до налаштування гіперпараметрів, особливо враховуючи невеликий розмір вихідної підвибірki [70]. Таким чином у роботі було приділено особливу увагу коригуванню глибини дерев, кількості листків, швидкості навчання, а також частки підсемплювання прикладів і ознак, оскільки саме ці параметри визначають баланс між гнучкістю та узагальнюваністю.

Як результат, при помірній кількості ознак і розмірі вибірки XGBoost демонструє високу роздільну здатність без складної інженерії ознак, що й пояснює його популярність у медичних задачах класифікації та ризик-стратифікації [55].

3.5.6 LightGBM

LightGBM включено до дослідження як альтернативний представник сімейства градієнтного бустингу, що використовує відмінну від XGBoost

стратегію побудови дерев — по-листо́ве зростання (leaf-wise growth). Ця особливість дозволяє алгоритму будувати глибші та складніші дерева, потенційно виявляючи більш тонкі закономірності в даних [58].

Водночас, на малих вибірках leaf-wise стратегія може створювати високий ризик перенавчання. Тому включення LightGBM у набір моделей супроводжується жорсткою стратегією регуляризації. Основний акцент зроблено не на швидкості, а саме на обмеженні архітектури дерев: лімітуванні максимальної глибини (max_depth) та кількості листків (num_leaves), а також встановленні високого порогу для мінімальної кількості прикладів у листі (min_data_in_leaf).

За умови коректного налаштування цих параметрів, LightGBM поєднуватиме високу швидкість із конкурентною якістю, що робить його привабливим для виробничих конвеєрів і сценаріїв з ітеративною оптимізацією моделей [58].

3.6 Запобігання витоку даних і схеми підвибірок

У дослідженнях, що базуються на медичних зображеннях, однією з ключових загроз коректності результатів є витік даних (data leakage) між етапами навчання та оцінювання моделей. У розробленому застосунку на рівні конвеєра обробки даних реалізовано низку обмежень, які мають на меті мінімізувати такі ризики як на етапі формування підвибірок, так і під час подальшої побудови моделей.

Перш за все, базовою одиницею розподілу даних є пацієнт, а не окремий КТ-зріз. Це означає, що всі зрізи, які належать одному пацієнту, жорстко закріплюються за однією з підвибірок і, відповідно, не можуть одночасно потрапляти до різних частин вибірки. Це допомагає запобігати витоку інформації через корельовані зрізи одного й того самого обстеження, адже модель не «бачить» ті самі анатомічні структури пацієнта і під час навчання, і під час оцінювання.

Формування підвибірок тренувальної, валідаційної та тестової підвибірок (train, val, test) виконується з урахуванням як класового балансу, так і сайтів сканування. В цілому, розподіл побудовано таким чином, щоб з одного боку зберегти максимально можливу репрезентативність кожного класу в навчальній та валідаційній підвибірках, а з іншого – забезпечити наявність достатньої кількості прикладів для незалежної тестової оцінки, включно з рідкісними класами. Для бактеріальної пневмонії передбачено використання п'ятифолдової перехресної перевірки з групуванням за пацієнтом, яка використовується під час навчання та порівняння класифікаційних моделей, щоб хоч трохи покращити ситуацію з цим малочисельним класом.

Окремим рішенням є сайт-орієнтований розподіл, що використовується для наближення сценарію зовнішньої валідації. Таким чином, усі дослідження, виконані на томографах виробника HITASHI було цілеспрямовано віднесено до тестової підвибірки. Вибір саме цих сканерів було зумовлено тим, що кількість відповідних пацієнтів у вибірці є мінімальною (лише двоє), тож їх включення до навчальної сукупності практично не вплинуло б на здатність моделі «пристосуватися» до цього сайту. Натомість, винесення їх до тесту дає змогу використати ці рідкісні випадки як наближену зовнішню валідацію (квазі-«зовнішній» датасет) та перевірити наскільки стійкою є модель до міжсайтової варіабельності.

Окремо контролюється можливий витік через етап гармонізації міжсайтової варіабельності. Моделі гармонізації навчаються виключно на даних навчальної сукупності і лише після цього застосовуються до ознак валідаційної та тестової підвибірок. Жодні параметри гармонізації не оновлюються на тестових даних, а отже структура розподілів у тестовій вибірці не впливає на перетворення, які потім буде використовувати модель.

Зрештою, на етапі побудови моделей додатково використовується перехресна перевірка з урахуванням групування за пацієнтом. При цьому в межах навчальної сукупності застосовуються схеми крос-валідації, де всі зрізи

одного пацієнта завжди потрапляють в один і той самий фолд. Це дозволяє запобігти ситуації, коли частина зрізів пацієнта опиняється у навчальній, а частина – у валідаційній підвибірці, що у свою чергу, викликало б переоцінку якості моделі.

3.7 Захист даних і приватність

Під час розробки даної системи диференціальної діагностики захворювань легень окрему увагу було приділено захисту медичних даних та дотриманню принципів приватності. Таким чином архітектура конвеєра проектувалася за підходом *privacy by design*, де всі етапи збору, обробки та зберігання даних організовано таким чином, щоб мінімізувати ризики несанкціонованого доступу, повторної ідентифікації пацієнтів та витоку конфіденційної інформації.

Вихідний датасет для навчання моделей було повністю анонімізовано шляхом видалення з DICOM-файлів прямих ідентифікаторів особи (П.І.Б., номер медичної картки, контактні дані, точні дати народження тощо). Для подальшої роботи в системі вводиться внутрішній псевдонімізований ідентифікатор `patient_id`, який не дозволяє однозначно встановити особу пацієнта поза межами клінічної інформаційної системи. Усі похідні артефакти (маски сегментації, таблиці ознак, результати гармонізації та класифікації) оперують виключно цим «кодом», без зберігання будь-яких полів, які могли б прямо або опосередковано ідентифікувати пацієнта.

Важливо також зазначити, що розроблюваний програмний застосунок орієнтований на локальне використання в закладі охорони здоров'я. Таким чином, завантаження КТ-зрізів, обчислення ознак та класифікація патології будуть виконуватися на робочій станції користувача (у даному випадку лікаря) без передачі зображень або ознак у зовнішні сервіси чи хмарну інфраструктуру. Це зменшує поверхню атаки та виключає ризики, що можуть бути пов'язані з переданням чутливих медичних даних мережею. Щодо

результатів у вигляді PDF-звітів, то вони містять лише узагальнені результати класифікації та мінімальну технічну інформацію, що необхідна для інтерпретації моделі, та не потребують включення персональних даних. У разі потреби, буде можлива інтеграція з клінічною системою закладу охорони здоров'я, де прив'язка до конкретного пацієнта виконуватиметься лише на стороні медичної інформаційної системи, а не в самому застосунку.

З боку організаційних заходів передбачається, що доступ до системи матимуть лише уповноважені користувачі, які працюють у межах захищеної мережі закладу, а файлові сховища з DICOM-знімками та похідними таблицями ознак розміщуються на носіях із контрольованим доступом. Щодо середовищ реального впровадження, то цілком доцільним є використання засобів шифрування дисків, резервного копіювання та журналювання доступу до даних, однак ці механізми відносяться до рівня ІТ-інфраструктури медичного закладу, а тому не реалізуються безпосередньо у програмному застосунку.

Висновок до розділу 3

У даному розділі обґрунтовано вибір обчислювального середовища та програмного стеку, необхідних для локальної реалізації конвеєра диференціальної діагностики захворювань легень. Сформовано узгоджену архітектуру обробки даних, що інтегрує етапи сегментації, обчислення інтерпретованих текстурних ознак, їх гармонізацію та подальшу класифікацію.

Окрему увагу приділено параметризації методів виокремлення ознак та конфігурації алгоритму RELIEF як основного інструменту для зменшення міжсайтової варіабельності. Також було аргументовано вибір набору моделей машинного навчання, що здатні ефективно працювати в умовах обмеженої вибірки та дисбалансу класів.

Також було систематизовано принципи запобігання витоку даних та окреслено підхід до захисту медичної інформації через анонімізацію та локальне виконання обчислень. Сукупність усіх цих рішень формує цілісну методологічну й технічну основу для практичної реалізації необхідної системи.

РОЗДІЛ 4

ПРОГРАМНА РЕАЛІЗАЦІЯ ТА МЕТОДИКА РОБОТИ ПРОГРАМНОГО ЗАСТОСУНКУ

У даному розділі розглянуто програмну реалізацію розробленої системи диференціальної діагностики захворювань легень та узагальнено методику роботи програмного застосунку. Спочатку описуються етапи попередньої обробки даних, які забезпечують підготовку вхідних КТ-знімків та відповідних ознак для подальшого аналізу моделями машинного навчання. На цій основі подається архітектура програмного застосунку, що поєднує модулі обробки даних із засобами взаємодії з користувачем і наприкінці демонструються основні сценарії роботи системи, які ілюструють практичне використання реалізованого функціоналу.

4.1 Попередня обробка даних

Перед навчанням ML-моделей та їх використанням у програмному застосунку необхідно виконати деякі етапи з попередньої обробки даних, для того, аби забезпечити належну підготовку КТ-зрізів.

4.1.1 Сегментація масок легень

Першим етапом стала побудова області інтересу (ROI) на КТ-зображеннях легень за допомогою бібліотеки lungmask, що використовує згорткові нейромережі сімейства U-Net для автоматичного виділення легеневої паренхіми й часток з урахуванням патологій. Відповідні маски генерувалися для кожного зрізу та зберігалися як двійкові зображення (фон/легені) у форматі PNG із тим самим найменуванням, що й вихідний файл, задля спрощення подальшої прив'язки до зрізів і використання в пайплайнні ознак.

Серед доступних моделей lungmask є [31, с. 46]:

- U-Net (R231) – сегментує окремі зрізи, виділяючи праву/ліву легеню, включно з повітряними кишнями та об’ємними утвореннями [31, с. 46].
- U-Net (LTRCLobes) – виконує сегментацію за частками, але чутливіша до щільних інфільтратів і може втрачати контури при виражених патологіях [31, с. 46].
- U-Net (LTRCLobes_R231) – поєднує результати двох попередніх підходів [31, с. 47].
- U-Net (R231CovidWeb) – донавчена на даних із COVID-19 для кращої стійкості до специфічних змін легеневої тканини [31, с. 47].

Оскільки датасет охоплює як здорові легені, так і випадки з інфекційно-запальними ураженнями (COVID-19, пневмонія) та постінфекційними змінами (long-COVID), сегментація має бути стійкою до консолідацій, помутнінь типу «матового скла» та інших структурних змін. І хоча базова R231 зазвичай формує коректні маски й на патологічних знімках, на зображеннях зі специфічними ураженнями модель R231CovidWeb здатна підвищити точність, оскільки має здатність краще відмежовувати паренхіму від уражених ділянок [28, с. 3; 31, с. 47; 71; 74]. Тож враховуючи все вищеописане було обрано саме модель U-Net (R231CovidWeb).

4.1.2 Обчислення текстурних ознак

Для виокремлення текстурних ознак було використано ru feats – це Python-пакет для радіоміки, що обчислює класичні текстурні дескриптори з 2D-зображень КТ. Загалом, бібліотека підтримує досить широкий спектр класичних сімейств ознак і може працювати як із цілим зображенням, так і в межах довільної ROI, якщо передати маску.

Оскільки ru feats не включає препроцесингу знімків, то перед тим, як обчислювати текстурні характеристики, було реалізовано багатоступеневу нормалізацію даних. По-перше, значення у Hounsfield одиницях обмежувались

фіксованим діапазоном $[-1350; 150]$ («легеневе вікно»). По-друге, відбувалось квантування інтенсивностей до $NG_LEVELS = 64$ рівнів сірого. По-третє (і це критично важливо), було реалізовано робастну обробку маски: пікселі поза ROI кодувалися як 0, тоді як усі 64 рівні паренхіми були зсунуті на +1 (у діапазон $[1, 64]$). Це дозволило pyfeats ігнорувати фоновий нульовий рівень і обчислювати ознаки виключно на основі легеневої тканини, уникнувши появи фонових значень у статистиці.

У результаті, для подальшого аналізу сформовано робочий піднабір із 36 текстурних ознак. Він включає:

- 14 ознак GLCM: ASM_Mean, Contrast_Mean, Correlation_Mean, SumOfSquaresVariance_Mean, InverseDifferenceMoment_Mean, SumAverage_Mean, SumVariance_Mean, SumEntropy_Mean, Entropy_Mean, DifferenceVariance_Mean, DifferenceEntropy_Mean, Information1_Mean, Information2_Mean, MaximalCorrelationCoefficient_Mean;
- 5 ознак GLDS: Homogeneity, Contrast, ASM, Entropy, Mean;
- 11 ознак GLRLM: Short Run Emphasis, Long Run Emphasis, Gray Level Non Uniformity, Run Length Non Uniformity, Run Percentage, Low Gray Level Run Emphasis, High Gray Level Run Emphasis, Short Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Long Run High Gray Level Emphasis [76, 77];
- 6 ознак LBP: R_1_P_8_energy, R_1_P_8_entropy, R_2_P_16_energy, R_2_P_16_entropy, R_3_P_24_energy, R_3_P_24_entropy.

Усі значення було збережено у CSV-файлах у форматі «один рядок на зріз» із додатковими службовими полями задля спрощення подальшого злиття наборів, коректної стратифікації та формування підвибірок.

4.1.3 Гармонізація між сайтами

У попередніх розділах вже згадувалось, що для гармонізації міжсайтових особливостей було використано такий метод, як Removal of Latent Inter-scanner Effects through Factorization. Проте є один нюанс: оригінальна реалізація RELIEF відбувається на мові програмування R, а на момент виконання даної роботи адаптацій для Python не було. У зв'язку з цим було прийнято рішення спробувати реалізувати таку адаптацію самостійно в межах магістерської роботи.

Під час адаптації методу до Python було важливо зберегти його основну постановку, проте її довелось дещо спростити. Стандартизована матриця ознак Y розглядається як сума спільної низькорозмірної компоненти R , набору сайт-специфічних низькорозмірних матриць I_b та залишкового шуму без явного моделювання перехоплень, коваріат та адитивних сайт-ефектів. Натомість, перехоплення функціонально вже враховане через попереднє центрування, а коваріати не були включені, оскільки нас цікавив саме міжсайтовий ефект у ознаках, а не корекція за клінічними чи демографічними змінними, як вік, стать тощо.

Оцінювання цих матриць реалізовано у функції `relief_decompose` як розв'язання задачі мінімізації квадратичної похибки з ядроно-нормовою регуляризацією для спільної компоненти R та кожного сайт-специфічного доданка I_b . Чисельне розв'язання виконується ітераційно за допомогою `soft-threshold`-ування сингулярних значень (SVD або randomized SVD) до досягнення збіжності за відносним покращенням цільової функції. Параметри штрафів `lam_shared` та `lam_per_site` обчислюються на основі робастної оцінки та кількості спостережень у кожному сайті, що узгоджується з оригінальною постановкою RELIEF.

Також Python-адаптацію було доповнено деякими етапами, щоб зменшити вплив викидів, стабілізувати масштаб ознак і забезпечити відтворюваність гармонізації на всіх підвбірках. Зокрема, перед декомпозицією ознаки послідовно проходять `winsorization` за MAD

(winsorize_by_mad), симетризоване логарифмування $\text{sign} * \log_{1p}(|x|)$ та робастне центровання й масштабування відносно медіани й MAD за допомогою `robust_center_scale_fit` та `robust_center_scale_apply`, що дозволяє працювати в узгодженому робастному просторі навіть за наявності шумних або сильно варіативних ознак.

Модуль має два етапи: `learn_global` для навчання та `apply_wide` для застосування гармонізації. На етапі `learn_global` перелік ознак, медіани та масштаби у SymLog-просторі, оцінені `lam_shared` і `lam_per_site` зберігаються в JSON-файлах для того, щоб мати змогу застосовувати одну й ту саму трансформацію до різних підвбірок без постійного перенавчання моделі гармонізації. Потім вмикається режим `apply_wide`, де ці артефакти повторно використовуються для підвбірок `train`, `val` і `test`. У разі, якщо у валідаційній чи тестувальній підвбірці наявні сайти, що відсутні у `train` (як HITCHI), тоді автоматично застосовується спільний штраф `lam_shared`, що дає змогу узгоджено обробляти «нові» сканери. Після віднімання сайт-специфічної компоненти дані повертаються у вихідний простір ознак через обернене робастне масштабування та інверсію SymLog ($\text{sign} * \expm1(|x|)$), і гармонізовані значення підміняють початкові ознаки в таблиці даних.

Таким чином, розроблена Python-адаптація є дещо спрощеною реалізацією оригінального методу RELIEF, що поєднує його основну логіку з додатковими кроками робастної підготовки даних. При цьому вбудована архітектура `learn/apply` спрямована на мінімізацію витоку даних під час навчання та узгоджену обробку спостережень із «нових» сайтів, застосовуючи раніше збережені артефакти. Метою такого підходу є формування більш однорідного простору ознак для подальшої побудови та оцінювання класифікаційних моделей.

4.1.4 Стратегія формування навчальної, валідаційної та тестової підвбірок

Формування підвбірок відбувається у два кроки. Спочатку пацієнти, обстежені на томографах HITCHI, цілеспрямовано відносяться до тестової

підвибірки, що дозволяє використати ці дослідження як окремий набір для підсумкової оцінки стійкості моделей до міжсайтової варіабельності. Далі, для решти пацієнтів виконується стратифікований поділ з урахуванням діагностичного класу та групування за `patient_id`: кожен пацієнт потрапляє лише до однієї з підвибірок, а частки класів у `train`, `val` і `test` наближено відтворюють початковий розподіл у сукупності.

Як результат, отримуються три окремі списки ідентифікаторів пацієнтів `train_ids`, `val_ids`, `test_ids`, які зберігаються в окремих файлах формату `csv`. Це забезпечує єдиний, фіксований варіант розподілу даних, що робить порівняння різних конфігурацій гармонізації та класифікації коректним і відтворюваним.

4.2 Архітектура програмного застосунку

Застосунок побудовано як двомодульну систему інференсу, яку запускає стартовий сценарій `main.py`. Перший модуль `MULTI` розв’язує мультикласову задачу класифікації (COVID-19, Здорова легеня, Пневмонія, Long-COVID), у той час як другий модуль `BINARY` виконує бінарне виявлення ознак СЗЛ у підмножині пацієнтів із COVID-19. Обидві ці гілки поділяють спільні підсистеми обробки даних, керування моделями та формування підсумкових результатів, що забезпечує єдині правила відтворюваності та масштабованості.

Підсистему інтеграції реалізовано як шар, незалежний від конкретних моделей, тож він може працювати з будь-якими сумісними класифікаторами. Під час першого запуску інференсу цей шар автоматично підхоплює доступні моделі з налаштованого каталогу й реєструє їхні конфігурації, після чого для коректного зіставлення прогнозів узгоджує порядок класів за атрибутом `classes_` у фінальному елементі моделі чи конвеєра. На виході імовірності нормуються та агрегуються в ансамблевий підсумок, що зменшує

варіативність окремих класифікаторів і підвищує стійкість системи до локальних збоїв.

Фінальна інтерпретація результатів уніфікована доменною логікою, тому обчислювальне ядро працює з імовірностями, а прикладний рівень відповідає за семантичну подачу. У мультикласовому модулі це забезпечує стислий висновок: якщо найімовірнішим є клас «Здорова легеня», виводиться нейтральне формулювання «Жодних змін структури легень не виявлено», в інших випадках зазначається відповідний патологічний клас. У бінарному модулі користувач отримує одну з двох узгоджених відповідей: «Є ознаки СЗЛ» або «Немає ознак СЗЛ».

4.3 Демонстрація функціоналу програмного застосунку

Основною метою було розробити інтуїтивно зрозумілий та не складний графічний інтерфейс для користувача, що не потребував би додаткового навчання аби успішно з ним працювати.

Під час запуску програмного застосунку першим відкривається головне вікно (рис. 4.1).

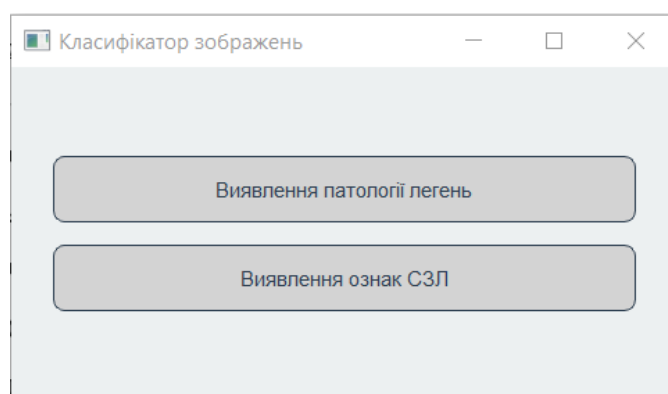


Рисунок 4.1 – Головне вікно ПЗ

Тут розміщені лише 2 кнопки: Виявлення патології легень (для мультикласифікації) та Виявлення ознак СЗЛ (для бінарної задачі).

Натискаючи на них, користувач переходить у відповідне вікно, що буде детальніше описано в наступних пунктах.

4.3.1 Мультикласова класифікація

Кнопка Виявлення патології легень, відкриває нове вікно під назвою «Класифікація зображень – мультиклас». Детальніше з його виглядом можна ознайомитись на рис. 4.2:

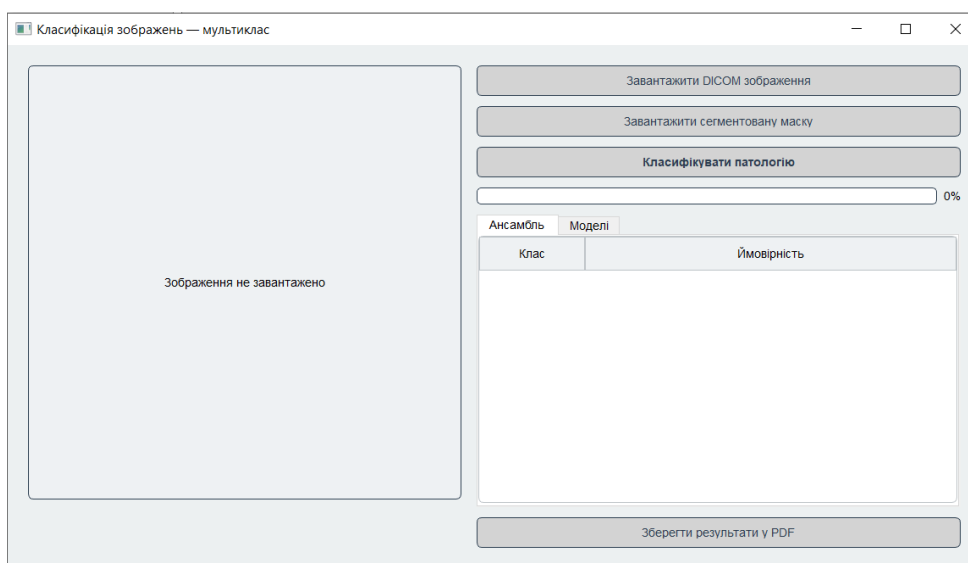


Рисунок 4.2 – Вікно для виконання мультикласифікації

Як можна побачити, візуально робоча область даного вікна розділена на два сегменти приблизно однакового розміру. Ліворуч розміщена область, де буде відображатися завантажений знімок. На рис. 4.2 вона відображена зараз як чорна тонка рамка із надписом «Зображення не завантажено». Відповідно, праворуч розташована область для завантаження КТ зрізів та роботи з ними. Тут користувач завантажує КТ зріз, його маску, бачить прогрес виконання класифікації на progress bar, переглядає результати ансамблю і моделей, а також зберігає звіт.

Для того, аби завантажити КТ зрізи, потрібно натиснути відповідну кнопку – Завантажити DICOM зображення. Після цього для користувача відкривається відповідна директорія (було задано для зручності) та

відображаються файли, що відповідають формату DICOM. Коли потрібний файл обрано і завантажено – обране зображення одразу починає відображатися у відповідній області, як це показано на рис. 4.3.



Рисунок 4.3 – Відображення завантаженого КТ зрізу у відповідній області

Після того користувач має завантажити сегментовану маску легень. Якщо ця маска не відповідатиме зображенню, яке було завантажено – з’являється повідомлення з попередженням (рис. 4.4).

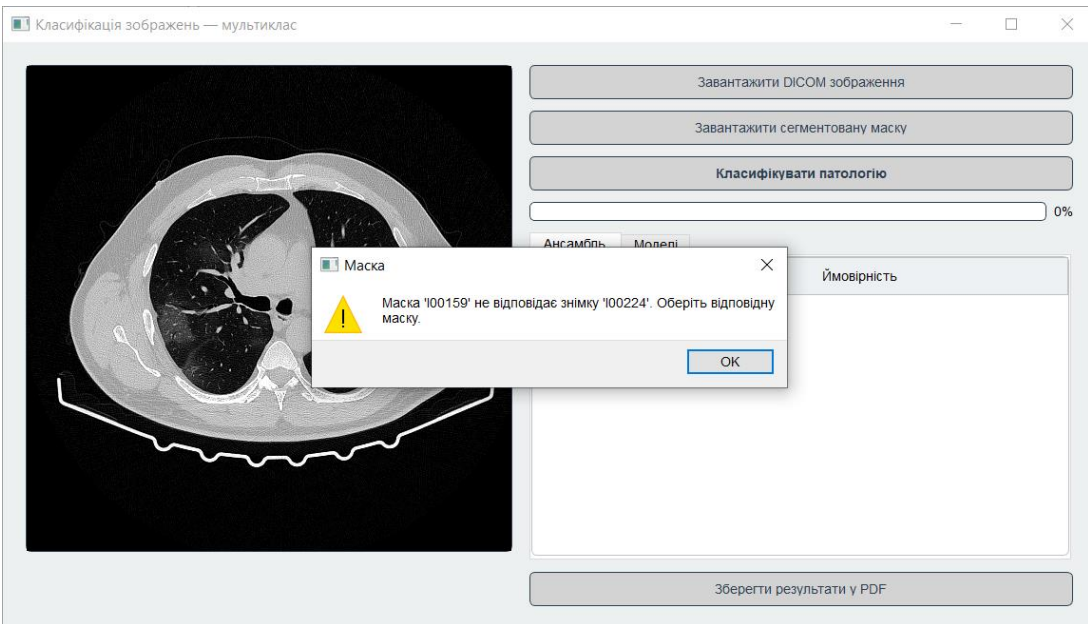


Рисунок 4.4 – Повідомлення про невідповідність маски до КТ зрізу

Після успішного завантаження зрізу та сегментованої маски можна переходити до класифікації – для цього варто натиснути на кнопку Класифікувати патологію. Як результат, модель видає свій кінцевий прогноз у випливаючому вікні (Message Box) (рис. 4.5), детальніше його можна переглянути у відповідній області.

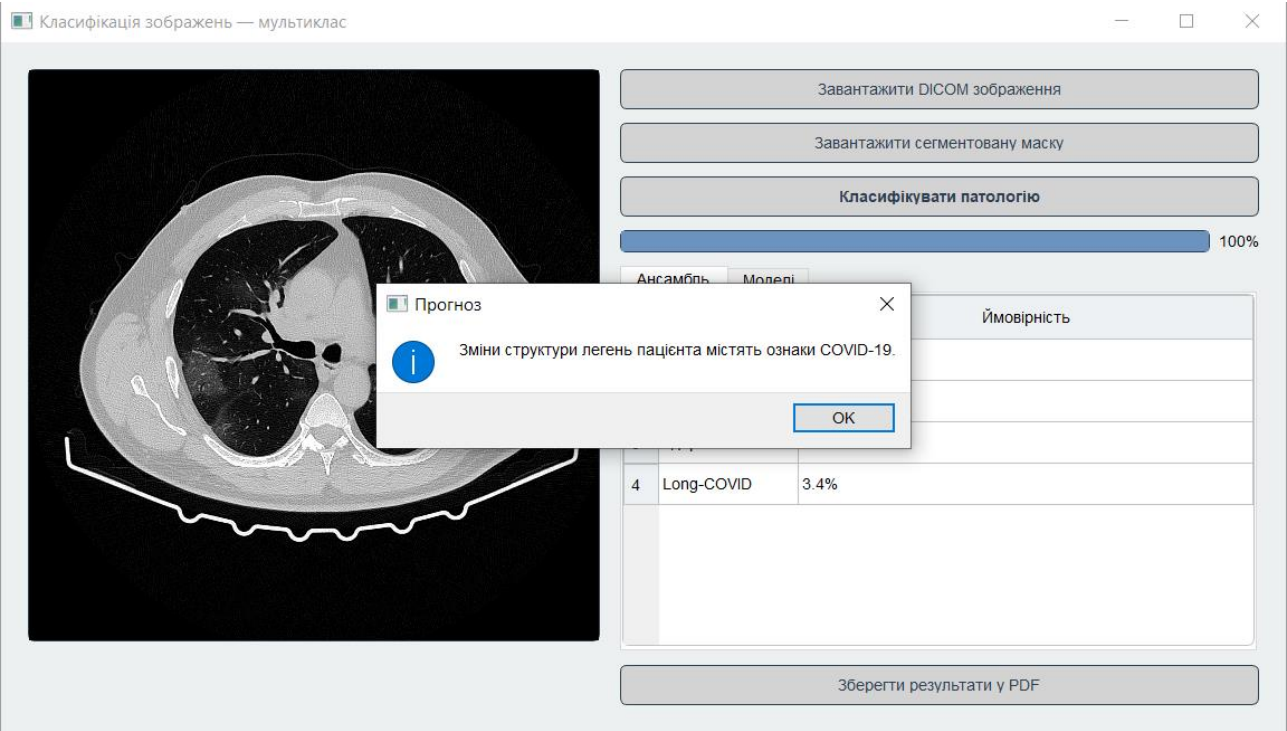


Рисунок 4.5 – Message Box з прогнозом багатокласової класифікації

Область для виведення результатів класифікатора поділена на 2 вкладки: Ансамбль (рис. 4.6) та Моделі (рис. 4.7).

Ансамбль		Моделі
	Клас	Ймовірність
1	COVID-19	83.3%
2	Пневмонія	8.1%
3	Здорова легеня	5.1%
4	Long-COVID	3.4%

Рисунок 4.6 – Відображення результатів у вкладці Ансамбль

Ансамбль		Моделі		
	Модель	Прогноз	Ймовірність	Час (с)
1	KNN	COVID-19	1.0000	0.0067
2	LightGBM	COVID-19	0.8913	0.0066
3	LR	COVID-19	0.4162	0.0013
4	Random Forest	COVID-19	0.9933	0.0713
5	SVM	COVID-19	0.9034	0.0049
6	XGBoost	COVID-19	0.7964	0.0024

Рисунок 4.7 – Відображення результатів у вкладці Моделі

У вкладці Ансамбль користувач бачить «сукупне» рішення всіх шести моделей та з якою ймовірністю моделі прогнозують виявлену патологію. Наприклад, як видно на рис. 4.6 ансамбль моделей прогнозує, що у пацієнта наявні ознаки COVID-19 з ймовірністю 83,3%. Решта класів також розпізнаються (оскільки присутня кореляція між ознаками патологій), але з набагато меншою «впевненістю», тож можна дійсно зробити висновок, що у пацієнта саме COVID-19, а не якийсь інше захворювання.

Крім того, є можливість переглянути більш детальний результат кожної з моделей, для цього варто перейти у вкладку з відповідною назвою – Моделі. Тут відображається таблиця (рис. 4.7), у якій вказана назва моделі, прогноз який вона зробила (COVID-19, Здорова легеня, Пневмонія, long-COVID), ймовірність цього прогнозу з точністю до десятитисячних та час за який його було виконано (в секундах).

4.3.2 Визначення ознак СЗЛ

Якщо ж лікар точно знає, що у пацієнта COVID-19 і необхідно перевірити наявність синдрому зникаючої легені, то у початковому вікні йому варто натиснути на кнопку Виявлення ознак СЗЛ. Загалом вікно для бінарної класифікації не відрізняється за виглядом і функціоналом від вікна з мультикласифікацією. Різниця полягає лише у класах, що визначаються.

Завантаження КТ зрізу з сегментованою маскою відбувається аналогічним чином, як і у сценарії для задачі мультикласифікації. Тому пропустимо даний етап і перейдемо до вже виконаної класифікації (рис. 4.8):

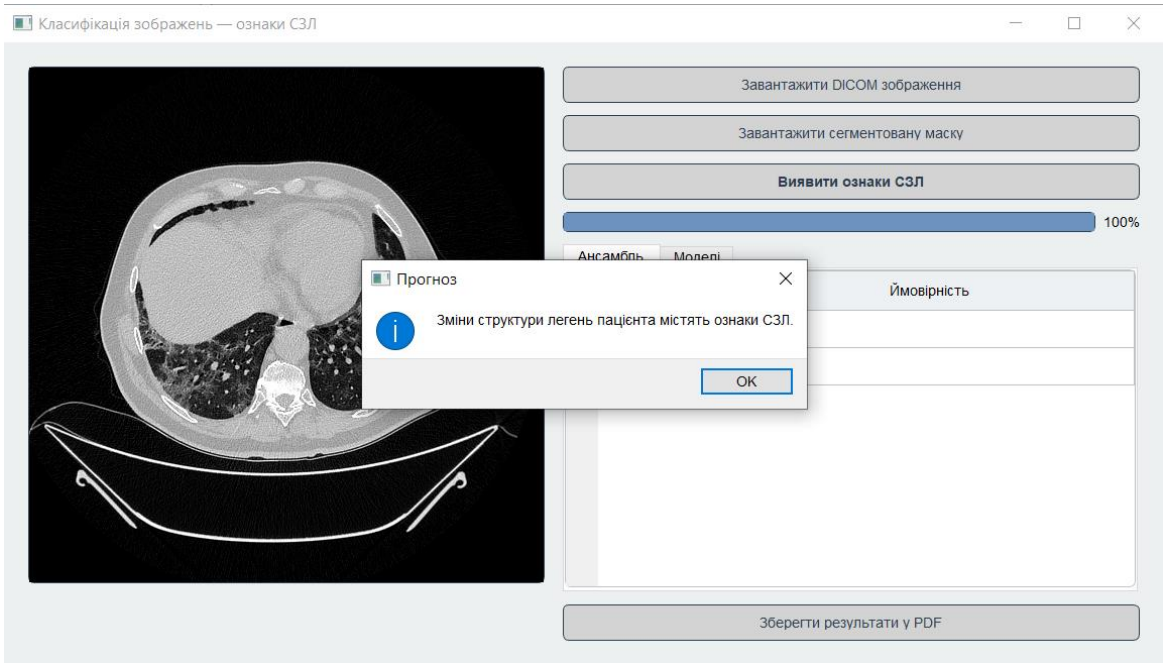


Рисунок 4.8 – Message Box з прогнозом бінарної класифікації

Вміст вкладки Ансамбль відображено на рис. 4.9. Згідно з отриманими результатами на завантаженому знімку є ознаки СЗЛ з імовірністю 78,4%.

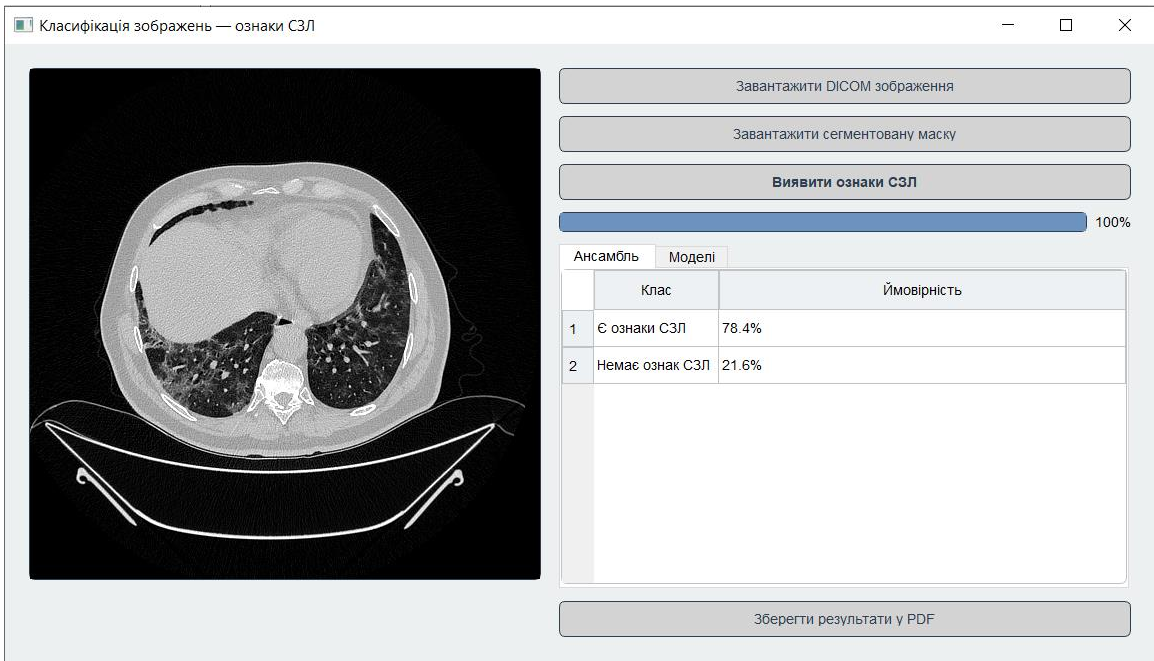


Рисунок 4.9 – Результат роботи класифікатора у вкладці Ансамбль

Відображення вмісту вкладки Моделі відображено на рис. 4.10. Як видно з наведеної таблиці, всі моделі, крім однієї прогнозують наявність ознак СЗЛ.

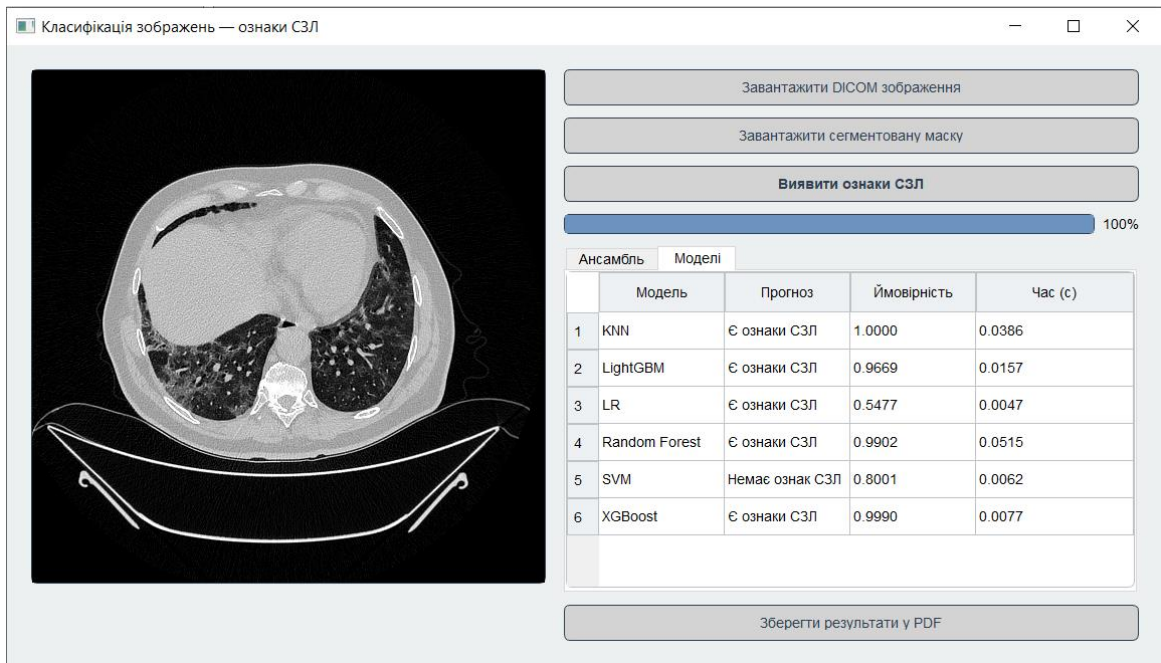


Рисунок 4.10 – Результат роботи класифікатора у вкладці Моделі

4.3.3 Генерація звіту

Кожне з вікон програмного продукту має кнопку «Зберегти результати у PDF». Вона розташована у нижній правій ділянці робочої області (рис. 4.11).

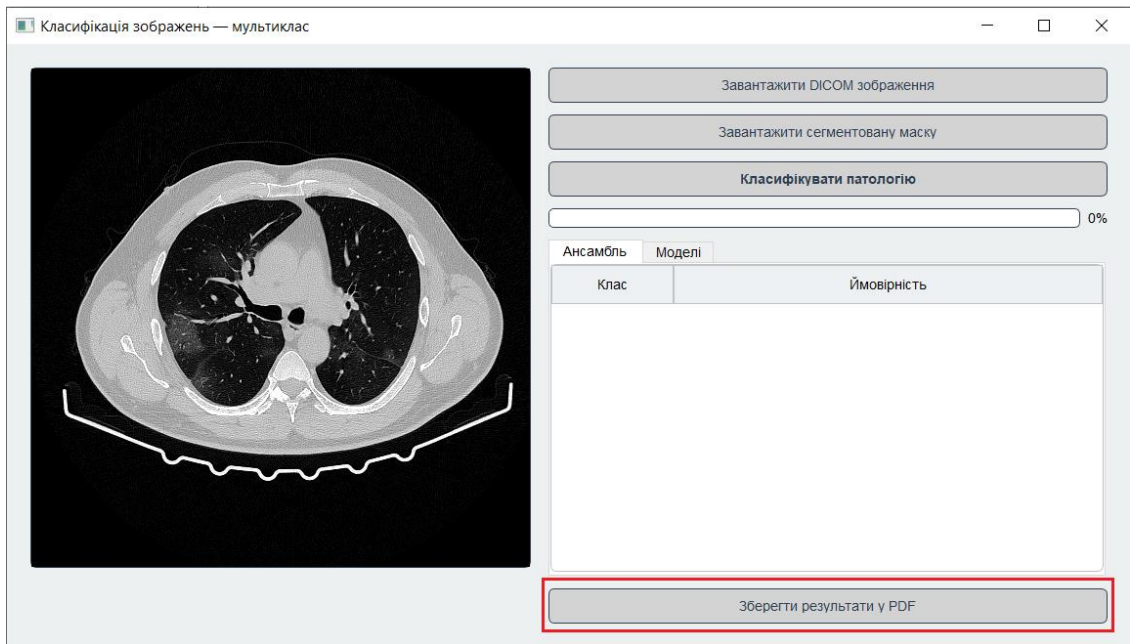


Рисунок 4.11 – Розміщення кнопки збереження звіту у вікні програми

Для того, аби його сформувати, користувач має лише натиснути на дану кнопку. Перед користувачем відкриється стандартне діалогове вікно для збереження файлу, де у рядку імені файлу автоматично вказуватиметься формат назви у вигляді «result_YYYYMMDD-HHMMSS.pdf» або «VLSresult_YYYYMMDD-HHMMSS.pdf» (рис. 4.12).

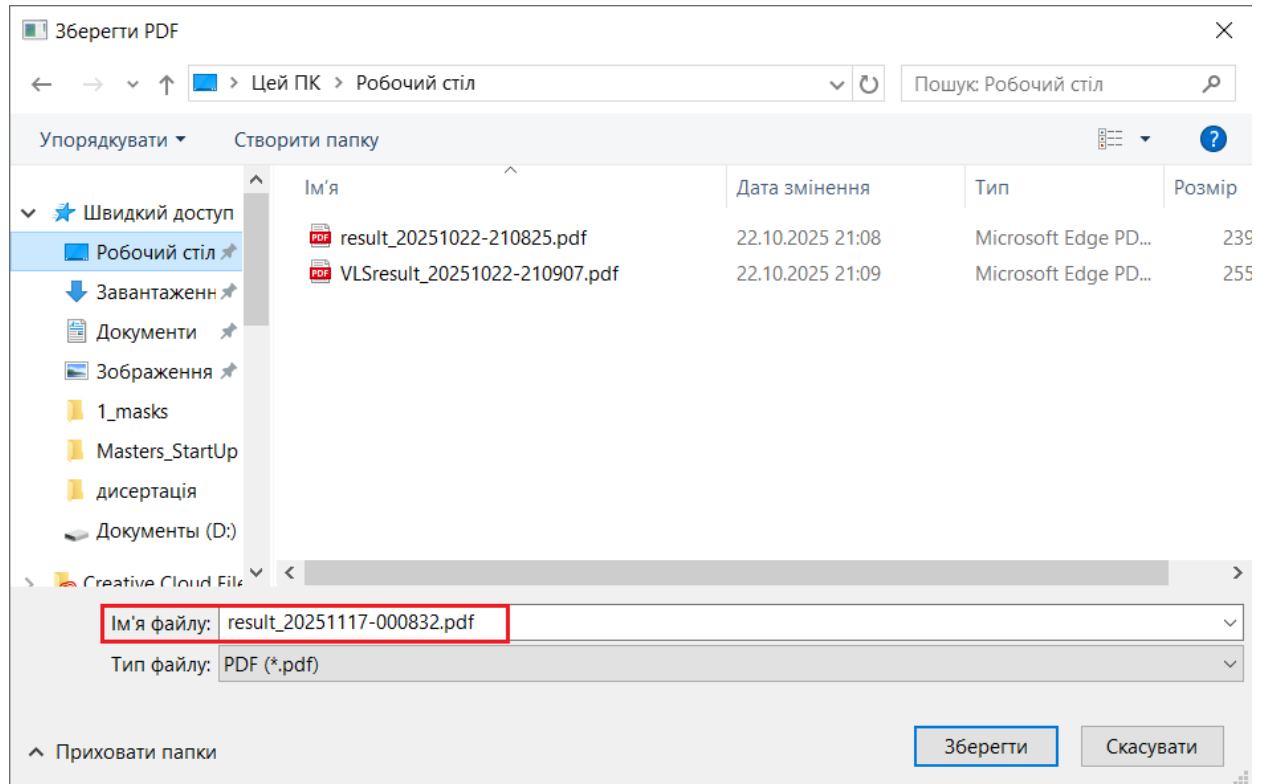


Рисунок 4.12 – Діалогове вікно з параметрами збереження звіту класифікації

Незалежно від типу задачі, звіт містить у собі дату проведення класифікації, зображення, на якому вона виконувалась, таблицю підсумкових ймовірностей ансамблю моделей, таблицю результату класифікації по кожній моделі окремо, а також прогноз класифікатора. Прогнозом є речення, що підсумовує до якої нозології відносяться патологічні зміни на знімку. Наприклад, «Зміни структури легень пацієнта містять ознаки COVID-19». Якщо ж пацієнт здоровий, то у прогноз записується фраза «Жодних змін структури легень не виявлено».

Приклад збереженого PDF-звіту можна переглянути на рис. 4.13.

Дата класифікації: 2025-10-15 17:21:12



Підсумкові ймовірності (ансамбль):

Клас	Ймовірність
COVID-19	83.3%
Пневмонія	8.1%
Здорова легеня	5.1%
Long-COVID	3.4%

Результати класифікації (моделі):

Модель	Прогноз	Ймовірність	Час (с)
KNN	COVID-19	1.0000	0.0067
LightGBM	COVID-19	0.8913	0.0066
LR	COVID-19	0.4162	0.0013
Random Forest	COVID-19	0.9933	0.0713
SVM	COVID-19	0.9034	0.0049
XGBoost	COVID-19	0.7964	0.0024

Прогноз класифікатора:

Зміни структури легень пацієнта містять ознаки COVID-19.

Рисунок 4.13 – Приклад збереженого PDF звіту

Висновок до розділу 4

Під час практичної частини роботи було розроблено модель бізнес-процесів, яка охоплює повний цикл локальної диференціальної класифікації КТ-зрізів – від імпорту даних до формування структурованого PDF-звіту. Це включає сегментацію легень, виділення інтерпретованих текстурних ознак і їх гармонізацію для зменшення міжсайтової варіабельності.

Як результат, було розроблено двомодульний застосунок, що містить мультикласову гілку та окремий бінарний модуль скринінгу ознак СЗЛ. Обидва вони спільно використовують контроль узгодженості пари «зображення – маска», ансамблеве зведення ймовірностей і генерацію звітів із усією інформацією стосовно прогнозу моделі. Крім того, графічний інтерфейс було створено максимально простим та інтуїтивно зрозумілим, щоб будь-який користувач міг вдало його опанувати.

РОЗДІЛ 5

ОБГОВОРЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

У цьому розділі узагальнюються результати магістерського дослідження та оцінюється ступінь виконання поставлених завдань. Окрему увагу приділено інтерпретації експериментальних результатів роботи моделей машинного навчання для мультикласифікації та бінарного виявлення ознак ураження, з акцентом на їх діагностичну спроможність.

5.1 Виконання вимог технічного завдання

У ході виконання магістерської роботи реалізовано всі основні вимоги технічного завдання як на рівні алгоритмічного конвеєра, так і на рівні прикладного програмного забезпечення. Згідно із завданням, система мала забезпечувати диференціальну діагностику захворювань легень за даними КТ із використанням інтерпретованих радіомічних ознак та класичних моделей машинного навчання. Для цього побудовано повний методологічний конвеєр обробки даних: від завантаження знімків та масок легенів до формування підсумкового рішення і структурованого звіту.

У рамках цього конвеєра реалізовано виокремлення масок легеневої паренхіми, розрахунок набору текстурних ознак методами текстурного аналізу GLCM, GLDS, GLRLM та LBP, і модуль гармонізації міжсайтової варіабельності, що узгоджується з теоретичними засадами, викладеними у попередніх розділах. Усі ознаки зберігаються у табличному форматі (CSV) з прив'язкою до ідентифікаторів пацієнтів та зрізів, що відповідає вимогам до подальшого аналізу та забезпечує відтворюваність експериментів.

Щодо формування навчальної, валідаційної та тестової підвбірок, то воно відбувається на рівні пацієнта, аби унеможливити ситуацію де один і той самий пацієнт ніколи не потрапляє одночасно в різні підвбірки. Це відповідає

вимогам технічного завдання щодо коректної експериментальної методології та запобігання витоку даних.

Функціональні вимоги до модулів класифікації також було виконано. Система реалізує мультикласову класифікацію КТ-зрізів, а також є окремий модуль для бінарної оцінки наявності синдрому зникаючої легені в підмножині COVID-19-пацієнтів. Для обох цих задач було використано узгоджений набір моделей машинного навчання (мультиноміальна логістична регресія, Random Forest, SVM, KNN, XGBoost, LightGBM), налаштування й оцінювання яких виконувалися з урахуванням сильного дисбалансу класів. Якість моделей було перевірено за допомогою крос-валідації та підсумкової оцінки на відкладеному тестовому наборі, що відповідає вимогам технічного завдання щодо кількісної валідації.

Окремий блок вимог стосувався програмної реалізації системи, відповідно було розроблено десктопний застосунок, який працює локально на робочій станції без потреби у підключенні до мережі. Він дозволяє завантажувати DICOM-знімки легень разом з їх сегментованими масками, запускати обраний модуль інференсу та переглядати результати як ансамблю моделей, так і кожної з них окремо.

Також, реалізовано формування PDF-звіту українською мовою з фіксацією всіх прогнозів моделей, що дещо спрощує подальше документування результатів. Варто зазначити, що обробка даних відбувається локально, ідентифікатори пацієнтів використовуються у псевдонімізованому вигляді, що відповідає вимогам до конфіденційності та безпечної роботи з медичною інформацією.

Таким чином, усе вищеописане в сукупності забезпечує належне виконання ключових вимог технічного завдання відповідно до теми моєї роботи як у частині функціоналу, так і в частині організації експериментів та роботи з даними.

5.2 Аналіз отриманих результатів моделей

5.2.1 Мультикласифікація

Мультикласова задача класифікації виконувалась як на гармонізованому наборі ознак, так і на ознаках з оригінальних DICOM зрізів, де було застосовано лише базову обробку. Результати тестової підвибірki можна побачити у табл. 5.1 та у табл. 5.2. Усі результати подано на тестових підвибірках.

Таблиця 5.1. Результати мультикласифікації після гармонізації (адаптований RELIEF)

Модель	Метрики класифікації					
	Accuracy	Recall	Precision	Specificity	F1-weighted	F1-macro
Мультиноміальна логістична регресія	0.6040	0.4648	0.4732	0.8508	0.6131	0.4659
Випадковий ліс	0.5151	0.3371	0.3566	0.7992	0.5045	0.3427
SVM	0.5740	0.3960	0.4325	0.8253	0.5654	0.4024
KNN	0.5234	0.3142	0.3459	0.7968	0.4883	0.3121
XGBoost	0.5958	0.3472	0.4115	0.8102	0.5319	0.3377
LightGBM	0.5912	0.3398	0.4401	0.8034	0.5224	0.3289

Таблиця 5.2. Результати мультикласифікації на даних без гармонізації

Модель	Метрики класифікації					
	Accuracy	Recall	Precision	Specificity	F1-weighted	F1-macro
Мультиноміальна логістична регресія	0.6349	0.5242	0.4990	0.8642	0.6440	0.5065
Випадковий ліс	0.4844	0.3300	0.3357	0.7911	0.4756	0.3288
SVM	0.5653	0.3916	0.4043	0.8182	0.5443	0.3889
KNN	0.5054	0.3068	0.3122	0.7914	0.4760	0.3013
XGBoost	0.5960	0.3467	0.4088	0.8101	0.5319	0.3368
LightGBM	0.5853	0.3383	0.4211	0.8020	0.5183	0.3284

Як видно з результатів, адаптований і дещо спрощений метод гармонізації RELIEF по-різному вплинув на моделі. Для чотирьох із шести досліджених нелінійних моделей, а саме: Random Forest, SVM, KNN та LGBM

– спостерігалось легке покращення інтегральних метрик порівняно з «сирими» даними, що є показником успішного, хоч і дещо недостатнього, очищення сигналу.

Оскільки зростали всі ключові показники, то можна зробити висновок, що розподіли ознак все ж було вирівняно. Зокрема, F1-weighted значно зросла, наприклад, для Випадкового лісу з 0.4756 до 0.5045 (одночасно з цим Accuracy зростає з 0.4844 до 0.5151), а для SVM – з 0.5443 до 0.5654 (див. табл. 5.1 та табл. 5.2 відповідно). Як підтвердження цієї позитивної динаміки змін є помітне зростання Precision та помірне для Recall. Ці результати демонструють, що навіть в умовах невеликої вибірки, гармонізація все ж трохи вирівняла дисперсії та розподіли ознак, що є критичним для моделей, які базуються на відстані та ансамблевих методах, дозволяючи їм точніше знаходити біологічно значущі розділяючі межі. І хоча F1-масго залишась низькою через проблемний клас PNEUMONIA (бактеріальна пневмонія), ця міра також показувала загальну позитивну тенденцію.

Щодо моделі Мультиноміальної логістичної регресії, то вона продемонструвала зниження F1-weighted з 0.6440 на оригінальних ознаках до 0.6131 на гармонізованих. Таке падіння можа назвати очікуваним, оскільки на сирих даних логістична регресія, як лінійна модель, імовірно, використовувала лінійні зміщення, спричинені сканерами, як сильні, але хибні предиктори діагнозу, що підвищувало показник F1-weighted за рахунок зниження узагальнювальної здатності. Натомість, після застосування адаптованої версії RELIEF модель була змушена спиратися переважно на біологічні відмінності, що могло дещо погіршити результати за використання тих самих налаштувань. Тож можна припустити, що цей результат також свідчить про хоч і часткове, проте усунення лінійних змішувальних факторів. У той самий час, модель XGBoost залишалася стабільною, підтверджуючи свою високу стійкість до варіацій у вхідних даних.

Найгострішою проблемою в задачі мультикласифікації лишається ідентифікація класу PNEUMONIA (бактеріальна пневмонія) з показником

support N=256 на тесті. Більшість моделей (LR, SVM, XGB, LGBM) демонстрували нульові показники Recall та F1-score як на гармонізованих, так і на сирих даних. Таким чином, жоден з випадків PNEUMONIA не був коректно класифікований, що знову ж таки вказує на критично низьку представленість класу, що значно ускладнює його диференцію.

5.2.2 Бінарна класифікація

Аналогічний експеримент було проведено і для задачі бінарної класифікації. Результати за основними метриками на тестових підвбірках відображено в табл. 5.3 та табл. 5.4.

Таблиця 5.3. Результати бінарної класифікації після гармонізації (адаптований RELIEF)

Модель	Метрики класифікації					
	Accuracy	Recall	Precision	Specificity	F1-weighted	F1-macro
Лінійна логістична регресія	0.6686	0.5949	0.6018	0.7859	0.6628	0.5973
Випадковий ліс	0.7573	0.6592	0.7252	0.9134	0.7369	0.6726
SVM	0.8138	0.7449	0.7968	0.9233	0.8049	0.7621
KNN	0.7945	0.7650	0.7585	0.8414	0.7958	0.7616
XGBoost	0.7407	0.6800	0.6928	0.8372	0.7363	0.6852
LightGBM	0.7312	0.6791	0.6829	0.8141	0.7298	0.6809

Таблиця 5.4. Результати бінарної класифікації на даних без гармонізації

Модель	Метрики класифікації					
	Accuracy	Recall	Precision	Specificity	F1-weighted	F1-macro
Лінійна логістична регресія	0.7427	0.7078	0.7001	0.7981	0.7451	0.7034
Випадковий ліс	0.7841	0.6722	0.7975	0.9619	0.7564	0.6908
SVM	0.8477	0.7758	0.8569	0.9619	0.8331	0.8007
KNN	0.8010	0.7635	0.7664	0.8607	0.8005	0.7649
XGBoost	0.7384	0.6834	0.6908	0.8259	0.7358	0.6867
LightGBM	0.7358	0.6806	0.6876	0.8235	0.7333	0.6837

Аналіз порівняння метрик між гармонізованими (див. табл. 5.3) та «сирими» (див. табл. 5.4) даними виявив, що застосування гармонізації за допомогою адаптованого RELIEF призвело до загального зниження інтегральних метрик для бінарної класифікації. Зокрема, лінійна модель LR продемонструвала найбільш виражене падіння F1-weighted з 0.7451 без гармонізації до 0.6628 з нею. Це падіння є значним і вказує на те, що на сирих даних логістична регресія успішно підлаштовувалася під лінійні зміщення сканерів, використовуючи їх як сильний, але хибний предиктор, а їх видалення під час гармонізації «оголило» менш виражену лінійну біологічну кореляцію. Одночасно, нелінійні моделі, такі як SVM, RF та KNN, що мали найвищі показники, також продемонстрували падіння F1-weighted після гармонізації (наприклад, SVM з 0.8381 до 0.8049). Таке масове зниження всіх метрик може свідчити про те, що site-ефекти не були чистим шумом, а містили тонкі нелінійні патерни, які допомагали цим моделям досягати вищої роздільної здатності на відомій вибірці, а адаптована версія RELIEF видаляючи латентні сканер-специфічні варіації, могла ненавмисно видалити й ці корисні нелінійні патерни.

5.2.3 Пропозиції щодо покращення результатів

З огляду на проведений аналіз ефективності розробленої системи та виявлені методологічні обмеження, можна сформулювати низку рекомендацій, реалізація яких дозволить підвищити діагностичну точність та надійність програмного продукту в майбутньому.

Першочерговим напрямком є розширення та балансування навчальної вибірки. Критично мала кількість прикладів для класу «бактеріальна пневмонія» обмежує здатність моделей до узагальнення специфічних патернів цієї нозології. Залучення нових даних з інших медичних центрів дозволить стабілізувати навчання та підвищити чутливість системи до менш поширених патологій.

Окремим напрямком може бути подальше удосконалення підходів до гармонізації міжсайтової варіабельності ознак. Доцільно було б на наступних

етапах провести більш систематичний підбір параметрів адаптованого RELIEF (розмір околу, кількість сусідів, поріг відбору ознак тощо) із використанням структурованих валідаційних схем, а також розглянути альтернативні або додаткові методи гармонізації чи підходи доменної адаптації. Можливим є також комбінування параметричних моделей корекції сайт-ефекту з подальшим фільтраційним відбором ознак. Такий підхід дозволить гнучкіше керувати компромісом між зменшенням технічних зсувів та збереженням діагностично важливої інформації без зміни базової архітектури конвеєра.

Крім того, доцільним є впровадження інтелектуального відбору репрезентативних зрізів для кожного пацієнта. Таким чином, замість використання всіх доступних зрізів пропонується алгоритмічно відбирати фіксовану кількість найбільш інформативних зображень (наприклад, з найбільшою вираженістю патологічних змін). Це дозволить зменшити шум від неінформативних апікальних та базальних зрізів, збалансувати внесок кожного пацієнта у навчання моделі та сфокусувати алгоритми класифікації на найбільш значущих патологічних змінах.

Важливий внесок може дати перехід від по-зрізового (2D) аналізу до об'ємного (3D) аналізу текстурних ознак. Наразі реалізована система розглядає кожен зріз ізольовано, що спрощує обчислення, але ігнорує важливу просторову інформацію про структуру ураження вздовж вертикальної осі (Z-axis). Впровадження 3D-розрахунку матриць для, зокрема, GLCM та GLRLM може дозволити враховувати об'ємну форму та зв'язність патологічних ділянок, що є потенційно значущим фактором для диференціації уражень.

Також наразі дуже перспективним є напрямок інтеграції новітніх візійно-мовних моделей (Vision-Language Models). Замість використання лише візуальних ознак, VLM дозволяють поєднувати аналіз зображень із семантичною інформацією з текстових описів або медичних знань. Такий мультимодальний підхід може значно покращити якість класифікації, особливо для складних або рідкісних випадків, дозволяючи моделі не просто класифікувати зображення, а й формувати опис виявлених змін природною

мовою, що є наступним кроком у розвитку систем підтримки лікарських рішень [72].

5.3 Порівняння з аналогами та практична цінність

Наведені в розділі 1 програмні аналоги демонструють широкий спектр можливостей – від повністю автоматизованих AI-екосистем із глибоким навчанням та інтеграцією в PACS/RIS до універсальних середовищ кількісного аналізу й ручної розмітки на кшталт Dragonfly. Такі системи, як правило, орієнтовані на масштабні лікарні зі стійкою інфраструктурою, забезпечують 3D-сегментацію, heatmap-візуалізації, розширений перелік нозологій і триаж потоку досліджень, але водночас залишаються «важкими» у розгортанні, потребують ліцензування, стабільного мережевого з'єднання та адаптації під конкретний стек обладнання.

Більшість із них базуються на глибинних нейромережевих моделях, які працюють безпосередньо з 3D-об'ємами в шкалі HU та специфічних протоколах сканування, а в даній роботі, навпаки, використано радіомічні текстурні ознаки з легеневого вікна та класичні моделі машинного навчання. За таких відмінностей повноцінне кількісне порівняння діагностичних метрик є некоректним, тому акцент робиться насамперед на функціональному та концептуальному позиціонуванні запропонованої системи.

Запропонований у роботі MVP займає іншу, більш «легку» нішу. Він спирається на інтерпретовані текстурні ознаки (GLCM, GLRLM, GLDS, LBP) та класичні моделі машинного навчання, що дає змогу прозоро простежити шлях від зображення до прогнозу та контролювати кожен етап конвеєра. Концепція двоетапної роботи передбачає, що після мультикласової класифікації радіолог за потреби може окремо запустити модуль бінарної оцінки синдрому зникаючої легені для підмножини зрізів пацієнтів із COVID-19. У поточній реалізації цей скринінг не виконується автоматично для всіх COVID-випадків, а застосовується вибірково, за ініціативи користувача, що

відображає роль системи як допоміжного інструмента, а не автономного діагностичного рішення. Додатково, на відміну від розглянутих рішень, де питання міжсайтової варіабельності не акцентується як окремий етап обробки, у запропонованому MVP гармонізацію радіомічних ознак явно реалізовано як самостійний крок конвеєра, що зменшує вплив відмінностей між КТ-сканерами та протоколами.

З практичного погляду система орієнтована на використання в умовах обмежених ресурсів, передусім у пульмонологічних центрах, де може не бути доступу до дорогих ліцензійних AI-рішень чи хмарної інфраструктури. Програмний застосунок працює локально на робочій станції, приймає готові DICOM-серії та маски легеневих ROI й видає структурований PDF-звіт українською мовою з імовірнісним висновком та зазначенням використаних моделей з їх прогнозами. Такий формат знижує міжпостерігачеву варіабельність, полегшує документування випадків і сприяє включенню результатів автоматизованої класифікації в медичну документацію без додаткових проміжних кроків.

5.4 Обмеження і зовнішня валідність

Під час інтерпретації отриманих результатів необхідно враховувати низку факторів, що обмежують узагальненість висновків. Передусім йдеться про обсяг і структуру вихідних даних. Загальний розмір 97 пацієнтів є відносно невеликим для побудови узагальнюючих моделей машинного навчання. Однак критичним викликом є екстремальний дисбаланс класів, і зокрема той факт, що клас «бактеріальна пневмонія» представлений лише п'ятьма пацієнтами. Як вже неодноразово зазначалося, цей обсяг є недостатнім для будь-якої статистично значущої валідації. Тому всі результати та метрики, отримані для цього класу, слід розглядати виключно як попередні та дослідницькі, а не як доказ надійної роботи моделі. Для валідації здатності

системи диференціювати бактеріальну пневмонію потрібне суттєве розширення вибірки саме для цього класу.

Методологічні обмеження зумовлені вибраним конвеєром обробки даних. Сегментація легеневих масок виконується автоматично за допомогою нейромережевої моделі без подальшої ручної корекції, тому можливі хоч і не значні, але похибки сегментації (особливо при надто виражених консолідаціях або атипових змінах), які безпосередньо впливають на розрахунок текстурних ознак. Крім того, застосування класичних моделей машинного навчання забезпечує інтерпретованість і прозорий контроль над гіперпараметрами, однак водночас може поступатися глибинним архітектурам у здатності відтворювати складні просторові патерни.

Ще один важливий аспект стосується зовнішньої валідності. Наразі модель перевірено на певному, хоча й різномірному, але не вичерпному наборі сканерів, протоколів реконструкції та клінічних ситуацій. Додатково, оцінювання проводилося в ретроспективному режимі без проспективного тестування в реальному клінічному потоці, тому перенесення результатів на інші заклади, популяції пацієнтів або інші профілі захворюваності потребує окремого підтвердження.

Зважаючи на все, що було описано вище, результати даної магістерської роботи доцільно розглядати як демонстрацію життєздатності обраного радіомічного підходу та мінімально життєздатного продукту, що «відкритий» до подальшого розширення.

Висновок до розділу 5

У п'ятому розділі продемонстровано, що розроблена система в цілому виконує вимоги технічного завдання, адже було реалізовано повний конвеєр «КТ-знімок – маска – текстурні ознаки – гармонізація – інференс – звіт», два взаємодоповнювальні модулі класифікації (мультикласова задача та окремий модуль бінарної оцінки СЗЛ) і локальний десктопний застосунок із

формуванням структурованого PDF-звіту українською мовою. У задачі мультикласифікації система демонструє помірні, але відтворювані показники на відкладеному тесті в умовах вираженого дисбалансу, а для бінарної класифікації досягає вищих інтегральних метрик, що, у свою чергу, свідчить про практичну придатність запропонованого підходу для скринінгових сценаріїв і впровадження в клінічну практику.

Аналіз впливу гармонізації RELIEF-адаптації показав, що її ефективність залежить від постановки задачі. Для більш складної мультикласової диференціації гармонізація виконує роль «шумоглушника», послаблюючи лінійні та латентні site-ефекти й допомагаючи нелінійним моделям (Random Forest, SVM) краще вловлювати тонкі біологічні відмінності між класами. Натомість у грубому бінарному поділі наявності або ж відсутності ознак СЗЛ, припускається, що частина сканер-специфічних варіацій виявляється інтегрованою в емпірично корисний розділюючий патерн, і надмірне згладжування цих ефектів може знижувати ключові метрики оцінки. Це настановляє на думку про те, що рівень та форма гармонізації мають підбиратися з урахуванням гранулярності цільової задачі, а не як універсальний режим для всього.

Порівняння з програмними аналогами показало, що запропонований MVP займає проміжну нішу між важкими DL-платформами та інструментами ручного аналізу. Реалізована у роботі система спирається на інтерпретовані радіомічні ознаки, класичні моделі машинного навчання та явно виділений етап гармонізації міжсайтової варіабельності, орієнтуючись на локальне використання в умовах обмежених ресурсів. Водночас результати розділу окреслюють ключові обмеження – ними є обмежений та незбалансований датасет, чутливість до похибок сегментації та обмежена зовнішня валідність.

Усі це задає подальший вектор розвитку, що першочергово полягає у розширенні та балансуванні вибірки (особливо класу бактеріальної пневмонії), інтелектуальному відборі інформативних зрізів, тонкому налаштуванні або комбінуванні методів гармонізації, поступовому

впровадженні 3D-аналізу текстурних ознак та мультимодальних підходів, зокрема на основі візійно-мовних моделей (VLM).

РОЗДІЛ 6

СТАРТАП-ПРОЄКТ З ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ ЛЕГЕНЬ «LUNGSCAN»

У розділі подано узагальнений опис стартап-проєкту, створеного на основі результатів магістерської роботи. Зокрема, проаналізовано загальну концепцію програмного продукту «LungScan», цільові категорії користувачів, основні елементи бізнес-моделі, організаційні та ресурсні передумови його впровадження, а також ключові аспекти економічного обґрунтування, ринкового позиціонування та подальшого розвитку.

6.1 Короткий опис стартап-проєкту

Проєкт передбачає створення програмного застосунку «LungScan», що призначений для автоматизованої диференціальної діагностики захворювань легень на основі аналізу КТ-знімків пацієнтів. У межах цієї системи реалізується підхід радіоміки та використовуються класичні моделі машинного навчання, що дає змогу не лише виявляти наявність патологічних змін, а й виконувати їх попередню класифікацію з формуванням структурованого звіту. У такий спосіб «LungScan» забезпечує підтримку прийняття клінічних рішень, сприяє підвищенню відтворюваності діагностики та зменшує навантаження на медичних фахівців.

Відмінними якостями продукту є:

- можливість автоматично класифікувати типи станів легень (COVID-19, бактеріальна пневмонія, long-COVID, відсутність виражених патологічних змін) за допомогою багатокласових моделей машинного навчання;
- наявність окремого модуля для виявлення ознак синдрому зникаючої легені (СЗЛ) у пацієнтів з підтвердженим COVID-19;

- спрощений інтерфейс із можливістю завантаження зображень та отримання попереднього результату у вигляді структурованого тексту;
- локальне офлайн-функціонування, що не потребує передавання конфіденційних медичних зображень в онлайн-середовище;
- можливість подальшого розширення функціоналу, зокрема візуалізації зони ураження та формування розширеного структурованого медичного звіту у наступних версіях продукту.

Запропонована інформаційна система орієнтована на підтримку лікаря-радіолога у швидкому попередньому аналізі КТ-знімків для виявлення та диференціації патологій без додаткового збільшення навантаження на робочий процес. Вона також враховує інтереси медичних закладів щодо підвищення точності діагностики, зменшення ймовірності суб'єктивних помилок під час оцінювання зображень та інтеграції результатів у медичні інформаційні системи за умови дотримання вимог до безпеки обробки медичних даних.

6.2 Бізнес-модель та цільові споживачі проєкту

6.2.1 Сегменти споживачів та ключові користувачі

Основним сегментом споживачів системи «LungScan» є медичні установи, що здійснюють комп'ютерну томографію для діагностики захворювань легень: державні й приватні лікарні, діагностичні центри, а також медичні заклади, які впроваджують цифрові інструменти у клінічну практику. Безпосередніми користувачами продукту в межах цього сегмента є лікарі-рентгенологи та пульмонологи.

Для зазначених установ характерні надмірне навантаження на лікарів-радіологів, потреба в швидкому прийнятті рішень у складних або невизначених клінічних ситуаціях, а також наявність ризиків суб'єктивності під час інтерпретації КТ-знімків. У зв'язку з цим спільною рисою для цього

сегмента є орієнтація на впровадження точних автоматизованих інструментів, які зменшують імовірність діагностичних помилок, скорочують час аналізу досліджень та знижують навантаження на медичний персонал. З погляду моделей придбання програмних продуктів такі установи зазвичай використовують централізовані закупівлі (тендерні процедури), прямі договори з постачальниками програмного забезпечення або підписки на спеціалізовані програмні рішення.

Окремий, але важливий сегмент становлять науково-дослідні установи, зокрема ДУ «Інститут фтизіатрії і пульмонології ім. Ф. Г. Яновського», які можуть використовувати систему для валідації алгоритмів, аналізу складних клінічних випадків, формування рекомендацій та супроводу навчального процесу. Хоча цей сегмент не є масовим, він має ключове значення для формування довіри до продукту та його подальшого розвитку.

6.2.2 Ціннісні пропозиції для цільових сегментів

Цінність програмного продукту «LungScan» полягає в реалізації радіомічного підходу до диференціальної класифікації КТ-зрізів легень із використанням класичних моделей машинного навчання. Система поєднує автоматизоване виокремлення інтерпретованих текстурних ознак із масок легень, гармонізацію міжсайтової варіабельності ознак та двоетапний алгоритмічний конвеєр, який забезпечує мультикласову класифікацію станів та окреме виявлення ознак СЗЛ у пацієнтів з COVID-19.

У такій конфігурації продукт орієнтований на виконання низки ключових вимог клінічної практики, а саме:

- забезпечення лікарю-радіологу можливості швидкого попереднього аналізу КТ-знімків легень для виявлення патологій без додаткового збільшення навантаження на робочий процес;
- формування структурованого звіту, який може бути інтегрований у медичні інформаційні системи;

- локальне офлайн-функціонування продукту, що забезпечує безпечну обробку медичних зображень без потреби у передаванні конфіденційної інформації через мережу.

6.2.3 Канали збуту та взаємодія з клієнтами

Для «LungScan» передбачаються такі канали збуту, як:

- прямі продажі ліцензій та підписок медичним установам (державним і приватним клінікам, діагностичним центрам) із можливістю технічної підтримки та локального налаштування;
- партнерство з телемедичними платформами та постачальниками медичного програмного забезпечення для інтеграції «LungScan» у наявні клінічні IT-рішення;
- участь у державних та міжнародних програмах цифрової трансформації охорони здоров'я, що підтримують впровадження інструментів штучного інтелекту;
- використання власного веб-ресурсу з описом функціоналу, умов ліцензування, прикладів використання та формою для подання запитів щодо впровадження;
- співпраця з дистриб'юторами та системними інтеграторами, які вже працюють з медичними установами і можуть пропонувати «LungScan» як складову комплексних рішень.

На етапі виведення MVP на ринок залучення перших користувачів передбачається здійснювати шляхом демонстрацій функціоналу системи представникам медичних установ під час галузевих заходів, виставок і конференцій, а також у форматі прямих консультацій із потенційними користувачами. Додатковим інструментом можуть бути партнерські програми, у межах яких окремі заклади отримують доступ до базової версії системи на пільгових умовах в обмін на структурований зворотний зв'язок, клінічну оцінку результатів або участь у спільних дослідницьких проєктах.

Крім того, розглядається можливість представлення результатів роботи системи у наукових публікаціях, як частина професійної комунікації, де згадка про «LungScan» сприятиме залученню нових зацікавлених фахівців.

Вже після впровадження взаємодія з користувачами передбачає надання консультацій електронною поштою та через онлайн-форму на офіційному сайті. Додатково забезпечуватиметься підтримка у вигляді інтерактивної бази знань з відповідями на поширені запитання та організовується дистанційне навчання користувачів. У разі нагальної потреби можливий навіть індивідуальний супровід на етапі інтеграції у клінічні процеси з подальшим розширеним сервісним обслуговуванням за окремими домовленостями.

6.3 Організаційне та ресурсне забезпечення стартап-проєкту

6.3.1 Ключові види діяльності

Ключовим видом діяльності стартапу є розробка, тестування та впровадження інформаційної системи для мультикласової диференціальної діагностики захворювань легень на основі аналізу КТ-знімків.

На початковому етапі до основних видів діяльності належать:

- проєктування архітектури програмного забезпечення та алгоритмічного конвеєра;
- розробка та навчання моделей машинного навчання для виявлення патологічних змін та класифікації нозологій;
- створення базової версії користувацького інтерфейсу та інструментів візуалізації результатів;
- тестування роботи алгоритмів на реальних медичних зображеннях у співпраці з медичним консультантом;
- підготовка технічної документації та проведення внутрішньої валідації точності;
- захист прав інтелектуальної власності на розроблений алгоритм та програмний код;

- демонстрація MVP цільовій аудиторії та збір зворотного зв'язку для подальшої ітеративної оптимізації продукту.

6.3.2 Ключові ресурси

Ключові ресурси, що знадобляться під час реалізації та впровадження стартап-проєкту можна розділити на декілька категорій: матеріальні, інтелектуальні, людські та фінансові.

З матеріальних ресурсів для реалізації «LungScan» необхідні будуть:

- обчислювальне обладнання: персональні комп'ютери або робочі станції з достатнім обсягом оперативної пам'яті (рекомендовано від 16 ГБ) для виконання радіомічного конвеєра та швидкого інференсу моделей;
- програмне забезпечення: операційні системи, середовища розробки та спеціалізовані бібліотеки для обробки медичних зображень і реалізації алгоритмів машинного навчання.

Інтелектуальні ресурси є дуже важливою складовою і включають в себе такі аспекти, як:

- структурований набір КТ-зображень та верифікованих експертами масок сегментації, отриманий у рамках співпраці з ДУ «Інститут фтизіатрії і пульмонології ім. Ф. Г. Яновського»;
- адаптована Python-реалізація алгоритму RELIEF із вбудованим конвеєром робастної попередньої обробки;
- файли з конфігураціями моделей машинного навчання для мультикласової та бінарної класифікації, які є результатом проведених експериментів;
- код десктопного застосунку «LungScan», що реалізує логіку інтерфейсу, обробки даних та генерації звітів.

До людських ресурсів були включені:

- 1) Керівник проєкту та Full-Stack розробник – виконує функції стратегічного управління та поєднує ролі інженера з машинного навчання і програміста.
- 2) Науковий консультант – забезпечує методологічний супровід, допомагає у виборі алгоритмічних підходів та здійснює валідацію наукової новизни рішень.
- 3) Лікар-рентгенолог – надає доменну експертизу, бере участь у розмітці даних, формуванні діагностичних критеріїв та клінічній оцінці результатів роботи системи.

Для етапу виходу на ринок та масштабування передбачається залучення додаткових фахівців:

- 4) Менеджер з маркетингу та продажів – для формування маркетингової стратегії, комунікації з медичними закладами та просування продукту.
- 5) Фінансовий менеджер – для управління бюджетом, розробки детальної моделі монетизації та супроводу тендерних процедур.

Таким чином, поточна штатна чисельність команди становить 3 особи (з урахуванням консультантів), із плановим розширенням до 5 осіб на етапі комерційного запуску.

Щодо фінансових ресурсів, то на початковому етапі фінансування стартапу «LungScan» здійснюється за рахунок власних ресурсів учасників команди, і ці витрати охоплюють такі елементи, як: базова інфраструктура для розробки, обчислювальні ресурси та мінімальні витрати на підготовку MVP.

У подальшому планується залучення зовнішніх фінансових ресурсів через такі канали, як:

- платна підписка та продаж ліцензій після запуску повноцінного продукту;
- грантові програми з підтримки інновацій у сфері охорони здоров'я, цифрової медицини та штучного інтелекту (Horizon Europe, USAID, EU4DigitalUA) тощо;
- державні ініціативи в Україні, спрямовані на цифровізацію охорони здоров'я (пілотні впровадження в закладах, співпраця з МОЗ України, участь у тендерах тощо).

6.3.3 Ключові партнери

До ключових партнерів відносяться установи та заклади, які надають експертну, медичну та науково-методичну підтримку, а саме:

- ДУ «Національний інститут фтизіатрії і пульмонології ім. Ф.Г. Яновського» [75; 78; 79] - надає анонімізовані КТ-знімки для тестування системи, а також є потенційною базою для валідації та впровадження продукту в умовах реального клінічного середовища;
- партнерські медичні установи (державні й приватні клініки та діагностичні центри, що здійснюють комп'ютерну томографію);
- університетські наукові лабораторії та ІТ-спільноти, з якими буде проводитися технічний аудит, оптимізація обчислювальних моделей та обмін досвідом.

6.4 Економічне обґрунтування та фінансова модель

6.4.1 Структура витрат на розробку та впровадження MVP

Витрати на організацію та реалізацію MVP стартап-проєкту «LungScan» на період близько 9 місяців орієнтовно становлять:

1. Організація процесу програмної розробки, доступ до сховищ даних і обчислювальних потужностей – ~250 000 грн.

2. Формування команди стартапу: керівник проєкту, інженер з ML, програміст, науковий консультант, медичний консультант, фінансовий менеджер та фахівець з маркетингу – ~684 000 грн.
3. Збір та анонімізація медичних зображень (КТ-знімків), формування навчального набору – безкоштовно, за договором з ДУ «Інститут фтизіатрії і пульмонології ім. Ф.Г. Яновського».
4. Розробка алгоритмів аналізу, інтерфейсу користувача, системи виведення результатів – включено до заробітної плати працівників.
5. Створення лендінгу, візуальних матеріалів, базове маркетингове просування – ~40 000 грн.
6. Резерв – непередбачувані витрати, юридичне та бухгалтерське забезпечення – ~40 000 грн.

Таким чином, загальний орієнтовний бюджет на перші 9 місяців орієнтовно становить близько 1 014 000 грн.

6.4.2 Формування роздрібної ціни ліцензії

Витрати підраховуються для конкретного інвестора за наданим шаблоном [73]:

Реалізація IC «LungScan»:

- Повна собівартість реалізованої продукції стартапу – 16 000 грн (одна ліцензія) [73];
- Прибуток стартапу – $251\,000 - 16\,000 = 235\,000$ грн [73];
- Оптова ціна виробника – 251 000 грн [73].

Непрямі податки [73]:

- Акцизні збори – не застосовуються до ПЗ = 0 грн [73];
- Податок на додану вартість – $0,2 * 251\,000 = 50\,200$ грн [73];
- Оптова ціна відпускна – $251\,000 + 50\,200 = 301\,200$ грн [73].

Посередницька надбавка до ціни [73]:

- Витрати посередника (5%) – $0,05 * 301\,200 = 15\,060$ грн [73];
- Прибуток посередника (7%) – $0,07 * 301\,200 = 21\,084$ грн [73];
- Податок на додану вартість посередника – $0,2 * (15\,060 + 21\,084) = 7\,228,8$ грн [73];
- Оптова ціна закупки – $301\,200 + (15\,060 + 21\,084) + 7\,228,8 = 344\,572,8$ грн [73].

Торгівельна надбавка до ціни [73]:

- Витрати торгівельної організації (5%) – $0,05 * 344\,572,8 = 17\,228,64$ грн [73];
- Прибуток торгівельної організації (7%) – $0,07 * 344\,572,8 = 24\,120,1$ грн [73];
- Податок на додану вартість торгівельної організації – $0,2 * (17\,228,64 + 24\,120,1) = 8\,269,75$ грн [73];
- Роздрібна ціна реалізації – $344\,572,8 + (17\,228,64 + 24\,120,1) + 8\,269,75 = 394\,191,29$ грн [73].

6.4.3 Оцінка окупності та фінансових перспектив

Дохід від впровадження системи диференціальної діагностики «LungScan» передбачається формувати за рахунок реалізації ліцензій на використання програмного продукту для медичних установ різних рівнів. Розглядається кілька моделей монетизації, зокрема продаж постійної ліцензії, підписка на оновлення (регулярні функціональні та безпекові оновлення системи з технічною підтримкою), разова оплата за впровадження, що охоплює налаштування, інтеграцію системи в локальну медичну інформаційну інфраструктуру та навчання персоналу, а також річні підписки за рівнем функціоналу (базова, стандартна та розширена конфігурації). Додатковим джерелом фінансування можуть виступати участь у державних програмах та

грантових ініціативах, спрямованих на підтримку впровадження інструментів штучного інтелекту в охороні здоров'я.

З урахуванням орієнтовного бюджету на розробку MVP у розмірі близько 1 014 000 грн та планової структури ціни, за якої прибуток стартапу від реалізації однієї ліцензії становить близько 235 000 грн, точка безбитковості досягається після продажу приблизно п'яти ліцензій. За консервативного сценарію реалізації однієї ліцензії на квартал строк виходу на безбитковість становитиме близько 15 місяців, тоді як за інтенсивнішого сценарію (одна ліцензія на місяць) цей строк може скоротитися до приблизно п'яти місяців. Починаючи з п'ятої угоди, сумарний прибуток стартапу перевищує стартові витрати і орієнтовно становить близько 161 000 грн, що створює підґрунтя для подальшого масштабування проєкту та розширення функціоналу системи.

6.5 Аналіз ринку та конкурентного середовища

До основних прямих та непрямих конкурентів можна віднести рішення таких компаній, як Thirona (LungQ), Siemens Healthineers (AI-Rad Companion) та Aidoc, які було детально розглянуто в розділі 1. Ці продукти є технологічно зрілими, проте вони здебільшого орієнтовані на високобюджетний сегмент ринку та вимагають інтеграції в хмарну інфраструктуру. Крім того, існують дослідницькі платформи на кшталт Dragonfly, які надають потужний інструментарій для аналізу, але є складними для рутинного використання лікарем.

Реалізована система диференціальної діагностики захворювань легень «LungScan» орієнтована на сегмент медичних установ, які використовують комп'ютерну томографію для діагностики захворювань легень і зацікавлені в автоматизованих інструментах підтримки диференціальної класифікації зображень. До цього сегмента належать державні та приватні лікарні, діагностичні центри, а також спеціалізовані пульмонологічні клініки. Окрему

нішу становлять науково-дослідні установи та університетські клініки, які використовують подібні системи не лише в рутинній практиці, а й для проведення досліджень, формування рекомендацій та навчання фахівців.

Аналіз ринку програмних рішень для аналізу КТ легень показав, що домінуючу позицію на ньому займають продукти великих міжнародних вендорів, інтегровані з PACS/MIS та орієнтовані переважно на глибинні нейронні мережі, які працюють безпосередньо з томографічними зображеннями. Значна частина цих систем сфокусована на виявленні окремих нозологій (наприклад, COVID-19 чи пневмонія), оцінюванні ступеня ураження легень або автоматизованому формуванні кількісних карт ураження. Для більшості таких продуктів характерні хмарна модель розгортання, орієнтація на великі медичні центри та потреба у розвиненій IT-інфраструктурі, що обмежує їх використання в закладах із обмеженими ресурсами та посилює вимоги до забезпечення захисту даних.

На цьому фоні «LungScan» позиціонується як локальне десктопне рішення радіомічного типу, яке реалізує класичні моделі машинного навчання на основі інтерпретованих текстурних ознак, гармонізованих між різними КТ-сканерами та протоколами. Система орієнтована на мультикласову класифікацію станів (COVID-19, бактеріальна пневмонія, постковідні зміни типу long-COVID, відсутність виражених патологічних змін) та окреме виявлення ознак синдрому зникаючої легені у пацієнтів з COVID-19 у межах єдиного інтерфейсу, що відрізняє її від рішень, зосереджених переважно на бінарних сценаріях. Локальне офлайн-функціонування, можливість адаптації до різноманітного парку КТ-обладнання та орієнтація на структурований текстовий звіт роблять систему потенційно придатною для впровадження в медичних закладах із різним рівнем ресурсного забезпечення. Водночас для виходу на більш конкурентний рівень необхідними є подальше розширення функціоналу та накопичення більших масивів даних для зовнішньої валідації.

6.6 Перспективи розвитку та масштабування стартап-проєкту

Перспективи розвитку стартап-проєкту «LungScan» пов'язані насамперед із поглибленням технічного функціоналу інформаційної системи. Подальші роботи можуть бути спрямовані на розширення алгоритмічного конвеєра за рахунок підтримки аналізу серій КТ-знімків, що дасть змогу враховувати просторовий розподіл патологічних змін у легенях та формувати більш узгоджені висновки щодо тяжкості ураження. Доцільним є також впровадження модулів сегментації зон ураження всередині легеневої паренхіми з подальшим обчисленням кількісних індикаторів (частка ураженого об'єму, розподіл патернів тощо) та розробка засобів візуального пояснення результатів класифікації у вигляді теплових карт у межах маски легень. Окремим напрямом є вдосконалення методів гармонізації міжсайтової варіабельності та уточнення схем навчання моделей на розширених вибірках для підвищення стійкості до відмінностей між сканерами і протоколами обстеження.

З програмно-інженерної точки зору перспективним є подальший розвиток десктопного застосунку шляхом поглиблення інтеграції з медичними інформаційними системами та інфраструктурою обробки зображень. Це охоплює реалізацію механізмів імпорту досліджень безпосередньо з PACS, експорту структурованих звітів форматах, сумісних з МІС, удосконалення інтерфейсу користувача для різних сценаріїв клінічного застосування, а також розширення інструментів журналювання та аудиту дій для забезпечення простежуваності використання системи. Важливими є подальше зміцнення заходів інформаційної безпеки, деталізація технічної та експлуатаційної документації, а також підготовка методичних матеріалів для навчання користувачів.

Щодо клініцистики, то подальший розвиток «LungScan» передбачає розширення співпраці з медичними закладами та дослідницькими центрами для проведення зовнішньої валідації на незалежних вибірках, включно з

багатофакторним аналізом роботи системи в умовах реальної клінічної практики. За результатами таких досліджень може бути уточнено налаштування моделей, порогові значення для прийняття рішень та формат представлення результатів для різних категорій користувачів. У довгостроковій перспективі, за наявності відповідних ресурсів, може розглядатися адаптація системи до інших нозологій легеневих захворювань і поступовий рух у бік відповідності вимогам регуляторних органів для медичних виробів програмного забезпечення.

Висновок до розділу 6

У розділі 6 на основі результатів магістерської роботи сформовано концепцію стартап-проєкту «LungScan», яка поєднує технічні напрацювання з елементами бізнес-моделі та планом подальшої комерціалізації. Стисло описано призначення та функціонал MVP-застосунку, окреслено цільові категорії користувачів, сформульовано ключові переваги локального офлайн-рішення для медичних закладів із різним рівнем ресурсного забезпечення, а також узагальнено основні компоненти бізнес-моделі, включно з цільовими сегментами, ціннісними пропозиціями, каналами збуту та форматами взаємодії з клієнтами.

В описаних підрозділах було деталізовано організаційне та ресурсне забезпечення стартапу: визначено ключові види діяльності, необхідні матеріальні, інтелектуальні й людські ресурси, а також окреслено коло партнерів, залучення яких є критичним для розгортання системи «LungScan» у клінічній практиці. Побудовано початкову фінансову модель, що включає структуру витрат на розробку та впровадження MVP, підхід до формування роздрібної ціни ліцензії та оцінку точки беззбитковості, яка демонструє можливість окупності проєкту за продажу обмеженої кількості ліцензій і створює підґрунтя для подальшого масштабування. На основі аналізу ринку та конкурентного середовища «LungScan» позиціоновано як локальне десктопне

радіомічне рішення, що використовує гармонізовані текстурні ознаки та класичні моделі машинного навчання і доповнює існуючі високобюджетні AI-системи. Окреслені напрями розвитку – підтримка аналізу серій КТ-знімків, кількісна оцінка обсягу ураження, візуальне пояснення результатів та розширення співпраці з медичними й науковими установами – підтверджують практичну значущість отриманих у роботі результатів і їхній потенціал до трансформації в життєздатний програмний продукт.

ЗАГАЛЬНІ ВИСНОВКИ

Відповідно до поставленої мети та задач магістерської дисертації було сформульовано наступні висновки:

1. У результаті аналізу літератури та наявних програмних аналогів показано, що для підтримки діагностики захворювань легень переважно застосовуються DL-рішення, інтегровані у великі програмні платформи (зокрема в зв'язці з PACS/RIS) з фокусом на автоматизовану оцінку ураження та формування звітів. При цьому підходи, що ґрунтуються на явному виокремленні інтерпретованих текстурних ознак та класичних статистичних/ML-методах, як правило, або не деталізуються, або взагалі не висвітлюються. Саме виявлений дефіцит прозорих радіомічних конвеєрів із чітко описаними ознаками та методикою обробки даних обґрунтував вибір напрямку даної роботи та постановку подальших задач щодо текстурного аналізу КТ, гармонізації ознак і побудови власної системи підтримки рішень.

2. В рамках роботи розроблено й реалізовано відтворювану методику попередньої обробки КТ-зображень та виокремлення інтерпретованих текстурних ознак, яка формує послідовний конвеєр від вихідних КТ-зрізів через побудову масок до обчислення текстурних ознак. На першому етапі виконується сегментація легеневої паренхіми за допомогою U-Net-моделі R231CovidWeb, після чого до отриманих зображень послідовно застосовуються фіксоване легеневе вікно, квантування інтенсивностей до 64 рівнів сірого та робастна обробка масок. У результаті такої послідовності кроків сформовано робочий набір із 36 інтерпретованих текстурних ознак (GLCM, GLDS, GLRLM, LBP), які кількісно описують мікроструктуру легеневої тканини й можуть бути узгоджені з клінічним тлумаченням.

3. З метою зменшення впливу міжсайтової варіабельності даних у системі було програмно реалізовано адаптований метод RELIEF, який інтегровано з механізмами захисту від витоку даних. У межах цього підходу

гармонізаційні моделі навчаються виключно на тренувальній вибірці, після чого застосовуються до валідаційних і тестових піднаборів, що наближає процедуру оцінювання до квазі-зовнішньої валідації. Експериментальні результати показали, що в мультикласовій задачі така гармонізація ефективно приглушує міжсайтові ефекти, тоді як у бінарному сценарії виявлення синдрому зникаючої легені ступінь згладжування потребує обережного налаштування, аби не нівелювати корисні розділювальні патерни.

4. На основі сформованих ознак побудовано та валідовано низку моделей машинного навчання (логістична регресія, Random Forest, SVM, KNN, XGBoost, LightGBM) для двох взаємопов'язаних задач: мультикласової диференційної діагностики (COVID-19, бактеріальна пневмонія, long-COVID, відсутність виражених патологічних змін) та бінарного виявлення ознак синдрому зникаючої легені. На реальному незбалансованому датасеті з 97 пацієнтів (понад 26 тисяч зрізів) із відкладеним тестовим піднабором моделі демонструють помірні, але відтворювані результати для мультикласифікації (акурату порядку 0.55, f1-weighted близько 0.50) та вищі інтегральні метрики у бінарній задачі (accuracy ≈ 0.71 , f1-weighted ≈ 0.67).

Для мультикласової класифікації найкращий показник F1-weighted був досягнутий моделлю Мультиноміальної логістичної регресії (близько 0,61) хоча модель SVM була обрана як найкраща для узагальнювальної здатності (f1-weighted ≈ 0.51). У бінарній класифікації найкращі інтегральні метрики продемонструвала модель SVM на (акурату та f1-weighted ≈ 0.81). Отже, запропонований радіомічний підхід є практично придатним для скринінгового виявлення СЗЛ та попередньої диференціації станів легень за умови подальшого розширення й балансування вибірки.

5. На основі розроблених алгоритмічних компонентів створено локальну програмну систему «LungScan» як мінімально життєздатний продукт, призначений для автоматизованої підтримки диференціальної діагностики. Реалізовано повну послідовність обробки даних, що починається з імпорту КТ-знімків і завершується формуванням структурованого звіту

результатів у форматі PDF. Загалом, реалізована система підтримує два взаємодоповнювальні модулі, один з яких виконує мультикласифікацію станів легень, а інший відповідає за бінарне виявлення ознак СЗЛ, що забезпечує узгоджений сценарій роботи для різних клінічних запитів. Завдяки своїй локальній обробці медичних даних та вбудованим механізмам мінімізації витоку інформації програмна система орієнтована на інтеграцію в робочий процес лікаря-радіолога у якості «AI-as-assist». У сукупності ці характеристики свідчать про те, що мінімально життєздатний продукт «LungScan» підтверджує життєздатність запропонованого підходу та може слугувати основою для подальшого масштабування стартап-проєкту, розширення функціоналу і проведення мультицентрової зовнішньої валідації.

Магістерська дисертація виконана на реальних даних Національного інституту фтизіатрії і пульмонології ім. Ф.Г. Яновського [75; 78; 79], що були надані згідно договору про співпрацю від 5 січня 2023 року, і куди були впроваджені попередні результати актом впровадження від 26 грудня 2024 року. Отримані в даній роботі результати і програмний застосунок також впроваджені в роботу Інституту (акт впровадження від ...).

Також робота виконується в рамках ініціативної наукової роботи кафедри НДР д/р № 0123U100866 «Методи та моделі ідентифікації станів об'єктів в задачах прийняття медичних рішень» (06.02.2023р.) [80].

За результатами виконаної роботи була опублікована 1 наукова стаття у фаховому журналі категорії Б:

1. Лутченко В. Г. Ефективність алгоритмів машинного навчання для класифікації змін структури легень у постковідних та гострих стадіях COVID-19 / В. Г. Лутченко, В. О. Бабенко, Є. А. Настенко, М. І. Линник // Біомедична інженерія і технологія. – Київ, 2024. - №15/2024. – С. 27–35 (стаття). – DOI: <https://doi.org/10.20535/.2024.15.306763> (дата звернення 15.11.2025) [28].

Була опублікована 1 наукова стаття у фаховому виданні категорії А (індексована в Scopus):

1. Іє.А. Nastenکو, Diagnosis of COVID-19-Associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: A Review of Methods and Future Research Directions / Іє.А. Nastenکو, М.І. Linnik, М.О. Honcharuk, І.В. Davydovych, V.H. Lutchenko, V.O. Babenko // Innovative Biosystems & Bioengineering. – 2025. – Vol. 9, № 3. – Р. 16-27 (стаття). DOI: <https://doi.org/10.20535/ibb.2025.9.4.325335> (дата звернення 27.11.2025) [74].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Li Y., Xia L. Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management [Електронний ресурс] // AJR American Journal of Roentgenology. — 2020. — Vol. 214, № 6. — С. 1280–1286. — DOI: 10.2214/AJR.20.22954. — Режим доступу: <https://www.ajronline.org/doi/10.2214/AJR.20.22954> (дата звернення: 03.11.2025).
2. Carotti M., Salaffi F., Sarzi-Puttini P. та ін. Chest CT features of coronavirus disease 2019 (COVID-19) pneumonia: key points for radiologists [Електронний ресурс] // Radiologia Medica. — 2020. — Vol. 125, № 7. — С. 636–646. — DOI: 10.1007/s11547-020-01237-4. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7270744/> (дата звернення: 03.11.2025).
3. Yousaf M. N., Chan N. N., Janvier A. Vanishing Lung Syndrome: An Idiopathic Bullous Emphysema Mimicking Pneumothorax [Електронний ресурс] // Cureus. — 2020. — Vol. 12, № 8. — Art. e9596. — DOI: 10.7759/cureus.9596. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7478520> (дата звернення: 03.11.2025).
4. Da-Ano R., Visvikis D., Hatt M. Harmonization strategies for multicenter radiomics investigations [Електронний ресурс] // Physics in Medicine & Biology. — 2020. — Vol. 65, № 24. — Art. 24TR02. — DOI: 10.1088/1361-6560/aba798. — Режим доступу: https://www.researchgate.net/publication/343086341_Harmonization_strategies_for_multicenter_radiomics_investigations (дата звернення: 04.11.2025).
5. Sawant P., Sreemathy R. A Review on Texture Feature Analysis of Chest Computed Tomography Images for Detection and Classification of

- Pulmonary Diseases [Электронный ресурс] // Communication and Intelligent Systems. — 2023. — (Lecture Notes in Networks and Systems; vol. 686). — С. 463–475. — DOI: 10.1007/978-981-99-2100-3_36. — Режим доступа: https://www.academia.edu/107289511/A_Review_on_Texture_Feature_Analysis_of_Chest_Computed_Tomography_Images_for_Detection_and_Classification_of_Pulmonary_Diseases (дата звернення: 04.11.2025).
6. Li X., Li Q., Xie X. та ін. Integrating CT radiomics and clinical data with machine learning to predict fibrosis progression in coalworker pneumoconiosis [Электронный ресурс] // Frontiers in Medicine. — 2025. — Vol. 12. — Art. 1599739. — DOI: 10.3389/fmed.2025.1599739. — Режим доступа: <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1599739> (дата звернення: 04.11.2025).
 7. Kazerooni E. A. High-resolution CT of the lungs [Электронный ресурс] // AJR American Journal of Roentgenology. — 2001. — Vol. 177, № 3. — С. 501–519. — DOI: 10.2214/ajr.177.3.1770501. — Режим доступа: <https://ajronline.org/doi/10.2214/ajr.177.3.1770501> (дата звернення: 05.11.2025).
 8. Cozzi D., Cavigli E., Moroni C. та ін. Ground-glass opacity (GGO): a review of the differential diagnosis in the era of COVID-19 [Электронный ресурс] // Japanese Journal of Radiology. — 2021. — Vol. 39, № 8. — С. 721–732. — DOI: 10.1007/s11604-021-01120-w. — Режим доступа: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8071755> (дата звернення: 05.11.2025).
 9. Koo H. J., Lim S., Choe J. та ін. Radiographic and CT Features of Viral Pneumonia [Электронный ресурс] // RadioGraphics. — 2018. — Vol. 38, № 3. — С. 719–739. — DOI: 10.1148/rg.2018170048. — Режим доступа: <https://kstr.radiology.or.kr/upload/notice/20200227111127.1000.0.1.pdf> (дата звернення: 05.11.2025).

- 10.Chen M., Wang Y., Wang Q. та ін. Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation [Електронний ресурс] // npj Digital Medicine. — 2024. — Vol. 7. — Art. 349. — DOI: 10.1038/s41746-024-01328-w. — Режим доступу: <https://www.nature.com/articles/s41746-024-01328-w> (дата звернення: 05.11.2025).
- 11.AuntMinnie.com staff writers. FDA clears updated version of Thirona’s chest CT software [Електронний ресурс] // AuntMinnie. — 22.01.2024. — Режим доступу: <https://www.auntminnie.com/clinical-news/ct/article/15662470/fda-clears-updated-version-of-thironas-chest-ct-software> (дата звернення: 09.09.2025)
- 12.LungQ™ Clinical Suite – Thirona [Електронний ресурс] // Health AI Register. — 06.06.2024. — Режим доступу: <https://healthairegister.com/radiology/products/thirona-lungq> (дата звернення: 09.09.2025).
- 13.Thirona. LUMC calculates lung damages COVID-19 patients with AI technology [Електронний ресурс]. — 04.03.2021. — Режим доступу: <https://thirona.eu/news/lumc-calculates-lung-damages-covid-19-patients-with-ai-technology/> (дата звернення: 09.09.2025).
- 14.Siemens Healthineers. AI-Rad Companion Chest CT [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://www.siemens-healthineers.com/en-th/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion/chest-ct> (дата звернення: 09.09.2025).
- 15.Imbio. Lung Density Analysis™ – Inspiration [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://www.imbio.com/products/lung-density-analysis-inspiration/> (дата звернення: 10.09.2025).
- 16.Imbio. Imbio lung AI models support lung disease diagnosis and care via Nuance AI Marketplace [Електронний ресурс]. — 29.06.2020. — Режим доступу: <https://www.imbio.com/news/imbio-lung-ai-models-support-lung->

- [disease-diagnosis-and-care-via-nuance-ai-marketplace](#) (дата звернення: 10.09.2025).
- 17.Imbio. CT LDA+ Software Manual (US) [Електронний ресурс]. — 2025. — Режим доступу: https://4dmedical.com/wp-content/uploads/2025/09/Imbio_CT_LDAlplus_SW_Manual_US.pdf (дата звернення: 10.09.2025).
- 18.Imbio. Lung Texture Analysis [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://www.imbio.com/products/lung-texture-analysis> (дата звернення: 10.09.2025)
- 19.Aidoc. Implementation of an AI algorithm in clinical practice to reduce missed incidental pulmonary embolisms on chest CT [Електронний ресурс] // Aidoc: Clinical Studies. — [б. р.]. — Режим доступу: <https://www.aidoc.com/learn/clinical-studies/implementation-of-an-ai-algorithm> (дата звернення: 11.09.2025).
- 20.Aidoc. Aidoc to market an AI algorithm under the FDA’s enforcement policy for imaging systems [Електронний ресурс] // Aidoc. — [б. р.]. — Режим доступу: <https://www.aidoc.com/about/news/fda-aidoc-covid19/> (дата звернення: 11.09.2025).
- 21.contextflow. ADVANCE Chest CT [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://contextflow.com/solution/advance-chest-ct-2/> (дата звернення: 11.09.2025).
- 22.contextflow ADVANCE Chest CT [Електронний ресурс] // Sectra Amplifier Marketplace. — [б. р.]. — Режим доступу: <https://amplifiermarketplace.sectra.com/application/contextflow-advance-lung-ct/> (дата звернення: 11.09.2025).
- 23.contextflow. Customer segment: Radiologists [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://contextflow.com/customer-segment/radiologists> (дата звернення: 11.09.2025).
- 24.Comet. Dragonfly [Електронний ресурс]. — [б. р.]. — Режим доступу: <https://dragonfly.comet.tech> (дата звернення: 12.09.2025).

25. Object Research Systems (ORS). Dragonfly Help 2024.1: Measurements and Scalar Data for Multi-ROIs [Електронний ресурс]. — 2024. — Режим доступу: <https://www.theobjects.com/dragonfly/dfhelp/2024-1/Content/ROIs%20and%20Multi-ROIs/Multi-ROIs/Measurements%20and%20Scalar%20Data%20for%20Multi-ROIs/Measurements%20and%20Scalar%20Data%20for%20Multi-ROIs.htm> (дата звернення: 12.09.2025)
26. Cui Y., Li Y., Xing D. та ін. Feature Extraction [Електронний ресурс] // Bio-protocol. — 2021. — Режим доступу: <https://bio-protocol.org/exchange/minidetail?id=9310738&type=30> (дата звернення: 16.09.2025).
27. Haralick R. M., Shanmugam K., Dinstein I. Textural Features for Image Classification [Електронний ресурс] // IEEE Transactions on Systems, Man, and Cybernetics. — 1973. — Т. 3, № 6. — С. 610–621. — Режим доступу: https://haralick.org/book_chapters/TexturalFeatures.pdf (дата звернення: 16.09.2025).
28. Лутченко В. Г., Бабенко В. О., Настенко Є. А., Линник М. І. Ефективність алгоритмів машинного навчання для класифікації змін структури легень у постковідних та гострих стадіях COVID-19 [Електронний ресурс] // Біомедична інженерія і технологія (категорія Б). — 2024. — № 15/2024. — С. 27–35. — DOI: 10.20535/.2024.15.306763. — Режим доступу: <https://doi.org/10.20535/.2024.15.306763>; <http://biomedtech.kpi.ua/article/view/306763/303424> (дата звернення: 10.11.2025).
29. Nugroho H. A., Haryanto I., Wahyono S., Andriani P. Multipatch-GLCM for texture feature extraction on classification of the colon histopathology images using deep neural network with GPU acceleration. [Електронний ресурс] // ResearchGate. — 2020. — Режим доступу до ресурсу: https://www.researchgate.net/publication/340484256_Multipatch-GLCM_for_Texture_Feature_Extraction_on_Classification_of_the_Colon

- [Histopathology Images using Deep Neural Network with GPU Acceleration/](#) (дата звернення: 16.09.2025).
- 30.Nabizadeh N., Kubat M. Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features. [Текст] // Computers & Electrical Engineering. – 2015. – Т. 45. – С. 286-301.
 - 31.Лутченко В. Г. Додаток для класифікації постковідних ознак змін структури легень порівняно з гострою фазою коронавірусу : дипломна робота бакалавра : 122 Комп'ютерні науки [Електронний ресурс] / Лутченко Вікторія Геннадіївна ; НТУУ «КПІ ім. Ігоря Сікорського». — Київ, 2024. — 77 с. — Режим доступу: <https://ela.kpi.ua/handle/123456789/67667>; <https://ela.kpi.ua/bitstreams/8cff9f8a-d1c7-4966-90c2-065c208abe38/download> (дата звернення: 10.11.2025).
 - 32.Oprisan A., Oprisan S. A. Bounds for Haralick features in synthetic images with sinusoidal gradients [Електронний ресурс] // Frontiers in Signal Processing. — 2023. — Vol. 3. — DOI: 10.3389/frsip.2023.1271769. — Режим доступу: <https://www.frontiersin.org/journals/signal-processing/articles/10.3389/frsip.2023.1271769/full> (дата звернення: 06.11.2025).
 - 33.Giakoumoglou N. G. PyFeats: Open-source software for image feature extraction. [Електронний ресурс] // PyPI. – Режим доступу до ресурсу: <https://pypi.org/project/pyfeats/> (дата звернення: 16.09.2025)
 - 34.Niu L., Qian M., Yang W., Meng L., Xiao Y., Wong K. K. L., Abbott D., Liu X., Zheng H. Surface Roughness Detection of Arteries via Texture Analysis of Ultrasound Images for Early Diagnosis of Atherosclerosis [Електронний ресурс] // PLOS ONE. — 2013. — Vol. 8, № 10. — e76880. — DOI: 10.1371/journal.pone.0076880. — Режим доступу: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0076880> (дата звернення: 05.11.2025).

35. Tang X. Texture Information in Run-Length Matrices [Электронный ресурс] // IEEE Transactions on Image Processing. — 1998. — Т. 7, № 11. — С. 1602–1609. — Режим доступа: https://www.researchgate.net/publication/3326868_Texture_information_in_run-length_matrices (дата звернения: 18.09.2025).
36. Nastenka I., Maksymenko V., Galkin A., Pavlov V., Nosovets O., Dykan I., Tarasiuk B., Babenko V., Umanets V., Petrunina O., Klymenko D. Liver Pathological States Identification with Self-organization Models Based on Ultrasound Images Texture Features. [Электронный ресурс] // Conference paper. — 2020, 23 грудня. — С. 401-418. — Режим доступа до ресурсу: https://link.springer.com/chapter/10.1007/978-3-030-63270-0_26 (дата звернення: 10.11.2025).
37. PyRadiomics. Radiomic Features [Электронный ресурс] // PyRadiomics documentation. — [б. р.]. — Режим доступа: <https://pyradiomics.readthedocs.io/en/latest/features.html> (дата звернення: 09.11.2025).
38. Albregtsen F. Statistical Texture Measures Computed from Gray Level Run Length Matrices [Электронный ресурс]. — Oslo, 1995. — Режим доступа: <https://pdfcoffee.com/glrlm-5-pdf-free.html> (дата звернення: 20.09.2025).
39. Ojala T., Pietikäinen M., Mäenpää T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns [Электронный ресурс] // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2002. — Т. 24, № 7. — С. 971–987. — Режим доступа: https://www.cse.msu.edu/~rossarun/BiometricsTextBook/Papers/Face/Ojala_LBP_PAMI02.pdf (дата звернення: 18.09.2025).
40. Ihalapathirana A. Understanding the Local Binary Pattern (LBP): A Powerful Method for Texture Analysis in Computer Vision. [Электронный ресурс] // Medium. — Режим доступа до ресурсу: <https://bit.ly/3xgt1kA> (дата звернення: 18.09.2025).

41. Park J. Y. RELIEF: R package for removal of latent inter-scanner effects (RELIEF harmonization) [Электронный ресурс] // GitHub. — [б. р.]. — Режим доступа: <https://github.com/junjypark/RELIEF> (дата звернения: 10.11.2025).
42. Zhang R., Oliver L. D., Voineskos A. N., Park J. Y. RELIEF: a structured multivariate approach for removal of latent inter-scanner effects [Электронный ресурс] // Imaging Neuroscience. — 2023. — Vol. 1. — С. 1–16. — DOI: 10.1162/imag_a_00011. — Режим доступа: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10503485/> (дата звернения: 10.11.2025).
43. Hu F., Chen A. A., Horng H. та ін. Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization [Электронный ресурс] // NeuroImage. — 2023. — Vol. 274. — Art. 120125. — DOI: 10.1016/j.neuroimage.2023.120125. — Режим доступа: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10257347/> (дата звернения: 03.09.2025).
44. GeeksforGeeks. Multinomial Logistic Regression in R [Электронный ресурс] // GeeksforGeeks. — 23.07.2025. — Режим доступа: <https://www.geeksforgeeks.org/r-language/multinomial-logistic-regression-in-r/> (дата звернения: 15.11.2025).
45. Wikipedia. Multinomial logistic regression [Электронный ресурс]. — [б. р.]. — Режим доступа: https://en.wikipedia.org/wiki/Multinomial_logistic_regression (дата звернения: 20.09.2025).
46. Wikipedia. Random forest [Электронный ресурс]. — [б. р.]. — Режим доступа: https://en.wikipedia.org/wiki/Random_forest (дата звернения: 20.09.2025).

47. What is random forest? [Электронный ресурс] // IBM. – Режим доступа до ресурсу: <https://www.ibm.com/topics/random-forest> (дата звернення: 20.09.2025)
48. Gunay D. Random Forest [Электронный ресурс] // Medium. — 11.09.2023. — Режим доступа: <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e> (дата звернення: 07.11.2025).
49. Wikipedia. Support vector machine [Электронный ресурс]. — [б. р.]. — Режим доступа: https://en.wikipedia.org/wiki/Support_vector_machine (дата звернення: 20.09.2025).
50. Tech-AI-Math. Support Vector Machines (SVM) in Depth. Part 1: Functional Margin, Hinge Loss, Dual Problem, Lagrange Multipliers [Электронный ресурс] // Artificial Intelligence in Plain English (Medium). — 05.07.2023. — Режим доступа: <https://ai.plainenglish.io/support-vector-machines-svm-in-depth-part-1-882448c8310c> (дата звернення: 03.11.2025).
51. McGregor M. SVM Machine Learning Tutorial – What is the Support Vector Machine Algorithm, Explained with Code Examples. [Электронный ресурс] // freeCodeCamp. – 2020, 1 липня. – Режим доступа до ресурсу: <https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/> (дата звернення: 20.09.2025).
52. Wikipedia. K-nearest neighbors algorithm [Электронный ресурс]. — [б. р.]. — Режим доступа: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm (дата звернення: 22.09.2025).
53. K-Nearest Neighbor Algorithm for Machine Learning. [Электронный ресурс] // JavaTpoint. – Режим доступа до ресурсу: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> (дата звернення: 22.09.2025).
54. KNN: What is the K-Nearest Neighbors Algorithm. [Электронный ресурс] // DataScientest. — Режим доступа до ресурсу:

- <https://datascientest.com/en/knn-what-is-the-knn-algorithm> (дата
звернення: 22.09.2025).
- 55.NVIDIA. What Is XGBoost and Why Does It Matter? [Електронний ресурс]
// NVIDIA Glossary. — [б. п.]. — Режим доступу:
<https://www.nvidia.com/en-us/glossary/xgboost/> (дата
звернення: 22.09.2025).
- 56.Zhou D. A hybrid ensemble method for pulsar candidate classification.
[Електронний ресурс] // ResearchGate. — 2019. — Режим доступу до
ресурсу: https://www.researchgate.net/publication/335483097_A_hybrid_ensemble_method_for_pulsar_candidate_classification/figures дата
звернення: 22.09.2025).
- 57.Malik Sh., Harode R., Singh Kunwar A. XGBoost: A Deep Dive into Boosting
[Електронний ресурс] // SFU Professional Computer Science (Medium). —
04.02.2020. — Режим доступу: <https://medium.com/sfu-csmp/xgboost-a-deep-dive-into-boosting-f06c9c41349> (дата звернення: 10.11.2025).
- 58.Microsoft/LightGBM developers. LightGBM Documentation (stable)
[Електронний ресурс]. — [б. п.]. — Режим доступу:
<https://lightgbm.readthedocs.io/en/stable/> (дата звернення: 23.09.2025)
- 59.GeeksforGeeks. LightGBM Feature parameters [Електронний ресурс]. —
23.07.2025. — Режим доступу: <https://www.geeksforgeeks.org/machine-learning/lightgbm-feature-parameters/> (дата звернення: 19.09.2025).
- 60.Soni P. Evaluation Metrics — Classification [Електронний ресурс]. —
16.02.2022. — Режим доступу:
<https://medium.com/@priyanshsoni761/classification-evaluation-metrics-7c3fe3b0119b> (дата звернення: 09.11.2025).
- 61.Bilginç IT Academy. WHAT IS PYTHON? [Електронний ресурс]. — [б.
п.]. — Режим доступу: <https://bilginc.com/ca/blog/what-is-python-5792>
(дата звернення: 15.11.2025).
- 62.JetBrains. Configure a virtualenv environment [Електронний ресурс] // PyCharm 2025.2 Help. — 04.11.2025. — Режим доступу:

- <https://www.jetbrains.com/help/pycharm/creating-virtual-environment.html#env-requirements> (дата звернення: 11.11.2025).
63. Weszka J. S., Dyer C. R., Rosenfeld A. A Comparative Study of Texture Measures for Terrain Classification [Електронний ресурс] // IEEE Transactions on Systems, Man, and Cybernetics. — 1976. — Т. SMC-6, № 4. — С. 269–285. — Режим доступу: <https://scispace.com/papers/a-comparative-study-of-texture-measures-for-terrain-tit3xdjwdo> (дата звернення: 17.09.2025).
64. Wang F., Cheng C., Ren S. та ін. Prognostic Evaluation Based on Dual-Time 18F-FDG PET/CT Radiomics Features in Patients with Locally Advanced Pancreatic Cancer Treated by Stereotactic Body Radiation Therapy [Електронний ресурс] // Journal of Oncology. — 2022. — Vol. 2022. — Article ID 6528865. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9303166/> (дата звернення: 10.11.2025).
65. Wady S. H., Yousif R. Z., Hasan H. R. A Novel Intelligent System for Brain Tumor Diagnosis Based on a Composite Neutrosophic-Slantlet Transform Domain for Statistical Texture Feature Extraction [Електронний ресурс] // BioMed Research International. — 2020. — Vol. 2020. — Article ID 8125392. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7369660/> (дата звернення: 11.11.2025).
66. Galloway M. M. Texture analysis using gray level run lengths [Електронний ресурс] // Computer Graphics and Image Processing. — 1975. — Vol. 4, № 2. — Р. 172–179. — Режим доступу: <https://www.sciencedirect.com/science/article/abs/pii/S0146664X75800086> (дата звернення: 05.09.2025).
67. Payong A., Mukherjee S. Mastering Logistic Regression with Scikit-Learn: A Complete Guide [Електронний ресурс] // DigitalOcean Community Tutorials. — 20.03.2025. — Режим доступу:

- <https://www.digitalocean.com/community/tutorials/logistic-regression-with-scikit-learn> (дата звернення: 09.11.2025).
68. Breiman L. Random Forests [Електронний ресурс] // Machine Learning. — 2001. — Vol. 45, № 1. — Р. 5–32. — Режим доступу: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf> (дата звернення: 09.11.2025).
69. Platt J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods [Електронний ресурс] // Advances in Large Margin Classifiers. — 1999. — Р. 61–74. — Режим доступу: <https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf> (дата звернення: 23.09.2025).
70. Киричек М. Градієнтний бустінг як інструмент для вирішення задач класифікації в умовах обмежених даних [Електронний ресурс] // Технології та інжиніринг. — 2025. — № 2. — С. 37–47. — Режим доступу: <https://technologies-engineering.com.ua/en/journals/t-26-2-2025/gradiyentny-busting-yak-instrument-dlya-virishennya-zadach-klasifikatsiyi-v-umovakh-obmezhenikh-danikh> (дата звернення: 12.09.2025).
71. Hofmanninger J. Automated lung segmentation in CT under presence of severe pathologies. [Електронний ресурс] // GitHub. — Режим доступу: <https://github.com/JoHof/lungmask/blob/master/README.md> (дата звернення: 18.09.2025).
72. Wu J., Wang Y., Zhong Z. та ін. Vision-language foundation model for 3D medical imaging [Електронний ресурс] // npj Artificial Intelligence. — 2025. — Vol. 1. — Article 17. — Режим доступу: <https://www.nature.com/articles/s44387-025-00015-9> (дата звернення: 11.11.2025).
73. Розробка стартап-проектів. Хлебінська Олександра Ігорівна, Чупріна Маргарита Олександрівна IATE_ФБМІ [Електронний ресурс] – Режим

- доступу: <https://do.ipk.kpi.ua/course/view.php?id=7206> (дата звернення: 28.05.2025).
74. Nastenko I., Linnik M., Honcharuk M., Davydovych I., Lutchenko V., Babenko V., Dolinchuk L. Results of Machine Learning Applications in the Study of COVID-19 Associated Cardiopulmonary Pathology Using Computed Tomography Data [Електронний ресурс] // Innovative Biosystems and Bioengineering. – 2025. – Т. 9, № 4. – Р. 16-27— DOI: <https://doi.org/10.20535/ibb.2025.9.4.325335>. Режим доступу: <https://doi.org/10.20535/ibb.2025.9.4.325335>; <https://ibb.kpi.ua/article/view/325335> (дата звернення: 27.11.2025)
75. Євтушенко Т. Ю. Система прогнозування синдрому зникаючої легені за ознаками консолідації при гострій формі COVID-19 : дипломна робота бакалавра : 122 Комп'ютерні науки [Електронний ресурс] / Євтушенко Тарас Юлійович ; КПІ ім. Ігоря Сікорського. — Київ, 2024. — 55 с. — Режим доступу: <https://ela.kpi.ua/handle/123456789/69720>; <https://ela.kpi.ua/bitstreams/c67f970d-f3e8-4dbc-87fd-cc457c66c5c8/download> (дата звернення: 19.11.2025).
76. Sollmann N., Rayudu N. M., Lim J. J. S. та ін. Multi-detector computed tomography (MDCT) imaging: association of bone texture parameters with finite element analysis (FEA)-based failure load of single vertebrae and functional spinal units [Електронний ресурс] // Quantitative Imaging in Medicine and Surgery. — 2021. — Vol. 11, № 7. — Р. 2955–2967. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8250011/> (дата звернення: 19.11.2025).
77. Качко І. В. Класифікація дифузних захворювань печінки за статистикою вищого порядку зображень сонографії : дипломна робота бакалавра : 6.050101 Комп'ютерні науки [Електронний ресурс] / Качко Ілля Володимирович ; КПІ ім. Ігоря Сікорського. — Київ, 2019. — 91 с. — Режим доступу: <https://ela.kpi.ua/handle/123456789/28982>;

https://ela.kpi.ua/jspui/bitstream/123456789/28982/1/Kachko_bakalavr.docx

(дата звернення: 19.11.2025).

- 78.Фещенко Ю. І., Линник М. І., Гуменюк М. І. та ін. Діагностика синдрому зникаючих легень як ускладнення негоспітальної пневмонії вірусної етіології (COVID-19) [Електронний ресурс] // Infusion & Chemotherapy. — 2021. — № 3. — С. 5–11. — Режим доступу: <http://www.ifp.kiev.ua/doc/journals/ic/21/pdf21-3/5.pdf> (дата звернення: 19.11.2025).

Інформаційні ресурси:

- 79.ДУ «Національний інститут фтизіатрії і пульмонології ім. Ф. Г. Яновського НАМН України». Офіційний вебсайт [Електронний ресурс]. — Режим доступу: http://www.ifp.kiev.ua/index_ukr.htm (дата звернення: 19.11.2025).
- 80.Кафедра біомедичної кібернетики ФБМІ КПІ ім. Ігоря Сікорського. Науково-дослідні теми та проекти кафедри [Електронний ресурс]. — Режим доступу: <https://bmc.fbmi.kpi.ua/naukovo-doslidni-temi-ta-projects/> (дата звернення: 19.11.2025).