

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова
праця на правах рукопису

ДАВИДОВИЧ ІЛЛЯ ВІКТОРОВИЧ

УДК 004.89:616-073.75

ДИСЕРТАЦІЯ
МЕТОД КЛАСИФІКАЦІЇ ОНКОЛОГІЧНИХ ЗАХВОРЮВАНЬ ЛЕГЕНЬ НА
ОСНОВІ СТРУКТУРНИХ ПАТЕРНІВ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ
ТОМОГРАФІЇ

122 Комп'ютерні науки
12 Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії.

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

_____ Ілля ДАВИДОВИЧ

Науковий керівник: Настенко Євген Арнольдович, д.б.н., к.т.н., професор

Київ – 2026

АНОТАЦІЯ

Давидович І.В. Метод класифікації онкологічних захворювань легень на основі структурних патернів зображень комп'ютерної томографії. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» (12 - Інформаційні технології). — Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2026.

Дисертаційна робота присвячена дослідженню даних комп'ютерної томографії легень у пацієнтів із раком легень та розробці методів диференціації типів захворювання.

Онкологічні захворювання легень залишаються однією з провідних причин смертності у світі. За глобальною статистикою у 2022 році було діагностовано близько 2,48 млн нових випадків та зафіксовано 1,8 млн смертей. В Україні значна частка випадків виявляється на пізніх стадіях, що підкреслює критичну потребу своєчасної діагностики. Важливу роль у виявленні патології відіграє медична візуалізація, зокрема комп'ютерна томографія (КТ) органів грудної клітини, яка використовується як базовий інструмент діагностичного уточнення та моніторингу стану захворювання, що здатні знижувати смертність завдяки ранньому виявленню пухлин. Водночас збільшення обсягів КТ-даних і складність їх інтерпретації створюють істотне навантаження на спеціалістів і підсилюють запит на автоматизовані методи, здатні виділяти діагностично значущу інформацію. Диференціація гістологічних типів визначає вибір тактики лікування та прогноз, тому підвищення точності інтерпретації КТ-даних є клінічно значущим.

Проведено аналітичний огляд сучасних підходів до цифрової обробки та аналізу КТ-зображень легень у задачах онкологічної діагностики, включно з радіомікою, текстурними характеристиками, методами глибокого навчання, та підходів на основі Bag-of-Visual-Words для виділення інформативних ознак на зображеннях комп'ютерної томографії.

За результатами проведеного огляду встановлено, що комп'ютерна томографія містить приховані структурні закономірності, які можуть бути використані для автоматизованої класифікації типів раку. Існуючі методи, зокрема глибокі нейронні мережі, демонструють високу точність, проте потребують великих обсягів якісно анотованих даних і часто не є у повній мірі прозорими у процесі прийняття рішень. Це ускладнює їх клінічну інтерпретацію та впровадження.

Розглянуті сучасні методи мають проблеми зі стійкістю до варіативності вхідних даних та дисбалансу діагностичних класів при диференціації різних гістологічних типів. Тому постає задача вдосконалення методів виділення ознак та класифікації для забезпечення високої точності та інтерпретованості діагностики в умовах обмежених вибірок.

Для підвищення інформативності даних розроблено метод селекції інтенсивностей пікселів для попередньої обробки зображень, що дозволяє виокремити найбільш репрезентативні діапазони яскравості для кожного класу патології. В алгоритмі попередньої обробки враховано необхідність зменшення розмірності простору інтенсивностей шляхом операції квантування та відбору пікселів, що формують стійкі структурні компоненти зображень, ігноруючи шумову складову. Селекція виконується шляхом узгодження інтенсивностей пікселів з урахуванням типових розподілів яскравостей для кожного класу, що зменшує міжсканерну варіативність та пригнічує артефакти, підвищуючи стабільність подальшого аналізу.

Проведено дослідження із застосування методів виділення локальних структурних патернів для ідентифікації специфічних закономірностей у будові пухлинної тканини. На відміну від традиційних методів текстурного аналізу, які узагальнюють просторові залежності пікселів у вигляді інтегральних статистик, запропонований підхід формує явний словник мікроструктур шляхом сканування зображення ковзним вікном і подальшого підрахунку частот появи типових локальних конфігурацій. Таке подання є дискретизованим описом текстури, де замість обмеженої кількості агрегованих показників зберігається інформація про

склад і розподіл конкретних локальних структур. Перевагою запропонованого підходу є використання методів статистичного зважування ознак, що дозволяє зменшити вплив фонових патернів та підсилити внесок рідкісних, але потенційно діагностично значущих структур. Виділення унікальних патернів дозволяє перейти від аналізу окремих пікселів до оцінки частотного розподілу локальних конфігурацій. Дискретний словник виділених структур дає змогу аналізувати внесок окремих патернів у рішення моделі та полегшує експертну верифікацію виявлених закономірностей.

Для числового подання виокремлених структур проаналізовано декілька альтернативних схем формування опису ознак, які відрізняються ступенем узагальнення та чутливістю до частотного профілю патернів. Побудовані моделі та методика формування простору ознак надає можливість впорядкувати патерни за їх дискримінативною здатністю та сформувати вектори ознак, на основі яких побудовано класифікатор. Проаналізовано підходи до нормування частот патернів, їх зважування та відбору найбільш інформативних ознак, що забезпечує формування компактного вектора у єдиному порядку компонент, визначеному спільним словником патернів для всіх класів й наборів даних.

На основі сформованого набору ознак проведено дослідження, спрямоване на розробку архітектури класифікатора для багатокласової діагностики гістологічних типів раку. Запропоновано реалізацію багатокласового класифікатора за стратегію каскадного «один проти решти» (One-vs-Rest), на базі логістичної регресії, що дозволяє послідовно відокремлювати класи, зменшуючи неоднорідність вибірки на кожному кроці. Для підвищення точності класифікації розроблено алгоритм автоматичного відбору порядку слідування класів та адаптивного підбору порогів прийняття рішень, що дозволяє враховувати специфіку дисбалансу даних та мінімізувати помилки розпізнавання для найбільш складних для розпізнавання класів.

Проведено порівняльне дослідження ефективності запропонованого підходу з низкою моделей машинного навчання та глибокого навчання, включно з лінійними методами, ансамблями дерев рішень і градієнтного бустингу,

багаторівневими ансамблями, а також нейромережевими архітектурами. Перевірку виконано на незалежній тестовій вибірці, що дозволило оцінити узагальнювальну здатність підходу на нових даних.

За результатами порівняльних експериментів запропонований підхід за обраних параметрів попередньої обробки забезпечив точність на рівні 0.949 ± 0.048 та макро-F1-міри 0.916 ± 0.074 . Додаткова перевірка в умовах ізольованої обробки навчальної та тестової вибірок підтвердила відтворюваність отриманих результатів і практичну придатність для сценарію застосування на нових даних. У додатковому, більш консервативному сценарії отримано точність на рівні 0.805, макро-F1-міра 0.766.

Метою роботи було підвищення точності та інтерпретованості автоматизованої диференційної діагностики гістологічних типів раку легень за даними комп'ютерної томографії шляхом розроблення методу багатокласової класифікації на основі локальних структурних патернів.

У дисертації отримані такі наукові результати:

1. Запропоновано метод селекції інтенсивностей пікселів КТ-зображень у межах області інтересу, який зменшує вплив випадкових коливань яскравості та узгоджує вхідні дані шляхом побудови нормованих розподілів інтенсивностей і їх послідовного усереднення. Це забезпечує подальший відбір найбільш репрезентативної частини діапазону відтінків сірого та пригнічення фонових й неспецифічних значень.

2. Розроблено спосіб виокремлення часто повторюваних локальних структурних патернів для подальшої їх селекції з метою отримання компактного та інформативного набору ознак. Запропонований спосіб відрізняється формуванням словника структур і поєднанням частотного та дискримінативного відбору.

3. Розроблено алгоритм багатокласової класифікації за стратегією каскадного «один проти решти» (One-vs-Rest), що відрізняється послідовним відокремленням класів з покроковим звуженням множини альтернатив на кожному етапі та автоматизованим підбором порядку проходження класів. Це забезпечує

зменшення кількості помилок для малопредставлених і тому складних для розпізнавання класів.

Практичне значення отриманих результатів.

Реалізовано технологію формування структурно орієнтованого опису ознак на основі виокремлення локальних клас-специфічних патернів в області інтересу з подальшою селекцією інформативних структур.

За результатами проведених досліджень розроблено інформаційну систему автоматизованого аналізу зображень комп'ютерної томографії легень, орієнтовану на багатокласову диференціацію гістологічних типів раку легень.

Доведено практичну цінність моделі багатокласової класифікації за стратегією каскадного «один проти решти», що забезпечує підвищення стабільності результатів за нерівномірної представленості різних нозологічних груп.

Ключові слова: нейронні мережі, машинне навчання, обробка медичних зображень, медична візуалізація, сегментація зображення, класифікація зображень, точність класифікації, розподіл частот інтенсивностей пікселів, статистичний розподіл, ознаки, прогнозування, стратегія лікування, рання діагностика, структурні патерни, рак легень.

ABSTRACT

Davydovych I. A Method for Classifying Oncological Lung Diseases Based on Structural Patterns of Computed Tomography Images. — Qualifying scientific work (manuscript).

PhD thesis submitted for the degree of Doctor of Philosophy in specialty 122 “Computer Science” (12 – Information Technologies). — National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, 2026.

The dissertation is devoted to the study of lung computed tomography (CT) data in patients with lung cancer and to the development of methods for differentiating disease types.

Lung cancer remains one of the leading causes of mortality worldwide. According to global statistics, in 2022 approximately 2.48 million new cases were diagnosed and 1.8 million deaths were recorded. In Ukraine, a significant proportion of cases is detected at late stages, which emphasizes the critical need for timely diagnosis. Medical imaging plays an important role in detecting pathology, in particular chest computed tomography (CT), which is used as a basic tool for diagnostic verification and for monitoring the course of disease and can reduce mortality through early tumor detection.

At the same time, the increasing volume of CT data and the complexity of its interpretation place a substantial burden on specialists and intensify the demand for automated methods capable of extracting diagnostically significant information. Differentiation of histological types determines treatment strategy and prognosis; therefore, improving the accuracy of CT interpretation is clinically important.

An analytical review of modern approaches to digital processing and analysis of lung CT images for oncological diagnosis was conducted, including radiomics, texture characteristics, deep learning methods, and Bag-of-Visual-Words–based approaches for extracting informative features from CT images.

Based on the review, it was established that CT contains latent structural regularities that can be used for automated classification of cancer types. Existing methods, including deep neural networks, demonstrate high accuracy; however, they

require large volumes of high-quality annotated data and are often not fully transparent in the decision-making process. This complicates their clinical interpretation and implementation.

The considered modern methods face challenges in robustness to input data variability and in handling diagnostic class imbalance when differentiating histological types. Therefore, there is a need to improve feature extraction and classification methods to ensure high accuracy and interpretability under limited-sample conditions.

To increase data informativeness, a pixel-intensity selection method for image preprocessing was developed, enabling the identification of the most representative brightness ranges for each pathology class. The preprocessing algorithm accounts for the need to reduce the dimensionality of the intensity space through quantization and selection of pixels forming stable structural components of images while ignoring the noise component. Selection is performed by aligning pixel intensities with regard to typical brightness distributions for each class, which reduces inter-scanner variability and suppresses artifacts, thereby increasing the stability of further analysis.

A study was carried out using methods for extracting local structural patterns to identify specific regularities in the structure of tumor tissue. Unlike traditional texture analysis methods that summarize spatial pixel dependencies as integral statistics, the proposed approach forms an explicit microstructure dictionary by scanning the image with a sliding window and subsequently counting the occurrence frequencies of typical local configurations.

This representation is a discretized description of texture in which, instead of a limited set of aggregated indicators, information about the composition and distribution of specific local structures is retained.

An advantage of the proposed approach is the use of statistical feature weighting methods, which reduces the influence of background patterns and strengthens the contribution of rare but potentially diagnostically significant structures. Extracting unique patterns makes it possible to move from pixel-level analysis to assessing the frequency distribution of local configurations.

The discrete dictionary of extracted structures enables analysis of individual patterns' contributions to model decisions and facilitates expert verification of the discovered regularities.

For numerical representation of the extracted structures, several alternative schemes for forming feature descriptors were analyzed, differing in the degree of generalization and sensitivity to the frequency profile of patterns. The constructed models and the methodology for forming the feature space make it possible to rank patterns by their discriminative power and to form feature vectors on the basis of which a classifier is built. Approaches to frequency normalization, weighting, and selection of the most informative features were analyzed, ensuring the formation of a compact vector with a unified component order defined by a common pattern dictionary for all classes and datasets.

Based on the formed feature set, a study was conducted aimed at developing a classifier architecture for multiclass diagnosis of histological cancer types. An implementation of a multiclass classifier using the One-vs-Rest strategy in a cascade, based on logistic regression, was proposed; it enables sequential separation of classes, reducing sample heterogeneity at each step.

To improve classification accuracy, an algorithm for automatic selection of the class order in the cascade and adaptive tuning of decision thresholds was developed, which allows accounting for data imbalance specifics and minimizing recognition errors for the most difficult-to-recognize classes.

A comparative study of the effectiveness of the proposed approach was performed against a range of machine-learning and deep-learning models, including linear methods, ensembles of decision trees and gradient boosting, multi-level ensembles, as well as neural network architectures. Validation was carried out on an independent test set, which made it possible to evaluate the generalization ability of the approach on new data.

According to the comparative experiments, under the selected preprocessing parameters the proposed approach achieved an accuracy of 0.949 ± 0.048 and a macro-F1 score of 0.916 ± 0.074 . Additional validation under conditions of isolated processing of the training and test sets confirmed the reproducibility of the results and practical

applicability for deployment on new data. In an additional, more conservative scenario, an accuracy of 0.805 and a macro-F1 score of 0.766 were obtained.

The aim of the work was to improve the accuracy and interpretability of automated differential diagnosis of histological types of lung cancer from computed tomography data by developing a multiclass classification method based on local structural patterns.

The dissertation presents the following scientific results:

A method for selecting pixel intensities of CT images within the region of interest is proposed, which reduces the influence of random brightness fluctuations and aligns the input data by constructing normalized intensity distributions and their sequential averaging. This ensures subsequent selection of the most representative part of the grayscale range and suppression of background and nonspecific values.

A method for extracting frequently recurring local structural patterns for their subsequent selection is developed in order to obtain a compact and informative feature set. The proposed method is distinguished by building a structure dictionary and combining frequency-based and discriminative selection.

A multiclass classification algorithm using the One-vs-Rest strategy is developed, which differs by sequential separation of classes with stepwise narrowing of the set of alternatives at each stage and automated selection of the class traversal order. This reduces the number of errors for underrepresented and therefore difficult-to-recognize classes.

Practical significance of the obtained results.

A technology for forming a structurally oriented feature description has been implemented based on extracting local class-specific patterns in the region of interest followed by selection of informative structures.

Based on the conducted studies, an information system for automated analysis of lung CT images was developed, oriented toward multiclass differentiation of histological types of lung cancer.

The practical value of the multiclass classification model based on the One-vs-Rest strategy has been demonstrated, providing improved stability of results under uneven representation of different nosological groups.

Keywords: neural network, machine learning, medical image processing, medical visualization, image segmentation, image classification, classification accuracy, pixel intensity frequency distribution, statistical distribution, features, prediction, treatment strategy, early diagnosis, structural patterns, lung cancer.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЙНОЇ РОБОТИ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Юхимюк Р, Шкепаст М, Настенко Є, Лінник М, Давидович І, Бабенко В. Використання глибоких нейронних мереж для порівняльного аналізу норми, пневмонії і COVID-19. Біомедична інженерія і технологія. 2023; (12): 56-64. doi:10.20535/2617-8974.2023.12.292729

Внесок кожного з авторів: Юхимюк Р. - проектування та реалізація автоенкодера, написання статті, Шкепаст М. - проектування та реалізація CNN, написання статті, Настенко Є. - остаточне схвалення роботи, Лінник М. - експертна думка, **Давидович І. - концепція роботи та дизайн**, Бабенко В. - критичний огляд роботи.

2. Давидович І, Бабенко В. Методології аналізу зображень для діагностичних застосувань (огляд). Біомедична інженерія і технологія. 2024; (14): 54-63. doi:10.20535/2617-8974.2024.14.302135

Внесок кожного з авторів: **Давидович І. - концепція та дизайн роботи, аналіз матеріалів, написання статті**, Бабенко В. - критичний огляд, написання статті.

3. Perepeka E, Lazoryshynets V, Babenko V, et al. Cardiomyopathy Prediction in Patients with Permanent Ventricular Pacing Using Machine Learning Methods. System Research and Information Technologies. 2024; 1: 33-41. doi:10.20535/SRIT.2308-8893.2024.1.03

Внесок кожного з авторів: Перепека Є. - концепція роботи та дизайн, написання статті, Лазоришинець В. - написання статті, Бабенко В. - критичний огляд, **Давидович І. - критичний огляд**, Настенко Є. - остаточне схвалення статті.

4. Nastenko I, Linnik M, Honcharuk M, Davydovych I, Lutchenko V, Babenko V, et al. Diagnosis of COVID-19-associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: a Review of Methods and Future Research Directions.

Innovative Biosystems and Bioengineering. 2025; 9(4): 16-27.
doi:10.20535/ibb.2025.9.4.325335

Внесок кожного з авторів: Настенко Є. - остаточне схвалення статті, Линник М. - експертна думка, Гончарук М. - проектування та реалізація алгоритму виділення ознак, написання статті, **Давидович І. - проектування та реалізація алгоритму виділення ознак, написання статті**, Лутченко В. - проектування та реалізація класифікаторів, Бабенко В. - написання статті, Долінчук Л. - критичний огляд.

5. Давидович І, Настенко Є. Клас-балансний відбір інтенсивностей у КТ-зображеннях легень. Біомедична інженерія і технологія. 2025; (20): 15-25.
doi:10.20535/.2025.20.344360

Внесок кожного з авторів: **Давидович І. - концепція роботи та дизайн, проектування та реалізація алгоритму фільтрації, написання статті**, Настенко Є. - критичний огляд, остаточне схвалення статті.

6. Давидович І, Настенко Є. Визначення специфічних структурних патернів гістологічно підтверджених форм раку легень за даними КТ-зображень. Наука і техніка сьогодні. 2025; (11(52)): 2036-2057. doi:10.52058/2786-6025-2025-11(52)-2036-2057

Внесок кожного з авторів: **Давидович І. - концепція роботи та дизайн, проектування та реалізація способу вилучення ознак, проектування та реалізація моделей класифікації, написання статті**, Настенко Є. - критичний огляд, остаточне схвалення статті.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

7. Давидович І, Настенко Є. Клас-балансний підхід до фільтрації інтенсивностей у КТ-зображеннях легень. In: Scientific Innovation: Theoretical Insights and Practical Impacts: Collection of Scientific Papers with the Proceedings of the 4th International Scientific and Practical Conference; 2025 Dec 8-10; Naples, Italy (online). European Open Science Space; 2025. p. 141-144. doi:10.70286/EOSS-08.12.2025. ISBN: 979-8-89704-952-3

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

8. Давидович І, Настенко Є. Локальні текстурні патерни КТ-зображень легень для розмежування гістологічних форм раку. In: Research in Science, Technology and Economics: Collection of Scientific Papers with Proceedings of the 5th International Scientific and Practical Conference; 2025 Dec 10-12; Luxembourg, Luxembourg (online). International Scientific Unity; 2025. p. 257-260. doi:10.70286/ISU-10.12.2025. ISBN: 979-8-89704-985-1

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

9. Давидович І, Настенко Є. Вилучення ознак із локальних структурних патернів частотних карт КТ-зображень легень у багатокласовій класифікації. In: Modern Perspectives on Global Scientific Solutions: Collection of Scientific Papers with the Proceedings of the 7th International Scientific and Practical Conference; 2026 Jan 26-28; Bergen, Norway (online). European Open Science Space; 2026. p. 60-64. doi:10.70286/EOSS-26.01.2026. ISBN: 979-8-89704-959-2

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, ОДИНИЦЬ ВИМІРЮВАННЯ ТА СКОРОЧЕНЬ.....	18
ВСТУП.....	19
РОЗДІЛ 1 СУЧАСНІ ПІДХОДИ ДО ЦИФРОВОЇ ОБРОБКИ ТА АНАЛІЗУ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ ЛЕГЕНЬ	24
1.1 Зображення комп'ютерної томографії як джерело діагностичних даних.....	25
1.2 Формування областей інтересу на КТ-зображеннях.....	27
1.3 Радіоміка як метод кількісного аналізу КТ-зображень.....	31
1.3.1 Статистики першого порядку	32
1.3.2 Статистики другого та вищого порядку	33
1.3.3 Підходи до стандартизації вилучених ознак.....	34
1.4 Методи машинного навчання для аналізу КТ-зображень	35
1.5. Методи глибокого навчання для аналізу КТ-зображень	37
1.6 Альтернативний підхід на основі Bag-of-Visual-Words.....	39
Висновок до розділу 1	41
РОЗДІЛ 2 СЕЛЕКЦІЯ ІНТЕНСИВНОСТЕЙ ПІКСЕЛІВ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ.....	43
2.1 Набір даних КТ-зображень легень	43
2.2 Попередня обробка зображень та побудова локальних частот.....	50
2.2.1 Розподіл інтенсивностей пікселів на рівні окремого КТ-зрізу	52
2.2.2 Агрегація інтенсивностей пікселів до рівня пацієнта.....	53
2.2.3 Агрегація інтенсивностей пікселів до рівня класу	54
2.3 Відбір найбільш представлених інтенсивностей пікселів	55
2.4 Результат роботи клас-балансного фільтра.....	58
2.5 Оцінка впливу клас-балансного фільтра	61
Висновок до розділу 2	67
РОЗДІЛ 3 ВИДІЛЕННЯ ЛОКАЛЬНИХ СТРУКТУРНИХ ПАТЕРНІВ ТА ФОРМУВАННЯ ОЗНАК НА ОСНОВІ ЧАСТОТНИХ КАРТ ІНТЕНСИВНОСТЕЙ	68

	16
3.1 Виокремлення локальних патернів патології	68
3.2 Пошук клас-специфічних патернів	74
3.3 Формування ознак на основі клас-специфічних патернів	76
Висновок до розділу 3	81
РОЗДІЛ 4 СТРАТЕГІЯ «ОДИН ПРОТИ РЕШТИ» З ПОСЛІДОВНИМ ВІДОКРЕМЛЕННЯМ КЛАСІВ ДЛЯ ЗАДАЧІ БАГАТОКЛАСОВОЇ ДІАГНОСТИКИ.....	83
4.1 Багатокласова класифікація за стратегією «один проти решти»	83
4.2 Класифікація за стратегією каскадного «один проти решти».....	84
4.3 Виведення багатокласових ймовірностей	86
4.3.1 Додаткові механізми калібрування рішень.....	87
4.4 Автоматичний відбір порядку слідування класів.....	88
Висновок до розділу 4	90
РОЗДІЛ 5 ДИФЕРЕНЦІАЦІЯ ТИПІВ РАКУ ЛЕГЕНЬ МЕТОДАМИ МАШИННОГО НАВЧАННЯ.....	91
5.1 Підготовка навчальної та тестової вибірок	91
5.2 Обрані конфігурації для класифікації	92
5.2.1 Підбір оптимальних параметрів фільтра.....	96
5.2.2 Вплив квантування на результати класифікації	97
5.2.3 Вплив режиму квантування та жорсткості фільтрації інтенсивностей на результати класифікації	99
5.2.4 Вплив способу представлення ознак на результати класифікації	100
5.2.5 Вплив архітектури класифікатора.....	101
5.3 Оцінка стійкості класифікаційних моделей.....	105
5.3.1 Етап 1. Фільтрація зображень.....	106
5.3.2 Етап 2. Пошук структурних патернів	108
5.3.3 Етап 3. Виокремлення клас-специфічних патернів.....	114
5.3.4 Результати класифікації для незалежно оброблених вибірках	118
5.4 Практичне використання класифікаторів.....	122
Висновок до розділу 5	126

	17
ВИСНОВКИ	128
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	130
ДОДАТОК А СПИСОК ПУБЛІКАЦІЙ ТА ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ	143
ДОДАТОК Б АКТИ ВПРОВАДЖЕННЯ	146

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, ОДИНИЦЬ ВИМІРЮВАННЯ ТА СКОРОЧЕНЬ

КТ — комп'ютерна томографія

ОГК — органи грудної клітки

DICOM — Digital Imaging and Communications in Medicine

ROI — область інтересу на зображенні (Region of Interest)

HU — одиниці Гаунсфілда (Hounsfield Units)

BoVW — мішок візуальних слів (Bag-of-Visual-Words)

TF-IDF — Term Frequency – Inverse Document Frequency

GLCM — матриця співвідношень рівнів сірого (Gray-Level Co-occurrence Matrix)

OvR — «один проти решти» (One-vs-Rest)

PNG — Portable network graphics

TIFF — Tagged image file format

NLST — National Lung Screening Trial

ВСТУП

Актуальність теми дослідження. Рак легень залишається однією з провідних причин онкологічної смертності у світі та характеризується високими показниками захворюваності й летальності [1]. В Україні проблема загострюється недостатнім рівнем раннього виявлення, унаслідок чого значна частина випадків діагностується на пізніх стадіях [2].

Серед методів медичної візуалізації комп'ютерна томографія (КТ) є одним із найпоширеніших інструментів раннього виявлення підозрілих змін у легенях та зниження смертності від раку легень [3]. Разом із тим зростання обсягів КТ-досліджень і складність оцінювання морфологічних проявів уражень посилюють навантаження на лікарів і зумовлюють потребу в комп'ютерних засобах підтримки прийняття рішень, здатних забезпечити стандартизований та відтворюваний аналіз [4].

Сучасні методи машинного навчання, зокрема моделі на основі радіомічних ознак та глибокі нейронні мережі, розширюють можливості автоматизованої інтерпретації КТ-зображень у задачах класифікації та прогнозування. [5]. Водночас підходи глибокого навчання потребують великих масивів якісно анотованих даних, отримання яких у медичній практиці є ресурсомістким через необхідність залучення фахівців і уніфікації критеріїв анотації патологічних зон [6]. Крім того, складність внутрішніх механізмів таких моделей обмежує прозорість формування діагностичних рішень, що ускладнює їх клінічне тлумачення та стримує впровадження у практику [7]. Це визначає актуальність розроблення підходів, які поєднують інформативність опису ознак з відтворюваністю та можливістю пояснення результатів.

Особливої практичної значущості набуває багатокласова диференціація гістологічних типів раку легень за даними КТ, оскільки міжкласові відмінності нерідко проявляються як тонкі текстурні закономірності та супроводжуються нерівномірною представленістю діагностичних груп у вибірці і вираженою міжпацієнтною варіативністю [10, 19]. За таких умов доцільними є підходи, що

забезпечують стійкість до змін характеристик даних і дозволяють інтерпретувати внесок ознак у рішення класифікатора, зокрема через структурно орієнтоване формування простору ознак на основі локальних структурних патернів та використання класифікаційних стратегій, чутливих до дисбалансу й різної складності розпізнавання класів. Таким чином, актуальність дослідження зумовлена потребою удосконалення методів автоматизованого аналізу КТ-зображень легень для підвищення відтворюваності, інтерпретованості та якості багатокласової диференціації гістологічних типів раку, а також зменшення навантаження на лікарів у клінічній практиці.

Зв'язок роботи з науковими програмами, планами, темами.

Робота виконана в межах науково-дослідного проекту кафедри біомедичної кібернетики КПІ ім. Ігоря Сікорського, номер державної реєстрації № 0123U100866, за темою: "Методи та моделі ідентифікації станів об'єктів в задачах прийняття медичних рішень", зареєстровано від 01 січня 2023 р.

Мета роботи.

Метою роботи було підвищення точності та інтерпретованості автоматизованої диференційної діагностики гістологічних типів раку легень за даними комп'ютерної томографії шляхом розроблення методу багатокласової класифікації на основі локальних структурних патернів.

Для досягнення поставленої мети необхідно було вирішити наступні **завдання**:

1. Проаналізувати сучасний стан досліджень у галузі аналізу КТ-зображень легень методами машинного навчання та визначити перспективні підходи для задач диференційної діагностики онкологічних патологій.
2. Дослідити особливості КТ-зображень і процедур їх попередньої обробки, що впливають на відтворюваність та точність автоматизованої класифікації.
3. Розробити метод формування ознак за даними зображень комп'ютерної томографії на основі локальних структурних характеристик.

4. Дослідити вплив різних методів формування комплексних ознак на їх здатність розділяти класи.
5. Розробити метод класифікації, що враховує різну складність міжкласового розмежування при обмеженому розмірі вибірки.
6. Реалізувати запропонований метод у вигляді системи диференційної діагностики.

Об'єкт дослідження – процес автоматизованої комп'ютерної діагностики та диференціації гістологічно підтверджених форм раку легень за даними КТ-зображень грудної клітки.

Предмет дослідження – методи та алгоритми виділення інформативних структурних ознак з КТ-зображень легень і побудови інтерпретованої багатокласової моделі класифікації.

Методи дослідження.

Методи цифрової обробки медичних зображень, методи статистичного аналізу та селекції ознак, методи машинного навчання для багатокласової класифікації.

Наукова новизна отриманих результатів

1. Запропоновано метод селекції інтенсивностей пікселів КТ-зображень у межах області інтересу, який зменшує вплив випадкових коливань яскравості та узгоджує вхідні дані шляхом побудови нормованих розподілів інтенсивностей і їх послідовного усереднення. Це забезпечує подальший відбір найбільш репрезентативної частини діапазону відтінків сірого та пригнічення фонових й неспецифічних значень.
2. Розроблено спосіб виокремлення часто повторюваних локальних структурних патернів для подальшої їх селекції з метою отримання компактного та інформативного набору ознак. Запропонований спосіб відрізняється формуванням словника структур і поєднанням частотного та дискримінативного відбору.
3. Розроблено алгоритм багатокласової класифікації за стратегією каскадного «один проти решти» (One-vs-Rest), що відрізняється послідовним відокремленням класів з покроковим звуженням множини альтернатив на кожному

етапі та автоматизованим підбором порядку проходження класів. Це забезпечує зменшення кількості помилок для малопредставлених і тому складних для розпізнавання класів.

Практичне застосування.

1. Реалізовано технологію формування структурно орієнтованого опису ознак на основі виокремлення локальних клас-специфічних патернів в області інтересу з подальшою селекцією інформативних структур.

2. За результатами проведених досліджень розроблено інформаційну систему автоматизованого аналізу зображень комп'ютерної томографії легень, орієнтовану на багатокласову диференціацію гістологічних типів раку легень.

3. Доведено практичну цінність моделі багатокласової класифікації за стратегією каскадного «один проти решти», що забезпечує підвищення стабільності результатів за нерівномірної представленості різних нозологічних груп.

Особистий внесок здобувача.

Здобувачем визначено мету дослідження, сформульовано завдання та розроблено методи і підходи до їх реалізації. Проведено аналіз сучасного стану застосування методів штучного інтелекту до задачі багатокласової диференціації гістологічних типів раку легень за даними комп'ютерної томографії. Розроблено підхід до попередньої обробки КТ-зображень та методику формування структурно орієнтованого опису ознак на основі локальних клас-специфічних патернів із подальшою селекцією інформативних компонент. Запропоновано та реалізовано стратегію багатокласової класифікації «один проти решти», що послідовно відокремлює класи з покроковим звуженням множини альтернатив на кожному етапі та автоматизованим підбором порядку їх класів. Побудовано статистичні моделі та проаналізовано можливості програмної реалізації алгоритмів, проведено експерименти на даних комп'ютерної томографії легень, проаналізовано та узагальнено результати досліджень. Загальний дизайн дослідження, статистичну та алгоритмічну постановку задач, а також розроблення моделей і технічних рішень виконано у співпраці з д.б.н., к.т.н. Настенко Є. А.

Апробація результатів дисертації.

Матеріали дисертаційної роботи доповідались та були обговорені на науково-практичних конференціях:

– Scientific Innovation: Theoretical Insights and Practical Impacts : Collection of Scientific Papers with the Proceedings of the 4th International Scientific and Practical Conference (December 8–10, 2025, Naples, Italy). – European Open Science Space, 2025. – 507 p. – ISBN 979-8-89704-952-3. – DOI: 10.70286/EOSS-08.12.2025.

– Research in Science, Technology and Economics : Collection of Scientific Papers with Proceedings of the 5th International Scientific and Practical Conference (December 10–12, 2025, Luxembourg, Luxembourg). – International Scientific Unity, 2025. – 735 p. – ISBN 979-8-89704-985-1. – DOI: 10.70286/ISU-10.12.2025.

– Modern Perspectives on Global Scientific Solutions : Collection of Scientific Papers with the Proceedings of the 7th International Scientific and Practical Conference (January 26–28, 2026, Bergen, Norway). – European Open Science Space, 2026. – ISBN 979-8-89704-959-2. – DOI: 10.70286/EOSS-26.01.2026.

Публікації.

Результати дисертаційної роботи опубліковані в 6 наукових працях, з яких 4 статті у наукових фахових виданнях України категорії Б за спеціальністю 122 – комп’ютерні науки, у 2 що включені до міжнародної наукометричної бази даних Scopus.

Впровадження.

Результати роботи були впроваджені в ДУ «Національний науковий центр фтизіатрії, пульмонології та алергології ім. Ф. Г. Яновського НАМН України» (акт впровадження від 09 грудня 2025 року).

Структура й обсяг дисертації.

Дисертаційна робота складається зі вступу, 5 розділів, висновків, списку літературних джерел. Загальний обсяг роботи складає 147 сторінки. Обсяг основного тексту становить 125 сторінок. Робота містить 11 таблиць, 58 формул та 32 рисунки. Список літературних джерел складається з 113 найменувань на 13 сторінках.

РОЗДІЛ 1

СУЧАСНІ ПІДХОДИ ДО ЦИФРОВОЇ ОБРОБКИ ТА АНАЛІЗУ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ ЛЕГЕНЬ

Онкологічні захворювання залишаються однією з провідних причин смертності у світі. Згідно з глобальною статистикою, у 2022 році було діагностовано близько 2,48 мільйона нових випадків раку легень та зафіксовано 1,8 мільйона смертей від нього – це робить рак легень найпоширенішим видом раку та головною причиною смерті від раку у світі [1]. Для порівняння, в Україні за 2022 рік зареєстровано 155 239 нових випадків злоякісних новоутворень та 84 153 пов'язаних із ними смертей. З них рак легень діагностовано у приблизно 18% чоловіків та лише 1,4% жінок. Така гендерна різниця пояснюється історично більш високою поширеністю факторів ризику, зокрема тютюнокуріння серед чоловіків. Водночас рак легень має вкрай несприятливий прогноз: в Україні частка випадків, виявлених на пізніх метастатичних стадіях, сягає 38,5% [2]. Ці факти підкреслюють критичну необхідність своєчасної діагностики та ефективних інструментів виявлення патології.

Медична візуалізація відіграє ключову роль у сучасній діагностиці та лікуванні онкологічних патологій. Методи променевої діагностики, зокрема комп'ютерна томографія (КТ) органів грудної клітини (ОГК), є базовим інструментом для виявлення та моніторингу раку легень [3,4]. Масові скринінгові програми, наприклад, низькодозова КТ для виявлення раку легень, показали здатність знижувати смертність на 20–43% завдяки виявленню пухлин на ранніх стадіях[5]. Однак зростання обсягів медичних зображень та їх складності створює неабияке навантаження на лікарів-рентгенологів. В умовах скринінгу спеціалісти змушені переглядати тисячі зображень, намагаючись розпізнати дрібні вузлики на тлі шуму і артефактів [6]. Людський фактор та суб'єктивність оцінки можуть призводити до пропуску патології або помилкової діагностики. В той же час за останні десятиліття було розроблено широкий спектр методологій аналізу медичних зображень – галузі, що поєднує методи обробки зображень,

комп'ютерного зору та штучного інтелекту для автоматичного виділення діагностично значущої інформації з медичних зображень. Підхід є ефективним для виявлення патологічних змін й підтримки клінічних рішень на основі візуальних даних [7].

Метою цього розділу є узагальнення сучасних методів та підходів автоматичного аналізу медичних КТ-зображень для діагностики раку легень, оцінка їх ефективності та тенденції розвитку у сфері медичних інформаційних технологій.

1.1 Зображення комп'ютерної томографії як джерело діагностичних даних

Комп'ютерна томографія є одним із провідних методів променевої діагностики патології органів грудної клітки та базовим інструментом виявлення й подальшої оцінки новоутворень у легенях [4]. На відміну від традиційної рентгенографії, КТ забезпечує пошарове відображення анатомічних структур без суттєвого накладання тіней, що підвищує чутливість до дрібних вогнищ, змін щільності та структурних деформацій у легеневій паренхімі. КТ є методом реконструкції поперечних зрізів на основі множинних вимірювань ослаблення рентгенівського випромінювання при обертанні джерела та детекторів навколо пацієнта. Отримані проекційні дані перетворюються у зображення зрізів, які відображають розподіл коефіцієнтів ослаблення у тканинах [8]. Результатом перетворення є цифрове зображення, у якому кожному пікселю відповідає числове значення, що характеризує рентгенівську щільність тканини. Для уніфікації інтерпретації застосовується шкала Гаунсфілда (Hounsfield units, HU), яка задає стандартизовану точку відліку відносно щільності води (0 HU) та повітря (-1000 HU). Таким чином, КТ-зображення є кількісним представленням відмінностей щільності, що має принципове значення для стандартизованого опису легеневих структур у клінічній практиці [9,10].

Оскільки діапазон HU значно ширший за можливості одночасного відображення на екрані, на практиці використовується операція вікно/рівень — відображення вибраного інтервалу щільності шляхом налаштування рівня (або центра встановленого на розподілі інтенсивностей) й ширини вікна. Це дозволяє цілеспрямовано підвищувати контраст релевантних тканин. Для візуалізації легеневих структур застосовують так зване легеневе вікно з центром приблизно $-600 \dots -700$ HU та шириною $1000 \dots 1500$ HU, що забезпечує одночасну видимість органів паренхіми з низькою щільністю та більш щільних компонентів, такі як судини, бронхіальні стінки, вузлики, підвищуючи контраст між ними [11].

Використання КТ у клінічній практиці на пацієнтах з підозрою на рак поділяється на кілька типових сценаріях. Перший — скринінг у групах високого ризику. Низькодозова КТ (LDCT) продемонструвала здатність зменшувати смертність від раку легень у рандомізованих дослідженнях, зокрема, у National Lung Screening Trial (NLST) порівняно з рентгенографією грудної клітки було продемонстровано зниження ризику смерті у групі LDCT [12], а дослідження [13] підтвердило наявність користі від об'ємної КТ-стратегії скринінгу в осіб високого ризику. Другий сценарій це випадкове виявлення новоутворень на КТ за інших причин діагностики. Через частоту таких знахідок важливими є стандартизовані підходи до подальшого спостереження та мінімізації надмірного втручання [14]. Третій сценарій це діагностичне уточнення та визначення стадії при відомому або підозрюваному раку легень. У цьому контексті КТ із контрастуванням використовується для оцінки локального поширення пухлини [15]. Насамкінець четвертий сценарій — моніторинг у динаміці та оцінка відповіді на лікування. Серійні КТ-дослідження дозволяють кількісно відстежувати зміну розмірів пухлинних вогнищ, появу нових уражень або прогресію [16].

Попри високу інформативність, застосування КТ у задачах онкологічної діагностики має низку обмежень, що потребують врахування при інтерпретації. До них належать залежність якості та порівнюваності результатів від параметрів сканування й реконструкції (товщина зрізу, алгоритм реконструкції, доза, артефакти руху), вплив контрастування, а також потенційна неоднозначність

морфологічних ознак при супутніх запальних або фібротичних змінах. Реконструкцією тут є математичне перетворення вихідних проекційних даних сканера (так звані *sinograms*) у поперечні зрізи або 3D-об'єм. Вибір алгоритму реконструкції (наприклад, *filtered back projection* або ітеративні методи) задає компроміс між шумом, просторовою роздільною здатністю та артефактами, що може впливати як на візуальну оцінку, так і на кількісні характеристики зображення [17]. У сучасних оглядах з онкоторакальної візуалізації підкреслюється необхідність коректного вибору протоколу й критичного аналізу обмежень методу при визначенні стадії та оцінці відповіді при лікуванні [11].

Узагальнюючи, комп'ютерну томографію у контексті раку легень слід розглядати як універсальний діагностичний інструмент, що одночасно забезпечує раннє виявлення, стандартизоване ведення інцидентальних вузлів, діагностичне уточнення та стадіювання, а також динамічний контроль плину хвороби у часі. Саме така роль цього методу створює методологічне підґрунтя для подальшого розгляду способів підвищення об'єктивності та відтворюваності аналізу КТ-зображень у прикладних дослідженнях.

1.2 Формування областей інтересу на КТ-зображеннях

У контексті комп'ютерного аналізу КТ-зображень дослідники спираються не лише на візуальну оцінку, а й використовують поняття ознак — кількісних характеристик зображення, які формалізують властивості тканин і можуть бути використані для подальшого статистичного аналізу. Застосування легеневого вікна під час аналізу КТ-зображень дає змогу обмежити дослідження переважно легеневою паренхімою та мінімізувати вплив сигналу від кісткових і структур м'яких тканин, які не є інформативними для оцінювання легневих процесів. Водночас, з огляду на анатомічну специфіку та складну геометрію легень, стандартна віконна фільтрація може частково включати до області аналізу позалегенові структури. Це, своєю чергою, створює ризик формування ознак, пов'язаних із зовнішніми артефактами, що потенційно зміщує фокус аналізу від

легеневих характеристик і може погіршувати узагальнюваність отриманих результатів [18].

Саме тому радіомічні ознаки обчислюються не по всьому зображенню, а в межах певної виділеної області інтересу (ROI). Фактично здійснюється бінарне маскування, тобто сегментація області інтересу та маскування нерелевантних для аналізу ділянок, де лишається виключно патологія яку надалі аналізуватимуть [19,20]. Варто також зазначити, що перед проведенням виділення областей інтересу, крім встановлення відповідного режиму вікна відображення до попередньої обробки також включають нормалізацію та ресемплінг. Нормалізація у контексті зображень КТ це приведення інтенсивностей до узгодженого представлення, наприклад, те ж обмеження діапазону HU або стандартизація рівнів сірого перед обчисленням ознак, щоб зменшити варіабельність, пов'язану з відмінностями протоколів і дискретизації сигналу [17]. Ресемплінг або перерахунок до єдиного розміру пікселя (часто ізотропного) виконується інтерполяцією для того, щоб мінімізувати залежність кількісних показників від товщини КТ-зрізу та кроку пікселя. Водночас він може змінювати структуру шуму й текстур, тому має бути чітко задокументований у методиці [21].

На практиці виділення областей інтересу може здійснюватися різними способами. Еталонним підходом, з погляду клінічної змістовності вважається ручне виділення, що здійснюється у присутності експерта. Саме такий підхід часто використовується як референт для оцінки автоматичних методів, проте він потребує контролю узгодженості та консенсусу, крім того може бути трудомістким та тривалим по часу [20].

Існують також напівавтоматичні методи, які поєднують алгоритмічну сегментацію з мінімальною участю експерта, задача якого зводиться до вибору початкової точки виділення, грубого обмеження області та контролю результатів. Це дозволяє зменшити трудомісткість порівняно з повністю ручним виділенням, водночас зберігаючи можливість клінічно осмисленого контролю меж сегментації [22]. Одним з найпоширеніших підходів є порогова сегментація (thresholding). Вона базується на тому, що тканини та середовища у КТ мають різні діапазони щільності

в одиницях Гаунсфілда, тому область інтересу можна виділити як сукупність пікселів, значення яких потрапляють у заданий інтервал. Важливо, що пороги можуть підбиратися вручну або автоматично [23,24]. Класичний приклад автоматичного підбору порога є метод Оцу, який знаходить поріг, що найкраще розділяє гістограму інтенсивностей на дві групи, умовно на об'єкт та фон, за статистичним критерієм міжкласової відмінності [25]. Перевагою напіваавтоматичних підходів із заданим або автоматичним порогом є швидкість та прозорість методу, однак вони погано працюють в умовах частково видимого об'єкта, наявності шуму, артефактів та при суміжності структур зі схожою щільністю. У задачах сегментації легеневих вузликів це проявляється як витік області виділення у судини або стінки бронхів, особливо коли контури вузлів нечіткі [26,27]. Близьким за логікою є метод нарощування області (region growing), зокрема у варіанті seeded region growing, коли експерт задає одну або кілька початкових точок (seed) всередині утворення, після чого алгоритм послідовно додає сусідні пікселі, які є подібними за інтенсивністю до вже включеної області [28]. Перевага цього підходу в тому, що він природно пристосовується до локальних властивостей конкретного ураження, водночас лишається чутливим до шуму й неоднорідності, а при анатомічній близькості судин та плеври може захоплювати сусідні структури, якщо їх щільність близька до щільності ураження. Дуже поширеною групою напіваавтоматичних підходів є методи активних контурів, які формують сегментацію як задачу оптимізації. Вони представляють межу об'єкта як криву, яка зростає під дією двох чинників: внутрішньої, що згладжує контур, та зовнішньої, що тягне його до меж об'єкта, зазвичай пов'язаних із градієнтом інтенсивності [29]. На практиці це корисно для уточнення межі ураження, коли експерт може грубо окреслити область, а алгоритм підтягує контур до найбільш імовірних країв. Обмеженням є залежність від заданої початкової точки та складність роботи зі слабо вираженими межами [30,31].

Зрештою застосовуються і повністю автоматичні підходи до сегментації із використанням сучасних нейронних мереж. Узагальнюючі огляди показують, що нейромережеві моделі здатні витягувати багаторівневі просторові ознаки з КТ-

даних і забезпечувати кращу узгодженість сегментації порівняно з класичними методами за умови коректного навчання і валідації [32]. У задачах аналізу зображень легень найчастіше автоматизують один або кілька рівнів виділення областей інтересу:

- маска легень, як первинне відокремлення паренхіми від решти грудної клітки [33];
- маски часток або сегментів для локалізації ураження в анатомічному контексті [34];
- маска підозрілого утворення, вузлика або пухлини [35].
- перитуморальна зона (кільце навколо пухлини) як окрема ROI у частині сучасних робіт [36–39].

На практиці часто застосовують конвеєрний підхід. Спершу сегментують легені, щоб обмежити область пошуку, потім виконується детекція кандидатів, і лише після цього — точну сегментацію конкретного утворення у вигляді маски [40,41]. Найвідомішою архітектурою для медичної сегментації є U-Net. Подібні архітектури мереж використовують механізми уваги для фокусування на найбільш релевантних ділянках, де енкодер виділяє ознаки та контекст, а декодер відновлює просторову деталізацію (рисунок 1.1) [42].

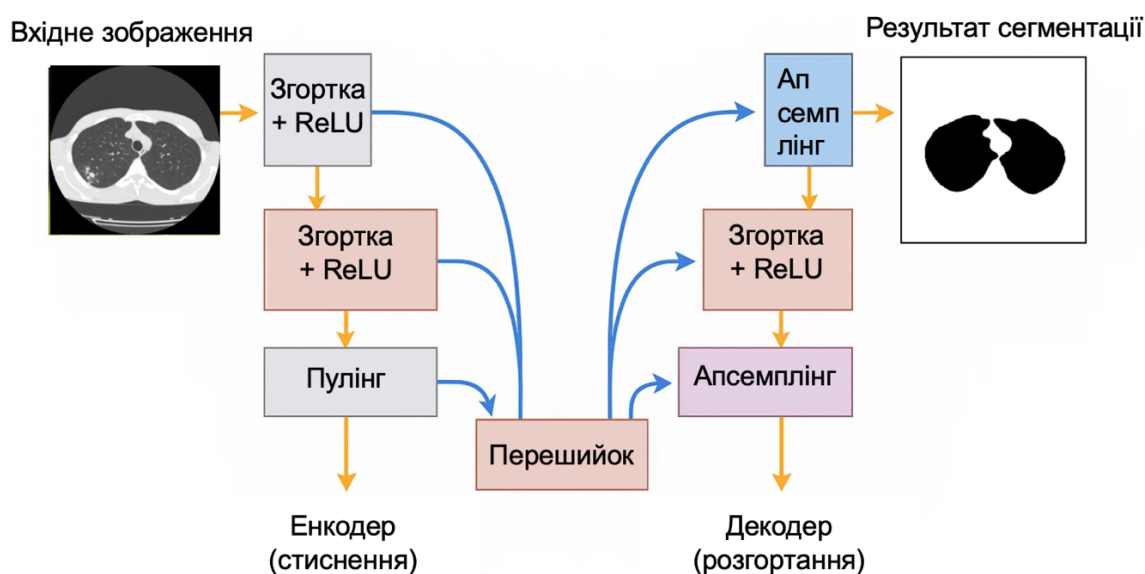


Рисунок 1.1 — Архітектура U-Net для сегментації зображень

Конкретно для сегментації легень, вже існують нейронні мережі на цій архітектурі, які здатні забезпечувати високу якість сегментації, до 96% ділянки легеневої паренхіми [43].

1.3 Радіоміка як метод кількісного аналізу КТ-зображень

Радіоміка — це методологія кількісного аналізу медичних зображень, що полягає в автоматизованому вилученні великої кількості стандартизованих ознак, наприклад, характеристик інтенсивності, текстурних та морфологічних тощо, із сегментованих ділянок області інтересу та їх подальшому використанні в статистичному або машинному моделюванні для опису, прогнозування чи стратифікації клінічних станів [44]. У контексті аналізу КТ-зображень легень із наявними онкологічними патологіями це дозволяє розглядати зображення як джерело кількісної інформації, що відображає фенотип пухлини та її внутрішню гетерогенність на рівні, який часто є недоступним або слабко відтворюваним при суто візуальній інтерпретації [45]. Емпіричну доцільність радіомічних підходів у онкології підсилюють роботи, де показано асоціації між комплексними наборами таких ознак і чітко визначеними клінічними результатами, з якими асоціюються кількісні характеристики, за умов належного контролю, які можуть спричиняти хибні зв'язки через специфіку даних або дизайну дослідження, а не через реальну закономірність [46–48].

Аналіз за допомогою радіоміки зазвичай описують як послідовність взаємопов'язаних етапів:

1. Отримання зображення, налаштування вікна відображення, нормалізація;
2. Визначення області інтересу;
3. Попередня обробка, фільтрація, дискретизація інтенсивностей тощо;
4. Обчислення ознак;
5. Відбір ознак та побудова моделі;
6. Валідація з оцінкою відтворюваності й узагальнюваності результатів.

Концептуальною особливістю тут є перетворення зображення, як візуального об'єкта на матрицю вимірюваних даних, де кожна ознака є формалізованою статистикою інтенсивностей, просторових взаємозв'язків або геометрії у межах області інтересу [49]. На практиці ознаки, що можна сформулювати, розділяють на три групи: статистика першого порядку (ознаки інтенсивності), статистики другого та вищого порядку (текстурні ознаки), ознаки форми та морфології.

1.3.1 Статистики першого порядку

Описують статистичні властивості розподілу інтенсивностей в межах визначеної області інтересу, тобто працюють з гістограмою розподілу значень без урахування просторових взаємозв'язків між пікселями. До цієї групи належать показники центральної тенденції (середнє, медіана), розсіювання (дисперсія, стандартне відхилення, міжквартильний розмах), форми розподілу (асиметрія, ексцес), а також показники, що відображають концентрацію чи неупорядкованість розподілу (енергія, ентропія, однорідність тощо) [50]. Перевагою цієї групи ознак з точки зору КТ-зображень є те, що інтенсивність має фізичну інтерпретацію у шкалі Гаунсфілда, що створює передумови для змістовного трактування, наприклад як кількісних маркерів щільності тканини, наявності компонентів різної щільності (некротичний, фіброзний, емфізематозний фон тощо), а також грубого рівня неоднорідності ураження. При цьому діагностична цінність таких характеристик у випадку легневих пухлин частіше пов'язана не з середньою щільністю як такою, а з комплексним описом розподілу інтенсивностей (перцентилі, характеристики хвостів розподілу, ентропія), який здатний відображати гетерогенність тканинного складу. [51].

Втім є і недоліки, що виражаються у чутливості до технічних факторів сканування. Зміни товщини КТ-зрізу, шуму та параметрів реконструкції можуть суттєво впливати на форму гістограми розподілу інтенсивностей пікселів, особливо на межах пухлин, що контактують з повітрям, або при наявних домішках судин, бронхів у сегментації.

Дослідження відтворюваності на повторних скануваннях та при альтернативних реконструкціях показують, що узгодженість ознак суттєво

погіршується, коли змішуються різні типи реконструкцій, і це стосується не лише текстур, але й частини статистик інтенсивності пікселів [52].

1.3.2 Статистики другого та вищого порядку

Дані статистики спрямовані на вилучення характеристик текстури та призначені для кількісного опису просторової організації інтенсивностей, їх упорядкованості або неоднорідності. Концептуально вони походять від класичних методів аналізу текстур у комп'ютерному зорі. Зокрема, ознаки Гараліка, що ґрунтуються на матрицях співвідношень інтенсивностей, історично є одним з фундаментів текстурної аналітики та згодом були адаптовані й стандартизовані для медичної радіоміки [53,54]. Цей підхід використовує низку матриць текстурних характеристик, серед яких найпоширенішою є матриця співвідношень рівнів сірого (Gray Level Co-Occurrence Matrix, GLCM). Вона оцінює ймовірність появи певних пар значень інтенсивності на заданій відстані та у визначеному напрямку в межах області інтересу [55]. Це дає змогу кількісно описати частоти переходів між інтенсивностями в сусідніх пікселях і на цій основі обчислювати показники контрастності, однорідності, кореляції, та ентропії переходів.

Менш поширеною, проте корисною в окремих сценаріях, є матриця довжин послідовностей рівнів сірого (Gray Level Run Length Matrix, GLRLM). Вона виявляє закономірності шляхом оцінювання просторових послідовностей пікселів з однаковою інтенсивністю та є чутливою до дрібно- й крупнозернистих текстур, а також до наявності протяжних однорідних ділянок [53,56]. Матриця розмірів зон зв'язних компонентів рівнів сірого (Gray Level Size Zone Matrix, GLSZM), забезпечує аналіз зображення через оцінювання розподілу ділянок однакової інтенсивності та дає змогу отримувати характеристики їхньої зв'язності й розмірних параметрів [57].

Окрему групу становлять матриця різниць рівнів сірого у локальному оточенні (Neighbouring Gray Tone Difference Matrix, NGTDM) та матриця залежності рівнів сірого (Gray Level Dependence Matrix, GLDM). Відповідні ознаки пов'язані з мірами грубості, зернистості та локальної структурованості [58].

Ключовою ідеєю текстурних ознак є те, що пухлина на зображеннях комп'ютерної томографії може демонструвати різні патерни неоднорідності (зумовлені некрозом, фіброзом, ангіогенезом, різною клітинністю та мікрооточенням), які не завжди надійно кодуються простими візуальними категоріями, але можуть бути частково формалізовані статистиками просторових залежностей інтенсивності. Саме тому застосування радіоміки в діагностиці онкології розглядається як підхід до вимірювання фенотипу пухлини за даними стандартної візуалізації [51].

1.3.3 Підходи до стандартизації вилучених ознак

Однією з ключових причин обмеженої відтворюваності результатів радіомічних досліджень між різними центрами обстеження є варіативність визначень однойменних ознак у різних програмних реалізаціях. Зокрема, відмінності можуть стосуватися математичного формулювання, способів дискретизації, правил обробки меж ROI, а також вибору двовимірної або тривимірної постановки. З метою усунення цієї проблеми було створено Image Biomarker Standardisation Initiative (IBSI) — міжнародну ініціативу, що стандартизує визначення радіомічних ознак і надає еталонні значення на референтних наборах даних для валідації та калібрування програмних реалізацій [50]. У базовому звіті IBSI було продемонстровано, що на початковому етапі між командами розробників існували суттєві розбіжності у значеннях частини обчислених ознак. Також показано, що досягнення консенсусу вимагає не лише чітких формальних визначень, але й наявності тестових об'єктів (цифрових або фізичних фантомів) та стандартизованих процедур верифікації.

Практичний висновок для дослідника полягає в тому, що в сучасних радіомічних роботах доцільно: 1) явно зазначати, чи відповідає інструмент вилучення ознак вимогам IBSI; 2) документувати ключові параметри обробки (ресемплінг, дискретизацію, фільтри); 3) за можливості виконувати технічну валідацію ознак на еталонних даних.

1.4 Методи машинного навчання для аналізу КТ-зображень

У межах радіомічного підходу ознаки є лише проміжним представленням між зображенням і рішенням. Вони забезпечують інформацію у вигляді компактного числового опису, яка стане основою для побудови моделей розрізнення патологічних станів або прогнозування діагностично значущої інформації. У такій постановці класичні методи машинного навчання зберігають методологічну значущість, оскільки забезпечують інтеграцію радіомічних і клінічних змінних у межах єдиної моделі та нерідко характеризуються вищою інтерпретованістю й кращим контролем ризику перенавчання порівняно зі складними високопараметричними підходами. У радіомічних дослідженнях онкології легень такі моделі застосовують як у діагностичних задачах, наприклад, оцінювання злоякісності легеневого вузлика, так і в прогностичних або предиктивних задачах, наприклад, прогноз ризику, формуючи підґрунтя для розроблення інструментів підтримки клінічних рішень [44]. З прикладної точки зору, специфіка радіомічних даних створює типові статистичні виклики, як висока розмірність, мультиколінеарність, обмежений розмір вибірки, а також ризики витоку інформації під час відбору ознак та налаштуванні гіперпараметрів. Тому вибір алгоритму має розглядатися разом із вибором стратегії регуляризації, схемою валідації та правилами відтворюваного препроцесингу [59].

Логістична регресія (ЛР) є базовою моделлю як для задач бінарної класифікації, у яких цільова змінна відображає наявність або відсутність події, так і для багатокласової класифікації, де цільова змінна набуває одного з кількох взаємовиключних класів. У багатокласовому випадку найчастіше застосовують підхід один-проти-всіх або мультиноміальну логістичну регресію, що оцінює ймовірності належності об'єкта до кожного класу [60]. Однак у радіомічній постановці часто спостерігається ситуація, коли наявно багато ознак при малій кількості спостережень, за якої класична логістична регресія має підвищений ризик перенавчання [51]. В таких випадках застосовують регуляризацію та/або попередній відбір ознак. Регуляризацію можна розглядати як введення до функції

вtrat штрафного члена, що обмежує складність моделі. Найпоширенішою є L1-регуляризація (Lasso), яка фактично відсіює зайві ознаки, стискаючи частину коефіцієнтів до нуля [61]. L2-регуляризація (Ridge) є доцільною за наявності мультиколінеарності та нестабільності оцінок, оскільки забезпечує усадку коефіцієнтів, зменшуючи їхні значення без явного занулення [62]. Elastic Net поєднує властивості L1 і L2 та застосовується у випадках великої кількості корельованих ознак, коли потрібні одночасно стабілізація параметрів і відбір інформативних предикторів [63]. У радіомічних дослідженнях комп'ютерної томографії легень логістична регресія з регуляризацією широко застосовується для побудови прогностичних і діагностичних моделей. Зокрема, у роботах показано, що використання LASSO-регуляризованої ЛР дозволяє ефективно відібрати інформативні радіомічні ознаки та побудувати номограми для прогнозу метастазів у лімфатичні вузли при ранній аденокарциномі легень, забезпечуючи стабільну дискримінаційну здатність і задовільне калібрування у зовнішніх вибірках [64]. В іншому дослідженні підхід застосовували для неінвазивного передбачення мутаційного статусу гена епідермального фактору росту за даними КТ, де поєднання радіомічних та клінічних змінних приводило до істотного підвищення точності моделі порівняно з використанням лише клінічних характеристик [65]. У масштабних скринінгових дослідженнях злоякісності легневих вузлів ЛР розглядалася як базова та інтерпретована модель і порівнювалася з методами на основі ансамблевих дерев, демонструючи зіставні результати точності за умови коректної перехресної валідації (cross-validation) та оцінки калібрування прогнозованих ризиків [66]. У літературі зустрічаються інші підходи, такі як метод опорних векторів (SVM), k-NN, наївний баєсів класифікатор, лінійні дискримінантні моделі, а також їх поєднання з різними стратегіями відбору ознак та балансування класів. Практична роль цих методів часто полягає в тому, що вони пропонують альтернативні припущення щодо характеру розмежування класів і можуть працювати конкурентно на невеликих вибірках [59,67]. Ансамблеві класифікатори ґрунтуються на поєднанні прогнозів кількох навчених моделей з метою підвищення узагальнювальної здатності та точності класифікації.

Поширеним прикладом ансамблевого класифікатора є Random Forest (RF) у якому фінальний результат отримується шляхом агрегації прогнозів великої кількості дерев рішень, збудованих на різних випадкових підвибірках даних і ознак. Підхід є практично цінним, оскільки добре моделює нелінійні залежності та взаємодії, стійкий до шуму та надлишковості ознак, може надавати оцінки важливості ознак. Основні методи регуляризації передбачають, зокрема, обмеження глибини дерева та встановлення мінімальної кількості зразків, необхідних для виконання поділу у вузлі дерева рішень [68]. Приклад застосування Random Forest продемонстровано у дослідженні [69], де використовується велика кількість кількісних характеристик, отриманих із комп'ютерної томографії, для виявлення певних генетичних змін у легеневій аденокарциномі. Продemonстровано AUC 0.802 для моделі на 94 радіомічних ознаках та покращення до 0.828 після додавання клінічних змінних (стать і анамнез куріння тощо). Це є типовим прикладом того, що ансамблеві дерева можуть ефективно працювати з високовимірними радіомічними даними, а інтеграція з клінічними предикторами часто додає стабільності й клінічній змістовності. Інший метод, градієнтний бустинг будує ансамбль дерев послідовно, де кожне наступне дерево навчається компенсувати помилки попередніх, оптимізуючи задану функцію втрат у термінах градієнтного спуску. Така модель машинного навчання часто дає дуже високу точність на табличних даних, але потребує ретельного налаштування (глибина дерев, learning rate, регуляризація, subsampling) [70]. XGBoost є популярною реалізацією градієнтного бустингу з додатковими регуляризаційними механізмами та оптимізацією, що зробило його поширеним у задачах медичної аналітики, коли є табличні ознаки та обмежені обсяги даних [71,72].

1.5. Методи глибокого навчання для аналізу КТ-зображень

Глибоке навчання — це підхід у машинному навчанні, за якого модель вчиться будувати представлення даних у кілька послідовних рівнів, де кожен наступний перетворює попередній на більш абстрактний опис [73]. Завдяки такій

багат шаровості модель може автоматично виділяти ознаки з сирих даних і використовувати їх для розпізнавання чи прогнозування [74]. Найпоширенішими моделями для глибокого навчання є нейронні мережі — багат шарові обчислювальні структури, які під час навчання підбирають параметри так, щоб мінімізувати помилку та узагальнювати закономірності даних [75]. У контексті задачі обробки зображень такі моделі неодноразово демонстрували високі результати та суттєве покращення якості порівняно з попередніми підходами, особливо після появи глибоких архітектур і великих навчальних наборів даних [75].

На відміну від підходів, де інформативні характеристики потрібно задавати вручну, нейронні мережі здатні навчатися виділяти корисні ознаки безпосередньо із зображень під час тренування, формуючи послідовність від простіших до складніших представлень. Це зменшує потребу в окремому етапі інженерії ознак і часто покращує узагальнення на нових даних. [76]. Практично важливо, що стабільність навчання та якість моделі можуть підвищуватися завдяки архітектурним прийомам нормалізації, які роблять оптимізацію більш керованою [77]. Для кращої роботи в умовах типових змін зображення (наприклад, зсуву, зміни масштабу, часткових деформацій) на практиці широко застосовують штучне розширення навчальних даних, а також спеціальні модулі, які навчаються компенсувати просторові перетворення [78,79].

Методи глибокого навчання вже давно широко використовуються для аналізу зображень комп'ютерної томографії органів грудної клітки. Систематичний огляд [80] узагальнює численні дослідження, в яких згорткові нейронні мережі демонструють високу ефективність у виявленні легеневих вузлів і підтримці діагностики раку легень на основі зображень комп'ютерної томографії. В іншому дослідженні [81] теж наведені подібні висновки – автори зазначають, що глибоке навчання стало домінуючою парадигмою в задачах аналізу зображень, зокрема сегментації та класифікації. Водночас як роботи [82,83], підкреслюють критичні обмеження подібних підходів, зокрема: залежність від значних обсягів якісно анотованих даних для навчання; недостатня для спеціалістів інтерпретованість прийнятих рішень від нейронних мереж (ефект «чорного

ящику»); чутливість моделей до міжсканерної та міжцентрової варіабельності зображень.

Зазначені обмеження стимулюють до пошуку методів, які б ефективно працювали на менших вибірках і забезпечували прозорішу інтерпретацію для експертів.

1.6 Альтернативний підхід на основі Bag-of-Visual-Words

Крім радіоміки, цікавою альтернативою глибокому навчанню є методологія «мішка візуальних слів» (Bag-of-Visual-Words, BoVW), що походить з комп'ютерного зору. Це підхід до подання зображень у вигляді фіксованого вектора ознак, який узагальнює локальні візуальні патерни за аналогією з моделлю «мішка слів» у текстовому аналізі. Метод передбачає виділення локальних дескрипторів у ключових точках або щільній сітці, квантування цих дескрипторів відносно заздалегідь сформованого словника, а далі агрегування інформації у вигляді гістограм частот появи візуальних слів (рисунок 1.2) [84]. Така репрезентація інваріантна до перестановки локальних елементів і є зручною для подальшої класифікації або кластеризації, хоча зазвичай втрачає явну просторову структуру зображення.

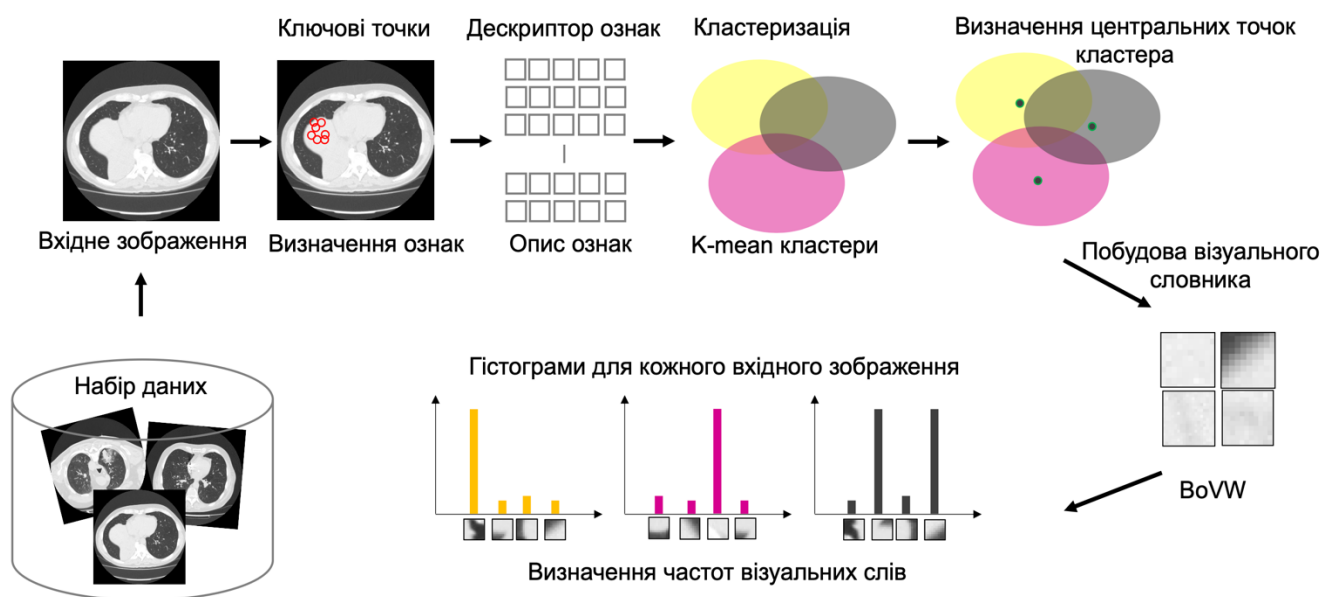


Рисунок 1.2 — Процес формування ознак методом Bag-of-Visual-Words

У дослідженні [84] використовували цю ідею до задач пошуку об'єктів на зображеннях та у відео. Локальні дескриптори квантуються в індекс візуальних слів, після чого кожне зображення описується гістограмою частот цих кодових слів. У медичній візуалізації такий підхід успішно використовували для класифікації зображень зі складною текстурою. Зокрема, для високороздільних КТ ОГК у роботі [85] автори розробили BoVW-алгоритми розпізнавання дифузних патологічних змін легеневої тканини, таких як консолидація, матове скло, емфізема і вузликові ураження. Використовуючи набір локальних статистичних ознак на основі значень КТ-інтенсивності, власних чисел Гессіана і словника візуальних слів, автори досягли точності розпізнавання 93,18%, що перевершило традиційні підходи з використанням глобальних ознак.

У дослідженні [86] своєю чергою запропоновано підхід з використанням BoVW-моделі на основі патчів зображення для класифікації патернів дифузних захворювань легень на КТ, де патчі — це локальні фрагменти зображення заданого розміру, з яких обчислюють дескриптори та формують візуальні слова. Подавши обчислені ознаки на вхід SVM-класифікатора під час навчання, автори продемонстрували точність у 93,9% на тестовій вибірці. Таким чином, підходи BoVW дозволяють описати складні текстури легеневої паренхіми у вигляді частотних гістограм локальних шаблонів. В огляді [87] наголошується на тому, що дискримінативну силу BoVW підходу можна підсилити за допомогою сегментації зображень, векторної квантизації, а також конструювання візуального словника.

У підходах BoVW для представлення зображень часто використовується схема зважування частот, запозичена з аналізу тексту – TF-IDF (Term Frequency - Inverse Document Frequency). Ідея полягає в тому, щоб збільшити вагу тих візуальних слів, які часто зустрічаються в документі (на зображенні), але рідкісні в межах усіх документів (зображень), і водночас зменшити вплив дуже поширених візуальних слів, які мало що розрізняють.

TF-IDF зважування гістограм візуальних слів дозволяє підсилити внесок унікальних, виразних ознак, що характерні для певного класу, і придушити вплив фонових патернів [84]. В контексті медичних зображень TF-IDF може бути

корисним, оскільки багато повторюваних структур, наприклад, нормальні анатомічні утворення, присутні в усіх зображеннях і не допомагають у класифікації, тоді як рідкісні патерни (специфічні ознаки ураження) є більш інформативними.

Висновок до розділу 1

Таким чином, огляд пов'язаних досліджень засвідчив, що, попри суттєвий прогрес у застосуванні методів глибокого навчання та радіоміки, зазначені підходи характеризуються низкою принципів обмежень.

Моделі глибокого навчання здатні забезпечувати високу діагностичну точність, однак потребують значних обсягів анотованих даних і, як правило, мають обмежену інтерпретованість.

Радіомічні ознаки, натомість, підвищують прозорість аналізу та можуть бути ефективними за умов відносно малих вибірок, проте залишаються чутливими до відмінностей у протоколах і параметрах сканування, а також до недостатньої стандартизації процедур виділення ознак, що ускладнює відтворюваність результатів.

Підходи, засновані на аналізі розподілу інтенсивностей і текстурних характеристик, також демонструють потенційну ефективність, однак часто потребують ретельної нормалізації даних та поєднання з іншими групами ознак для досягнення стабільної якості.

Нарешті, методи типу Bag-of-Visual-Words (BoVW) надають можливість узагальнювати складні локальні патерни у вигляді частотних представлень, проте їхня результативність істотно залежить від обраного типу дескрипторів, стратегії квантування, розміру словника та ступеня неоднорідності досліджуваної вибірки.

З огляду на зазначені обмеження, метою цієї роботи є розроблення підходу до кількісного аналізу КТ-зображень легень, який поєднує інтерпретованість процесу прийняття рішень для фахівців із можливістю ефективного застосування на відносно невеликих вибірках даних.

Запропонований метод інтегрує концепції радіоміки, аналізу текстурних характеристик і підходів типу Bag-of-Visual-Words, забезпечуючи подання унікальних локальних патернів зображення у вигляді частотних ознак із подальшим зважуванням за їхньою інформативністю. З теоретичної точки зору, таке представлення створює передумови для підвищення розпізнавання різних типів легневих уражень.

РОЗДІЛ 2

СЕЛЕКЦІЯ ІНТЕНСИВНОСТЕЙ ПІКСЕЛІВ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ

У цьому розділі представлено опис походження даних КТ-зображень легень та їх підготовку для проведення досліджень. Окрему увагу приділено етапам попередньої обробки зображень, зокрема сегментації регіонів з наявними патологічними змінами у тканинах з метою формування областей інтересу, виділенню режиму легеневого вікна, його нормалізації до 256 градацій сірого. На основі отриманих областей інтересу запропоновано метод фільтрації фонових й найменш інформативних пікселів. Отримані зразки трансформувалися у формат частотних карт інтенсивностей, де яскравість пікселя відображає частоту появи відповідного рівня сірого в усьому наборі даних.

Результати цього етапу створюють методологічне підґрунтя для подальшого застосування оброблених вихідних зображень для пошуку локальних, специфічних структур, які слугуватимуть ознаками для класифікації.

2.1 Набір даних КТ-зображень легень

У даній роботі використано клінічний набір даних зображень комп'ютерної томографії легень з наявними ураженнями ракових пухлин. Дані надані Національним науковим центром фтизіатрії, пульмонології та алергології імені Ф.Г. Яновського НАМН України [95].

Кожна серія знімків містить дані зрізів комп'ютерної томографії 17 пацієнтів з характеристиками щільності тканин для їх гістологічно та імуногістологічно підтверджених типів новоутворень легень.

КТ-знімки за типами патології були агреговані у аналітичні групи, на основі яких формувався розподіл по класах (табл. 2.1, [95]).

Агрегацію типів в укрупнені аналітичні групи було здійснено за рекомендацією лікаря.

Таблиця 2.1 — Опис набору клінічних даних

ID	К-сть зрізів	К-сть масок	Діагноз	Мітка класу
1_0020	472	25	Дрібноклітинний рак	0 - small
1_0021	492	70	Дрібноклітинний рак	0 - small
1_0000	446	44	Низькодиференційована аденокарцинома	1 - adeno
1_0002	391	19	Інвазивна аденокарцинома змішаної будови-муцинозної та колоїдної	1 - adeno
1_0005	436	23	Плеоморфна веретено клітинна та гігантоклітинна карцинома	1 - adeno
1_0009	523	18	Протокова карцинома слинної залози (аденокарцинома)	1 - adeno
1_0010	531	22	Низькодиференційована плеоморфна веретено клітинна карцинома	1 - adeno
1_0011	461	7	Аденоїд-кістозна карцинома	1 - adeno
1_0012	638	34	Аденокарцинома колоїдного типу	1 - adeno
1_0001	245	25	Бронхогенна диференційована плоскоклітинна карцинома	2 - flatO
1_0006	416	30	Бронхогенна плоскоклітинна карцинома з ороговінням	2 - flatO
1_0014	451	39	Плоскоклітинна карцинома із ороговінням	2 - flatO
1_0003	80	48	Плоскоклітинна карцинома без ороговіння	3 - flatNO
1_0004	456	26	Первинна плоскоклітинна карцинома	3 - flatNO
1_0015	441	14	Плоскоклітинна карцинома без ороговіння	3 - flatNO
1_0008	411	23	Карциноїд (доброякісна пухлина)	4 - carci
1_0013	582	36	Атиповий карциноїд (доброякісна пухлина)	4 - carci

Кодування аналітичних груп для формування міток класів здійснювалося за наступним принципом:

- "0 - small" відповідає діагнозу дрібноклітинного раку;
- "1 - adeno" - аденокарцинома;
- "2 - flatO" - плоскоклітинний рак з ороговінням;
- "3 - flatNO" - плоскоклітинний рак без ороговіння;
- "4 - carci" - карциноїд.

Варто відзначити, що карциноїд відноситься до нейроендокринних пухлин низького/помірного ступеня злоякісності і використовується лише як представник новоутворень легень для збільшення вибірки [88,89].

Визначення областей інтересу на наборі КТ-зображень здійснювалася шляхом експертної ручної розмітки, проведеної за участю провідних медичних фахівців у галузі КТ-діагностики легеневої патології для забезпечення високої клінічної достовірності дослідження. Основною метою було захопити якомога

більшу площу зони ураження. Потрапляння інших частин паренхіми не було критичним у цьому випадку, оскільки вплив буде усунено шляхом застосування алгоритму фільтрації, процес детально описано далі.

В результаті були сформовані бінарні маски, які надалі накладаються на оригінальні КТ-зрізи з ураженнями та дозволяють відсікати всю нерелевантну інформацію, зберігаючи лише ділянки патологічного процесу.

Приклади зображень з відміченими областями інтересу наведено на рисунках 2.1-2.3.

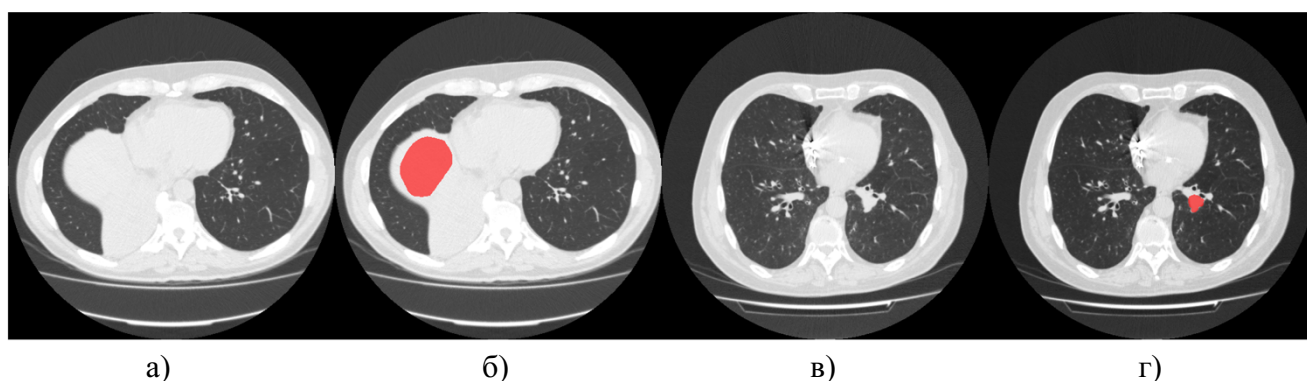


Рисунок 2.1 — Приклад зображення з виділеною зоною ураження: а - оригінальне зображення з наявним дрібноклітинним раком; б - виділена область ураження дрібноклітинного раку; в - оригінальне зображення з наявною аденокарциномою; г - виділена область ураження аденокарциноми

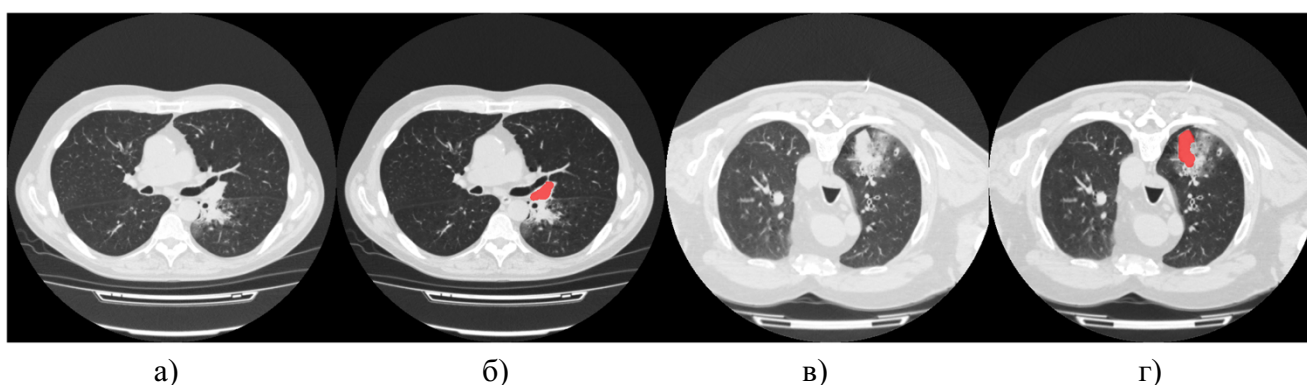
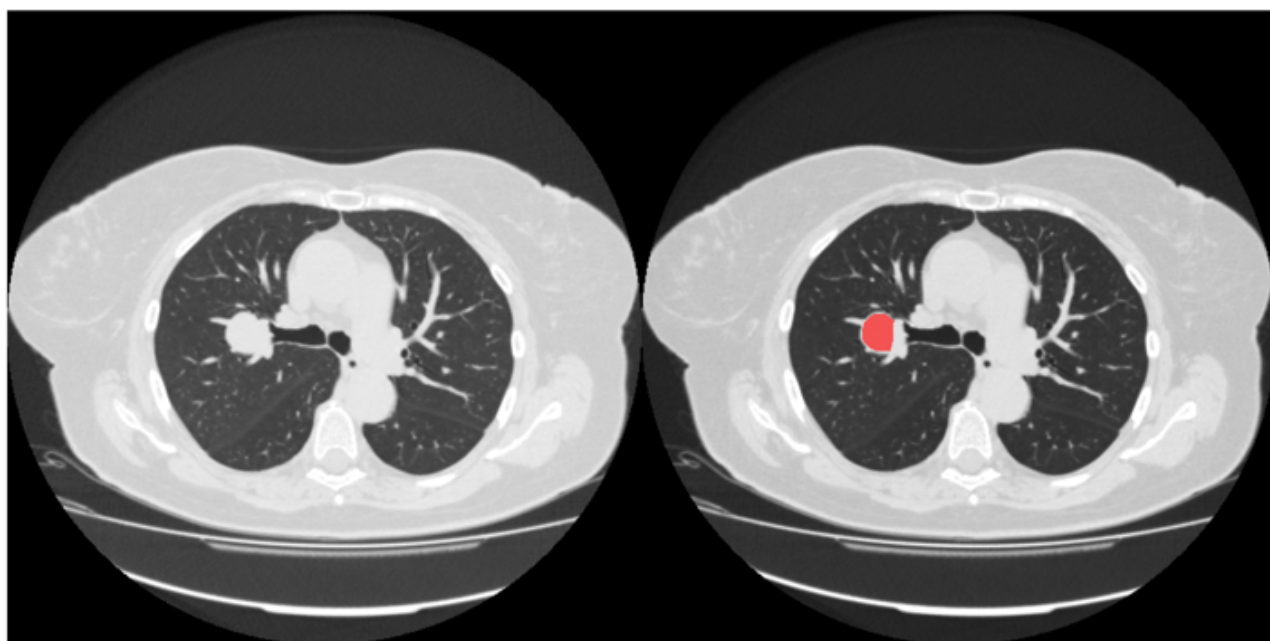


Рисунок 2.2 — Приклад зображення з виділеною зоною ураження: а - оригінальне зображення з наявним плоскоклітинним раком з ороговінням; б - виділена область ураження плоскоклітинного раку з ороговінням; в - оригінальне зображення з наявним плоскоклітинним раком без ороговінням; г - виділена область ураження плоскоклітинного раку без ороговіння



а)

б)

Рисунок 2.3 — Приклад зображення з виділеною зоною ураження: а - оригінальне зображення з наявним карциноїдом; б - виділена область ураження карциноїда;

Оригінальні зображення у форматі DICOM, розмірністю 512x512 пікселів були переведені у режим легеневого відображення (рівень -600 HU, ширина вікна 1500 HU, у значеннях в одиницях Гаунсфілда), що звело діапазон щільностей з (-2000 HU; 1000 HU) до (-1350 HU; 150 HU) [95]. Застосування легеневого вікна мало на меті оптимізувати діапазон вхідних даних, оскільки під час візуалізації та подальшого аналізу здійснюється його нормалізація. Одним із ключових завдань використання такого вікна є покращення локалізації патологічних змін. Налаштування параметрів контрастності та щільності забезпечує виразніше відображення структур у межах легеневої паренхіми, розширюючи діапазон значень, характерних для тканин легень. Крім того, приведення даних до легеневого вікна сприяє зменшенню впливу апаратних шумів, оскільки модифікація діапазону щільностей дозволяє відфільтрувати частину небажаних сигналів, що не є інформативними для подальшого аналізу. Усі КТ-зрізи були конвертовані у формат TIFF з подальшою нормалізацією діапазонів градацій сірого до рівня від 0 до 255 за формулою (2.1), що в результаті дало 8-бітне сірошкальне зображення.

$$I_{norm}(x, y) = \frac{I(x, y) - I_{min}}{I_{max} - I_{min}} \cdot 255 \quad (2.1)$$

де $I_{norm}(x, y)$ – значення інтенсивності пікселя за координатами (x, y) після нормалізації.

Нормалізовані зображення області інтересу, отримані з оригінальних КТ-зображень кожного типу новоутворень представлені на рисунку 2.4.

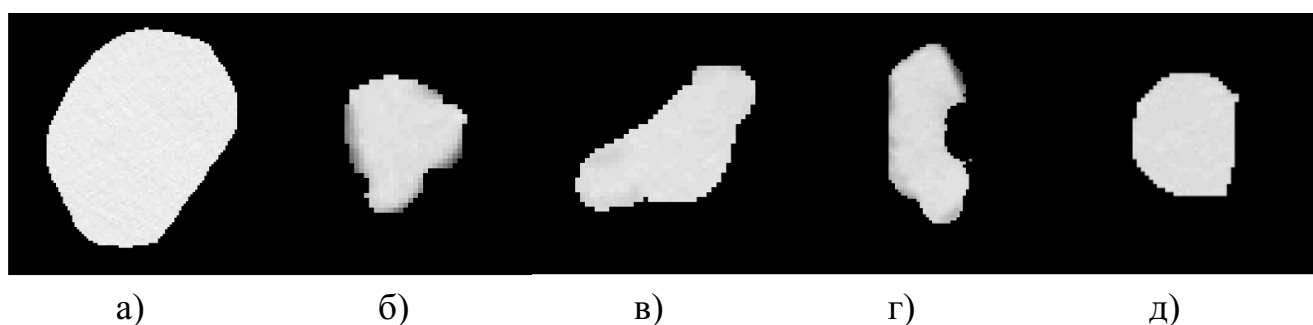


Рисунок 2.4 — Нормалізоване зображення виділеної області ураження: а - дрібноклітинний рак; б - аденокарцинома; в - плоскоклітинний рак з ороговінням; г - плоскоклітинний рак без ороговіння; д - карциноїд

Наявність невеликої та нерівномірно представлені за діагнозами вибірки КТ-зображень створює типову для медичної візуалізації проблему: подальші етапи аналізу, зокрема пошук локальних структурних патернів, можуть непропорційно підлаштовуватися під домінуючі класи або під випадкові флуктуації, що не відтворюються в інших підвибірках. У таких умовах доцільним є попередній крок, який статистично обмежує простір інтенсивностей і тим самим зменшує розмірність та варіативність вхідних даних перед виділенням структурних ознак. Комплексні дослідження [90,91] зазначають, що без спеціальних механізмів компенсації дисбаланс, навчання на подібних вибірках призводить до систематичного зміщення критеріїв відбору ознак і погіршення узагальнювальної здатності моделей.

Окремим джерелом варіативності є характеристики інтенсивностей пікселів на КТ-зображеннях, що залежать від параметрів сканування, реконструкції та

обраного протоколу попередньої обробки. У радіоміці це розглядається як ключовий чинник відтворюваності, тому наголошується про необхідність чітко визначати та контролювати етапи, пов'язані з шкалою інтенсивності й формуванням похідних кількісних описів [92]. Емпіричні дослідження також показують, що такі налаштування, як дискретизація рівнів сірого та пов'язані перетворення, можуть істотно впливати на стабільність радіомічних ознак у різних умовах [93], а підходи нормалізації, пов'язані з розмірами пікселів і кількістю рівнів сірого, є суттєвими для порівнюваності кількісних характеристик у дослідженнях раку легень [94].

Отже, введення контрольованого представлення інтенсивностей перед виділенням структурних ознак є обґрунтованим з позицій відтворюваності та зменшення некорисної варіативності.

У роботі запропоновано підхід до попередньої фільтрації зображень усієї вибірки, який орієнтований на статистично репрезентативні прояви інтенсивностей в області інтересу, із балансуванням внеску класів. Концептуально така підготовка даних виконує дві функції: 1) зменшує розмірність та варіативність простору інтенсивностей пікселів, знижуючи ризик того, що рідкісні або протокольно зумовлені інтенсивності визначатимуть подальші структурні висновки; 2) забезпечує більш справедливі умови для класів із різною кількістю пацієнтів/зображень, аби потенційно інформативні прояви міноритарних діагнозів не були поглинуті статистикою мажоритарних класів. Кожен рівень інтенсивності у градаціях сірого на КТ-знімках оцінюється з погляду його внеску в розрізнення класів на основі всього навчального набору. На цій підставі визначається глобальна вага класової значущості для відповідного рівня яскравості. Інтенсивності, що мають низьку вагу, тобто трапляються приблизно з однаковою частотою в усіх класах і не містять виразної дискримінаційної інформації, вилучаються з подальшого аналізу. Натомість інтенсивності з високою класовою вагою зберігаються, оскільки саме вони забезпечують найбільший внесок у розпізнавання [95]. Схематично підхід до фільтрації зображень показано на рисунку 2.5.

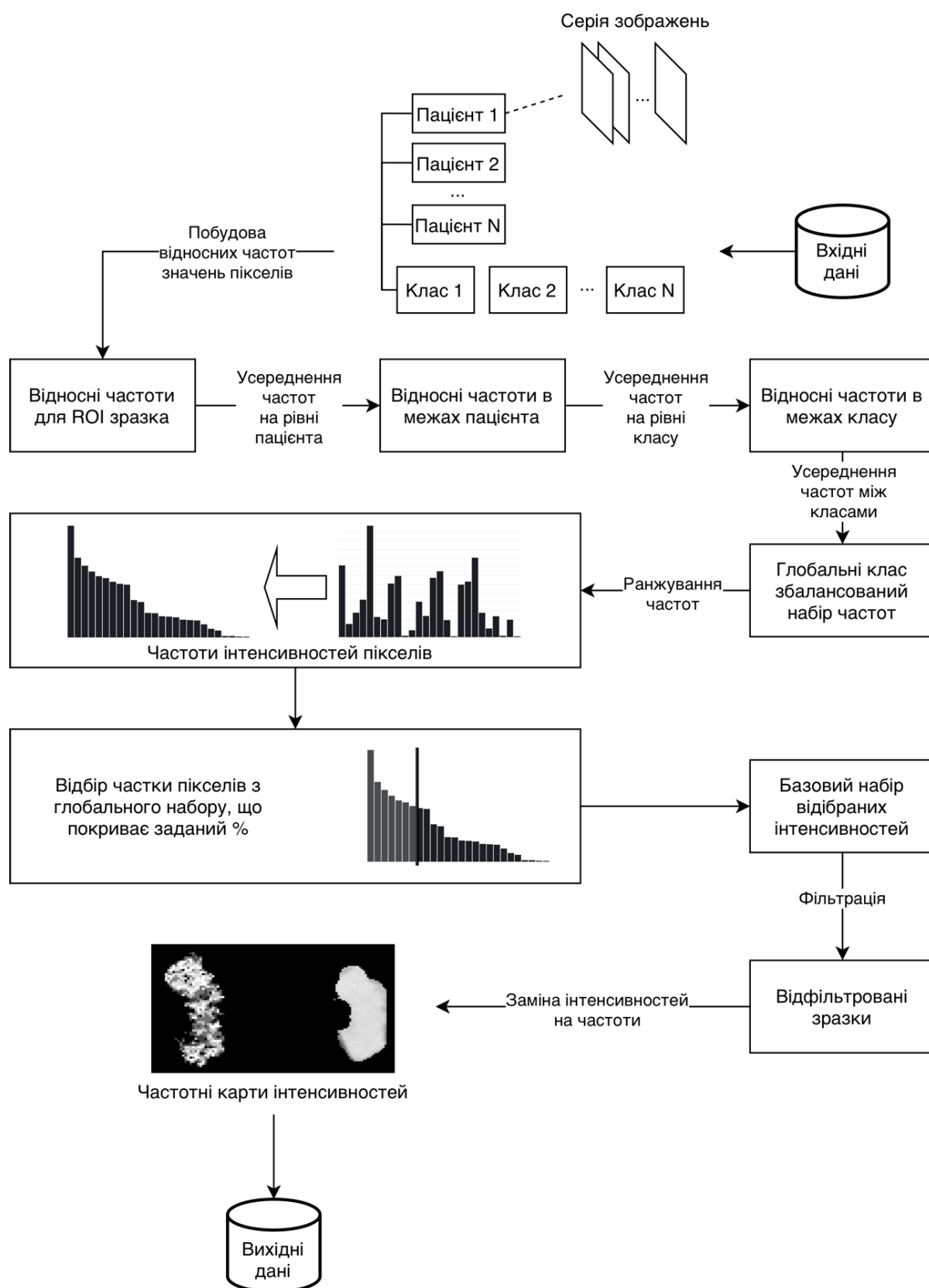


Рисунок 2.5 — Кроки фільтрації зображень

Аналогічно методу Оцу, який у задачі двокласової сегментації автоматично визначає поріг інтенсивності шляхом максимізації міжкласової дисперсії,

запропонований підхід узагальнює ідею порогового розділення інтенсивностей у багатокласовому випадку. Таким чином формується процедура клас-балансного відбору інтенсивностей, що дає змогу виокремити інформативні діапазони яскравості та на глобальному рівні збалансувати внесок різних класів, водночас усуваючи нерелевантні ділянки зображення [95].

2.2 Попередня обробка зображень та побудова локальних частот

Зображення пацієнтів з конкретними діагнозами новоутворень у легеневій паренхімі були агреговані у більш крупні аналітичні групи (класи за типами патології). Саме у такому вигляді, на цьому найвищому рівні укрупнення, дані будуть використовуватись для фільтрації найбільш інформативних пікселів у всьому наборі даних загалом.

Кожному знімку було привласнено унікальний ідентифікатор пацієнта та відповідного класу, що унеможлиблює перемішування вибірок між групами. Зображення опрацьовуються в сірошкальному форматі, тому всі подальші обчислення виконуються на основі однієї скалярної інтенсивності. Щоб зменшити варіативність даних та стабілізувати оцінки на малих вибірках, а також мінімізувати вплив артефактів зображення, було здійснено операцію рівномірного квантування інтенсивностей. Формально операцію квантування можна описати наступним чином, нехай L — кількість рівнів після квантування (деяке ціле число від 2 до 255), тоді $s_q = (L - 1)/255$ — коефіцієнт масштабування; $I_x(u) \in \{0, \dots, 255\}$ — інтенсивність пікселя з координатами u на зображенні x . Тоді, інтенсивність після квантування задається формулою [95]:

$$I_x^{(Q)}(u) = \frac{1}{s_q} \text{round}(s_q I_x(u)), \quad (2.2)$$

що стискає близькі значення до спільних інтенсивностей (бінів, якщо дивитись з точки зору гістограми розподілу градацій сірого) [95]. Результат роботи

операції квантування зі зменшенням градації сірого з 256 до 64-х, 32-х та 16-ти відтінків для одного і того ж зразка пацієнта з діагнозом дрібноклітинний рак продемонстровано на рисунку 2.6.

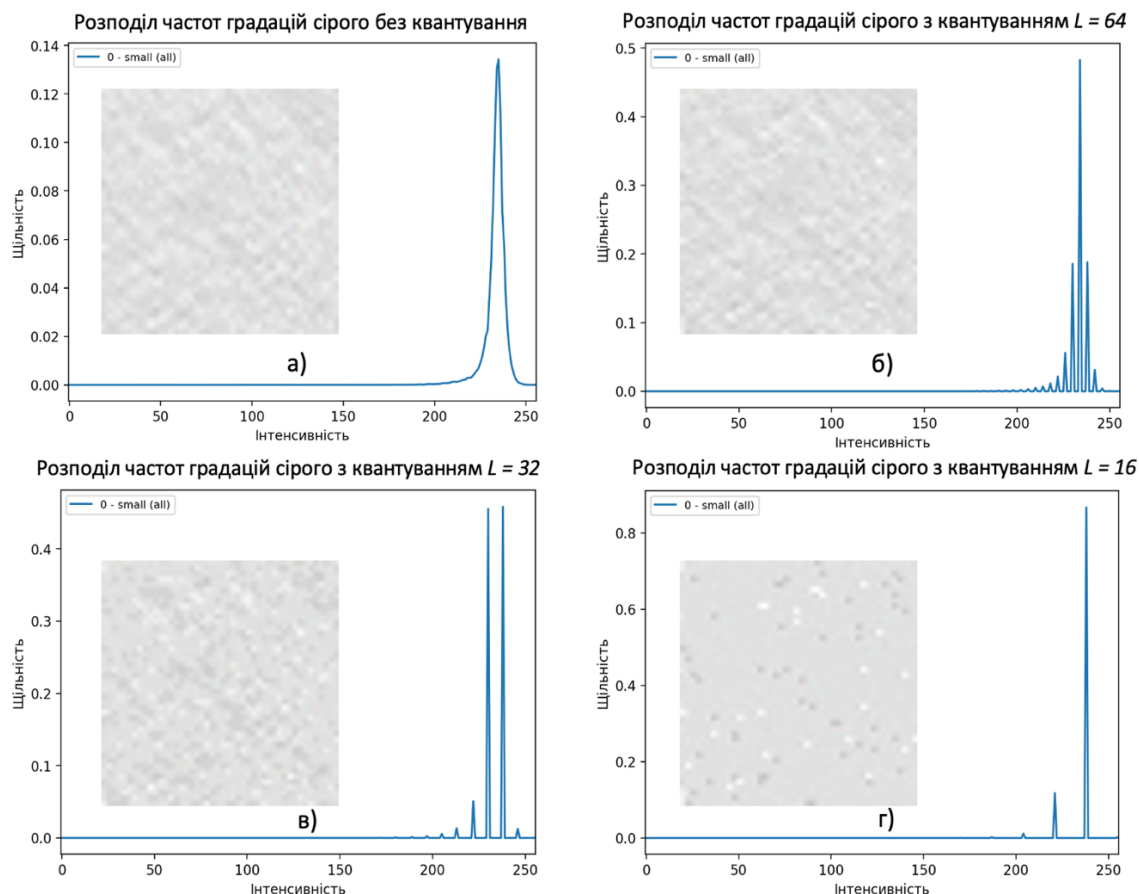


Рисунок 2.6 — Фрагмент зразка та його розподіл інтенсивностей при різному ступені квантування: а) зразок та розподіл без квантування; б) зразок та розподіл з квантуванням до 64-х відтінків; в) зразок та розподіл з квантуванням до 32-х відтінків; г) зразок та розподіл з квантуванням до 16-х відтінків

Можна спостерігати, що при різних рівнях $L=64$ (рисунок 2.6б), $L=32$ (рисунок 2.6в) та $L=16$ (рисунок 2.6г), відбувається послідовне зменшення дрібних флуктуацій між сусідніми рівнями, а також зменшення кількості унікальних інтенсивностей пікселів (бінів) порівняно з оригінальним неквантованим зразком (рисунок 2.6а).

Проте, занадто сильне квантування, особливо при малих значеннях L , наприклад 16, може утворювати вузькі піки й розмивати видимість тонких

структур, отже у цій процедурі необхідний певний компроміс між збереженням деталей та їх згладжуванням. Варто зазначити, що оптимальний рівень квантування обирався на основі результатів класифікації, що описано у подальших розділах.

2.2.1 Розподіл інтенсивностей пікселів на рівні окремого КТ-зрізу

На наступному кроці для кожного зображення будується 256-бінова гістограма $h_x(i)$, де $i \in \{0, \dots, 255\}$ — індекс біну, а $h_x(i)$ — кількість пікселів із інтенсивністю i [95].

Оскільки розміри виділеної області інтересу на різних зображеннях відрізняють, так само як і, потенційно, частоти інтенсивностей пікселів, порівнювати абсолютні їх значення недоцільно.

Щоб зробити оцінки між зображеннями співставними, вводиться нормований локальний розподіл частот $\chi = \{1, \dots, 255\}$, де значення фону, тобто $i = 0$, ігнорується, що запобігає домінуванню тривіально великих темних областей (фон поза межами області інтересу) [95].

Визначимо кількість інтенсивностей, що входить у нормування:

$$N_x = \sum_{i \in \chi} h_x(i) \quad (2.3)$$

а також частоту інтенсивностей i для зображення x :

$$t_x(i) = \begin{cases} \frac{h_x(i)}{N_x}, & i \in \chi \\ 0, & i = 0 \text{ і фон ігнорується} \end{cases} \quad (2.4)$$

Графічно частоти інтенсивностей пікселів можна зобразити у вигляді гістограми їх розподілу. Приклад подібного розподілу, отриманого з оригінальних

зображень комп'ютерної томографії кожного типу новоутворень (усі зображення класів) представлені на рисунку 2.7.

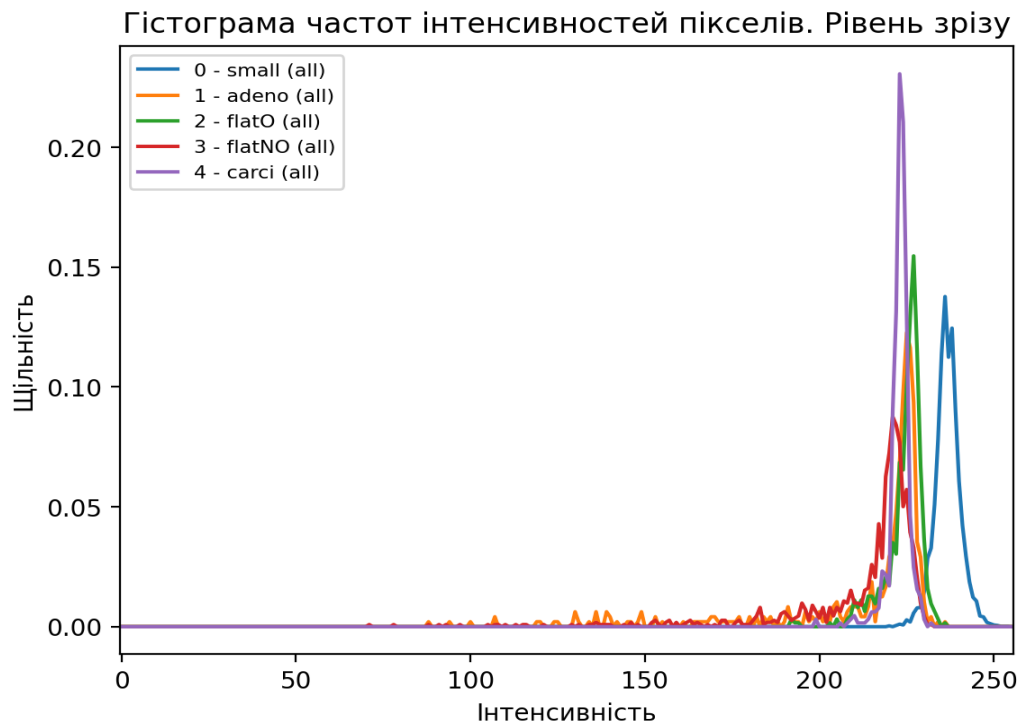


Рисунок 2.7 — Гістограма розподілу частот інтенсивностей пікселів для окремих КТ-зрізів пацієнтів різних класів

2.2.2 Агрегація інтенсивностей пікселів до рівня пацієнта

Кількість КТ-зрізів з областями інтересу, що відповідають певному діагнозу, може суттєво різнитися між пацієнтами, що неминує впливає на представленість відповідного класу. Щоб уникнути ситуації, коли один пацієнт непропорційно домінує у статистичній оцінці, частки інтенсивностей спочатку усереднюються всередині множини КТ-зображень кожного пацієнта. Нехай $\chi_{c,j}$ — множина зображень класу c , що належить пацієнту j , тоді середнє визначається як:

$$t^{(c,j)} = \frac{1}{|\chi_{c,j}|} \sum_{x \in \chi_{c,j}} t_x \quad (2.5)$$

де $|\chi_{c,j}|$ — це кількість знімків даного пацієнта [95].

Приклад агрегованого розподілу частот інтенсивностей усіх зразків пацієнтів, показано на рисунку 2.8.

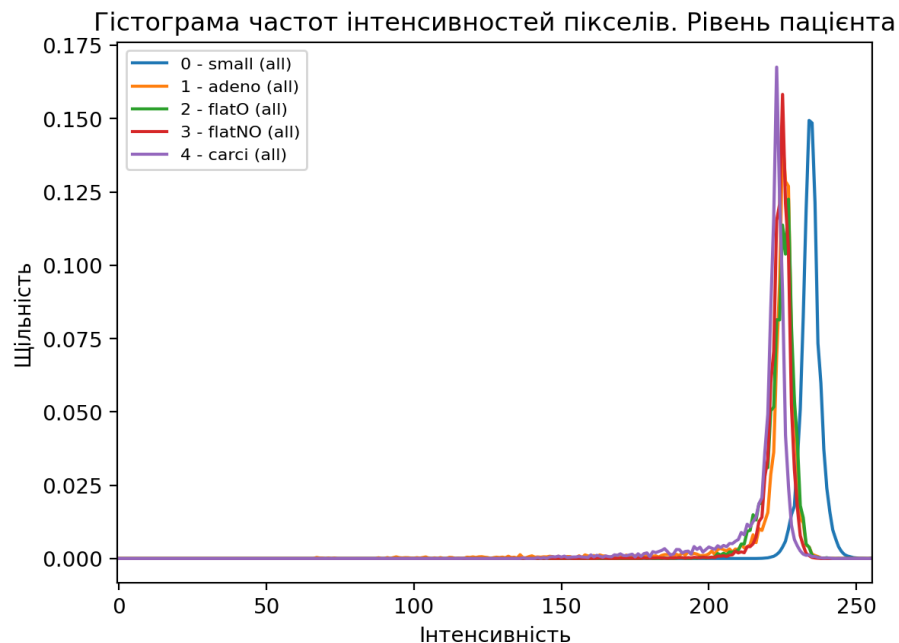


Рисунок 2.8 — Середні гістограми розподілу частот інтенсивностей пікселів по всіх КТ-зрізах в межах пацієнта

2.2.3 Агрегація інтенсивностей пікселів до рівня класу

Після цього така сама операція усереднення виконується на рівні пацієнтів у межах одного класу. Репрезентативний розподіл для класу отримується шляхом усереднення середніх розподілів інтенсивностей, обчислених для кожного пацієнта цього класу. Нехай P_c — множина пацієнтів у класі c , тоді:

$$t^{(c)} = \frac{1}{|P_c|} \sum_{j \in P_c} t^{(c,j)} \quad (2.6)$$

Вводиться повторне нормування і весь вектор масштабується так, щоб $\sum_{i=0}^{255} t^{(c)}(i) = 1$. Отримані значення є масштабно-інваріантним описом типового розподілу інтенсивностей у класі, який є нечутливим до внутрішньокласових дисбалансів [95]. У підсумку одержуємо для кожного класу профіль

інтенсивностей (надалі під терміном профіль інтенсивностей мається на увазі вектор значень), у якому всі пацієнти мають однакову вагу, приклад наведено на рисунку 2.9.

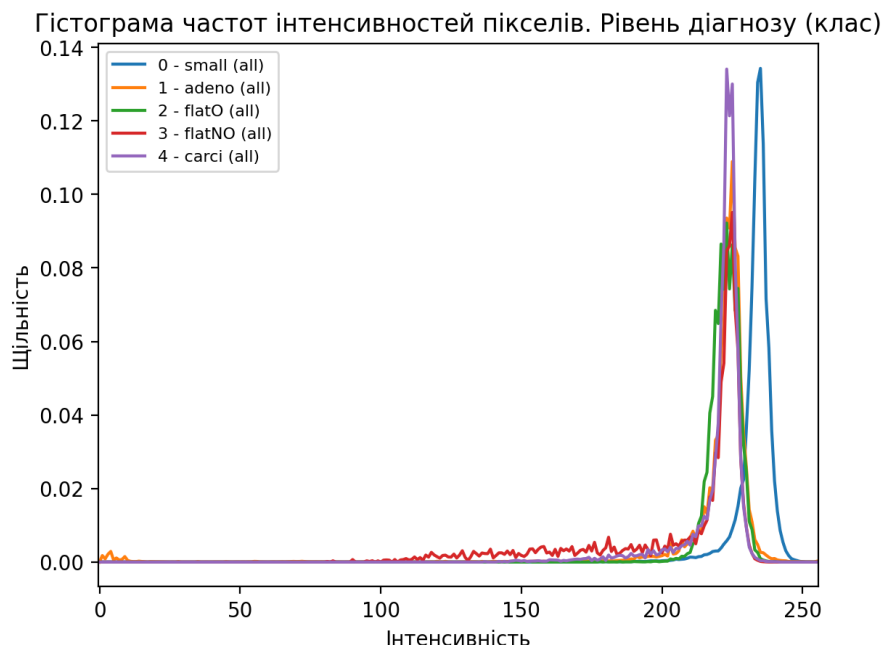


Рисунок 2.9 — Середні гістограми розподілу частот інтенсивностей пікселів по всіх знімках пацієнтів в межах класу

2.3 Відбір найбільш представлених інтенсивностей пікселів

Ключовим агрегованим показником є глобальний профіль інтенсивностей, сформований як усереднення за класами. Ця операція спрямована на отримання узагальненого частотного профілю рівнів сірого, який є інваріантним до дисбалансу зразків між класами і водночас зберігає репрезентативність для кожної патологічної групи. Така агрегація дозволяє інтерпретувати отримане значення для кожного біну інтенсивності як оцінку його масової частки за умови рівноправного внеску всіх класів, тобто як клас-збалансовану характеристику поширеності відповідного рівня сірого в загальній сукупності даних. Формально, нехай C — множина всіх класів, $|C|$ — кількість класів, $s_{global}(i)$ — глобальна збалансована оцінка між класами важливості бінів i . Визначається як [95]:

$$s_{global}(i) = \frac{1}{|C|} \sum_{c \in C} t^{(c)}(i), \sum_i s_{global}(i) = 1 \quad (2.7)$$

s_{global} відображає сумарну репрезентативність певного біна інтенсивності за умови рівноправного внеску кожного діагнозу. Вводиться поняття маси пікселів, під якою слід розуміти інтегральну статистичну вагу (відносну частоту) конкретного рівня яскравості, яка визначається як середнє значення його поширеності серед усіх досліджуваних патологічних груп. Зростання маси певної інтенсивності відображає те, що цей рівень сірого стабільно присутній у середніх розподілах кількох класів і, відповідно, є відтворюваною складовою зображень для різних груп. Натомість низька маса відповідає інтенсивностям, які трапляються рідко, є переважно пов'язаними з фоновими або шумовими компонентами. Проміжні значення інтерпретуються як частково узгоджені інтенсивності, що присутні в даних, однак не є домінантними або відтворюваними одночасно для більшості класів. Важливо, що така процедура принципово відрізняється від прямого усереднення за всіма зразками, оскільки запобігає ситуації, коли класи з більшим обсягом даних непропорційно визначають глобальну картину інтенсивностей.

Отриманий профіль уже містить узагальнену інформацію про інтенсивності, які найбільш притаманні кожному з діагностичних класів, але в той же час він усе ще описує повний спектр рівнів сірого, включно з тими, що можуть містити меншу диференційну цінність. Без додаткового відбору наступні кроки, що мають на меті пошук структурних особливостей на зразках неминуче почнуть відображати не лише структурно значущі компоненти, але й другорядні коливання інтенсивностей, які потенційно погіршують відтворюваність ознак і підсилюють чутливість методу до варіабельності. Поряд із цим, збереження нерелевантних елементів зображення підвищує загальну ресурсомісткість обчислень, що негативно впливає на швидкодію алгоритмів аналізу структурних патернів. Тому, виникає потреба сконцентрувати подальший аналіз на обмеженій множині рівнів сірого, які є найбільш репрезентативними.

Запропонований наступний етап селекції реалізує саме цю потребу, перетворюючи глобальний профіль на компактний словник інтенсивностей, на який надалі спиратимуться всі просторові перетворення. Метою є відібрати якомога менше інтенсивностей (бінів), водночас, зберегти задану частку глобальної маси пікселів. Іншими словами це відповідь на питання: як обрати мінімальний набір рівнів сірого, який зберігає потрібну частку змісту. Нехай $\alpha \in (0,1]$ — деяка бажана частка маси, $S_{base} \subseteq \{0, \dots, N\}$ — шуканий набір індексів бінів. Базовий набір визначається як найменша за частотою множина, для якої виконується нерівність [95]:

$$\sum_{i \in S_{base}} s_{global}(i) \geq \alpha \quad (2.8)$$

Практично виконується процедура жадібного алгоритму, де інтенсивності пікселів впорядковуються за спаданням $s_{base}(i)$ відносно їх частоти та послідовно додаються до моменту, поки кумулятивна сума не перевищуватиме встановлений поріг α . Після визначення найбільш представлених інтенсивностей на глобальному наборі, зображення фільтрується маскою, тобто пікселі, значення яких належать множині S_{base} , зберігаються, решта ігнорується. Фактично це є операцією фільтрації, де знімок очищується від усього, що є малозначущим для задачі. Якщо визначити операцію фільтрації як $F_x(u)$, тоді [95]:

$$F_x(u) = \begin{cases} I_x(u), & \text{якщо } I_x(u) \in S_{base} \\ 0, & \text{інакше} \end{cases} \quad (2.9)$$

Щоб інтерпретувати збережені пікселі й підкреслити їхню інформативність, за визначеним глобальним набором будується частотна карта інтенсивностей. Для цього використовується частотний профіль, вага біну з індексом i дорівнює глобальному лічильнику $L_i^{frequency} = H(i)$. На основі цього формується карта для зображення x , як $M_x(u) = L_{I_x(u)}^{frequency}$ для пікселів, інтенсивність яких належить до

S_{base} та $M_x(u) = 0$ поза межами набору. Фактично це є відображенням, наскільки часто певна інтенсивність трапляється у всьому наборі, чим частіше зустрічається бін, тим більша вага [95].

У підсумку процес можна подати як послідовність етапів: дані перетворюються в простіший формат для обробки, прибирається вплив фону та розміру, урівноважуються різниці за кількістю внесок зображень пацієнтів та класів, визначається, які рівні сірого трапляються найчастіше, відбираються найнеобхідніші з них. Частотна карта інтенсивностей додає пояснювальний шар та показує, де саме на зображенні присутні типові або, навпаки, рідкісні для всього набору значення пікселів [95].

2.4 Результат роботи клас-балансного фільтра

Щоб продемонструвати ефективність запропонованого алгоритму фільтрації було використано клінічний набір КТ-зображень легень із гістологічно підтвердженими типами раку (табл. 2.1). Обробка вхідних зображень виконувалась з використанням трьох різних параметрів налаштування. За фіксованих умов, при яких фон зображень ігнорується, квантування встановлене на $L=128$, розглянуто три пороги маси α [95]. Варто зазначити, що даний рівень квантування був обраний у ході проведених експериментів як найбільш оптимальний з точки зору редукції кількості градацій, що пришвидшує роботу алгоритму формування ознак та класифікації, при цьому зберігає достатню кількість інформації. Обґрунтування вибраного $L=128$ наведено у розділі 5. При встановленому параметрі $\alpha \approx 0.3$, що дає глобальне покриття 30.08% відбирається лише два біни у S_{base} . Під глобальним покриттям мається на увазі відсоток пікселів у всій сукупності КТ-зображень, які залишаються представленими (не відкидаються) після застосування запропонованої процедури відбору інтенсивностей.

Спостерігається, що така компресія інтенсивностей суттєво підсилює представленість класу карциноїда, для якого покриття сягає 47.23%, а також аденокарциноми — 35.83%. Для двох типів плоскоклітинного раку відповідні

значення становлять 32.69% та 31.98%. Водночас клас дрібноклітинного раку на цьому етапі практично не представлений, з покриттям лише 2.65%. При встановленому параметрі $\alpha \approx 0.5$, глобальне покриття сягає 50.82%, відбирається чотири біни. Покриття окремих класів при цьому також підвищується і становить 61.96% для аденокарциноми, 62.11% для плоскоклітинного раку з ороговінням, 52.37% для плоскоклітинного раку без ороговіння та 71.68% для карциноїда. Попри це, клас дрібноклітинного раку залишається недостатньо охопленим — 5.99%. При встановленому параметрі $\alpha \approx 0.7$ отримуємо вісім бінів у S_{base} і глобальне покриття 71.87%. Саме на цьому рівні спостерігається суттєве покращення міжкласового балансу: покриття дрібноклітинного раку зростає до 55.17%, аденокарциноми — до 78.25%, типів плоскоклітинного раку — до 82.86% та 61.79% відповідно, а карциноїда — до 81.26%. Узагальнені результати для кожної конфігурації параметрів наведено в табл. 2.2 [95].

Таблиця 2.2 — Параметри та покриття яскравості пікселів на глобальному наборі¹

Параметри			
α	0.3	0.5	0.7
К-ть бінів	2	4	8
Маса покриття, %	30,08	50,82	71,87
Покриття по класах, %			
0 - small	2,65	5,99	55,17
1 - adeno	35,83	61,96	78,25
2 - flatO	32,69	62,11	82,86
3 - flatNO	31,98	52,37	61,79
4 - carci	47,23	71,68	81,26

За встановленого значення параметра $\alpha \approx 0.7$ можна спостерігати, що додавання кількох найбільш масивних бінів дає змогу охопити характерні інтенсивності класів з відмінними профілями розподілу та фактично усунути основний міжкласовий дисбаланс. Таким чином, відсоток покриття для окремого класу визначається ступенем перетину його розподілу інтенсивностей із базовим набором: якщо моди розподілу класу припадають на найбільш представлені біни глобального профілю, відповідне покриття є високим [95]. Це наочно ілюструє

¹ Результати були відображені у матеріалах статті [95]

відмінності між інтенсивнісними профілями зображень різних класів. Метод забезпечує формування універсального правила фільтрації зображень для етапу інференсу, за якого мітка класу є апіорно невідомою, але гарантується досягнення щонайменше заданого рівня покриття інтенсивностей для кожного класу. Додатковою перевагою такого підходу є суттєве зменшення розмірності даних, що безпосередньо впливає на зниження обчислювальної складності подальшого аналізу. Відібрані інтенсивності застосовуються до вихідних зображень і зберігаються у вигляді частотних карт інтенсивностей, у яких значення кожного пікселя замінюється значенням ранжування відносно його глобальної частоти або маси відповідного рівня яскравості, за спаданням. Приклади вихідних зображень після фільтрації наведено на рисунок 2.10-2.12

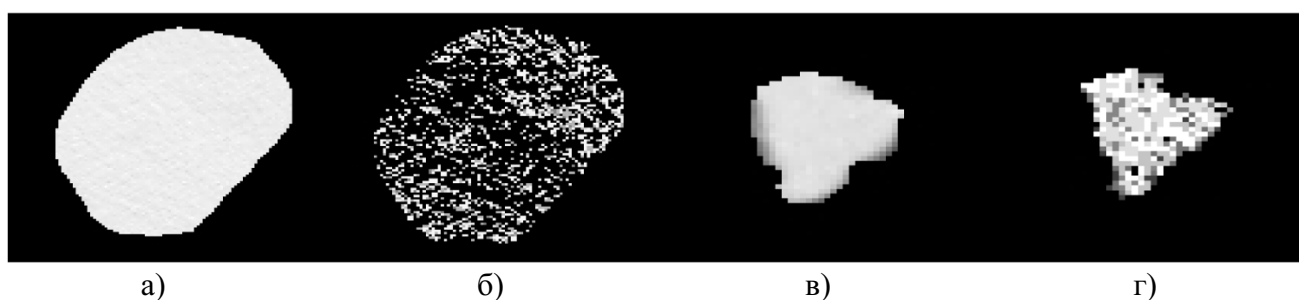


Рисунок 2.10 — Трансформовані зображення після етапу фільтрації: а - зріз дрібноклітинного раку; б - фільтрований зріз дрібноклітинного раку; в - зріз аденокарцинома; г - фільтрований зріз аденокарциноми

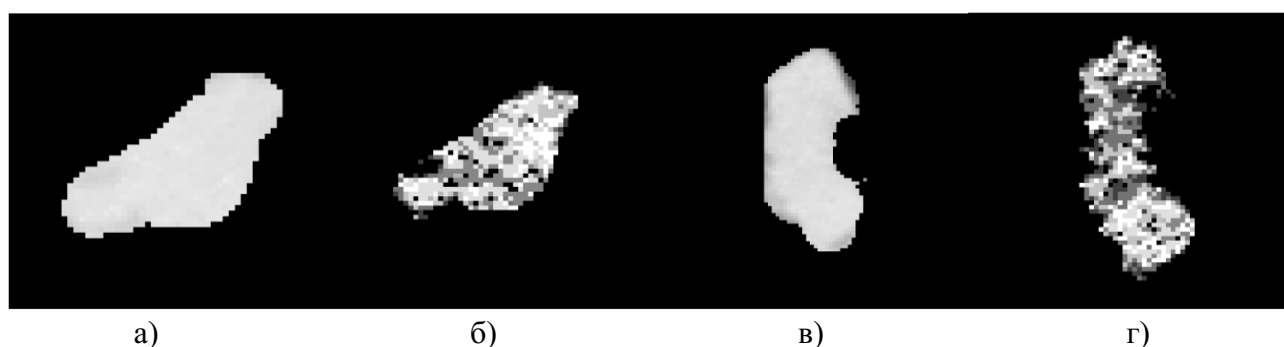


Рисунок 2.11 — Трансформовані зображення після етапу фільтрації: а - зріз плоскоклітинного раку з ороговінням; б - фільтрований зріз плоскоклітинного раку з ороговінням; в - зріз плоскоклітинним раком без ороговіння; г - фільтрований зріз плоскоклітинного раку без ороговіння

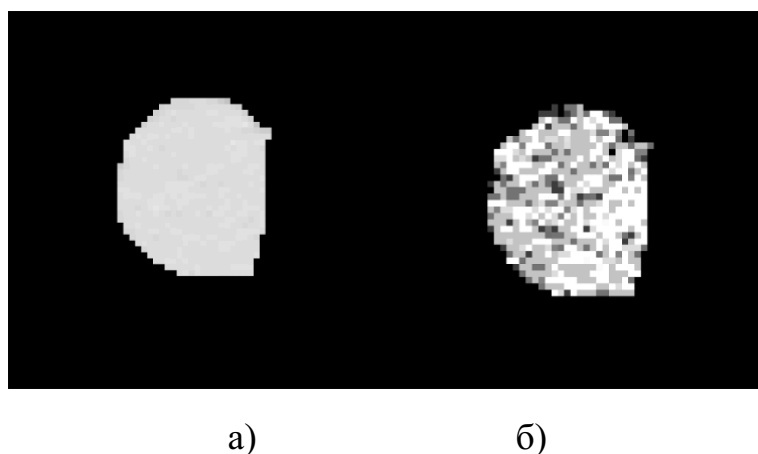


Рисунок 2.12 — Трансформовані зображення після етапу фільтрації: а - зріз карциноїда; б - фільтрований зріз карциноїда

Інтерпретація результатів у вигляді частотних карт була обрана з двох причин. По-перше, вони зберігають градації внеску всередині відібраних найбільш представлених інтенсивностей, що підвищує інтерпретованість результатів і дозволяє оцінювати не лише факт належності пікселя до набору, а й силу його типовості. По-друге, вони забезпечують узгоджений, відтворюваний спосіб привласнення ваги пікселям для подальших етапів обробки. Експериментальна проста бінаризація, тобто перетворення відібраних пікселів у маску 0, якщо піксель не належить до глобального набору, і 1 — якщо належить, не продемонструвала ніякої корисної інформації. Отримані структури виявлялися занадто великими, однорідними та не давали жодної диференційної інформації для подальшого аналізу.

2.5 Оцінка впливу клас-балансного фільтра

Для кількісного оцінювання ефективності роботи фільтра було використано вихідний набір даних, представлений за рівня квантування ($L = 128$) та значення параметра $\alpha \approx 0.9$. Обґрунтування вибору даних параметрів наведено в підрозділах 5.2.1–5.2.3.

У межах цього аналізу обчислено систему показників, що охоплює п'ять груп вимірювань. Перша група характеризує баланс представленості класів і відображає,

наскільки рівномірною є частка збережених інтенсивностей чи їх сумарної ваги в кожному класі.

Для цього використовуються коефіцієнт варіації покриття: $CV(coverage)$ [96] та співвідношення максимального і мінімального значень між класами (max/min) [97].

$$CV(coverage) = \frac{std(\tilde{c}_1, \dots, \tilde{c}_k)}{mean(\tilde{c}_1, \dots, \tilde{c}_k)} \quad (2.10)$$

де k — кількість класів, \tilde{c}_k — медіана частки збережених пікселів для зображення у k -му класі, $std(\cdot)$ та $mean(\cdot)$ — стандартне відхилення та середнє по множині $\{\tilde{c}_k\}_{k=1}^K$.

$$max/min = \frac{max_k \tilde{c}_k}{min_k \tilde{c}_k} \quad (2.11)$$

де $max(\cdot)$ та $min(\cdot)$ — найбільша та найменша медіани частки збережених пікселів серед усіх класів.

Друга група показників описує селективність відбору, тобто здатність фільтра відсіювати нерелевантні інтенсивності. Вона оцінюється через рівень покриття всередині кожного класу ($coverage$) [98].

$$coverage = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W M_{ij} \quad (2.12)$$

де H — висота частотної карти (зображення), W — ширина, HW — загальна кількість пікселів, $M_{ij} \in \{0,1\}$ — індикатор збереження пікселя (i, j)

Третя група характеризує концентрацію ваги, де вагами виступають скалярні значення частотної карти, присвоєні кожному збереженому пікселю та такі, що відображають його внесок у наборі [95]. Для цього використовуються показники:

Сумарної ваги ($sum\ values$ [92]):

$$Sum\ Values = \sum_{(i,j) \in \Omega} I_{ij} \quad (2.13)$$

де Ω — множина збережених пікселів, I_{ij} — значення частотної карти у пікселі i, j .

Концентрації енергії у нормованій формі (Energy concentration [54]):

$$EC = \sum_{(i,j) \in \Omega} p_{i,j}^2, \quad p_{i,j} = \frac{I_{ij}}{Sum\ Values} \quad (2.14)$$

де $p_{i,j}$ — нормована вага пікселя у позиції i, j , серед усіх збережених.

Ентропія нормованих ваг (entropy [99]):

$$H = - \sum_{(i,j) \in \Omega} p_{ij} \ln p_{i,j} \quad (2.15)$$

Коефіцієнт Джині (Gini [100]):

$$Gini = \frac{1}{N \cdot Sum\ Values} \sum_{n=1}^N (2n - N - 1) v_{(n)} \quad (2.16)$$

де N — кількість збережених пікселів, $v_{(n)}$ — n -та вага після сортування значень $\{I_{ij}: (i,j) \in \Omega\}$ за зростанням (тобто $v_{(1)} \leq \dots \leq v_{(N)}$), $n \in \{1, \dots, N\}$ — індекс у відсортованому ряді.

Четверта група метрик описує просторову організацію збережених областей маски, зокрема ступінь їхньої злитності, розмірності, гладкості контурів та топологічної складності. До неї належать щільність меж (Boundary Density [101]):

$$BD = \frac{\sum_{c=1}^C P_c}{A} \quad (2.17)$$

де C – кількість зв'язних компонент на зображенні, c – окремо взята компонента, A – загальна площа всіх окремих компонент, P_c – периметр окремої компоненти.

Частка найбільшої компоненти (Largest Component Ratio [102]):

$$LCR = \frac{\max_c A_c}{A} \quad (2.18)$$

де A_c – площа окремої компоненти.

Компактність (медіана по компонентах):

$$Compactness = median_c \left(\frac{P_c^2}{4\pi A_c} \right) \quad (2.19)$$

Показник просторової узгодженості (Moran's I [103]):

Для кожної частотної карти I будується бінарна маска збережених областей $M_{ij} = 1$ якщо піксель ненульовий, інакше 0. Просторова узгодженість збережених областей оцінюється показником Moran's I, який обчислюється для поля M з використанням локального 4-зв'язного сусідства:

$$I_M = \frac{\sum_{i,j} (M_{i,j} - \bar{M})(\overline{M - \bar{M}})_{N(i,j)}}{\sum_{i,j} (M_{i,j} - \bar{M})^2} \quad (2.20)$$

де $\bar{M} = \frac{1}{HW} \sum_{i,j} M_{i,j}$ – частка збережених пікселів у масці, $N(i,j)$ – множина 4-зв'язного сусідства пікселя i, j , $\overline{(M - \bar{M})}_{N(i,j)}$ – середнє центрованих значень у сусідстві.

П'ята група складається з ознак Гараліка (GLCM), які відображають властивості повторюваності, однорідності, контрасту та лінійної кореляції сусідніх інтенсивностей [54].

$$GLCM_{energy} = \sum_{a=0}^{L-1} \sum_{b=0}^{L-1} P_{ab}^2 \quad (2.21)$$

де P_{ab} – елемент матриці GLCM (ймовірність співвідношень рівнів a та b),
 L – кількість рівнів квантування для GLCM, $a, b \in \{0, \dots, L-1\}$ – індекс рівнів.

$$GLCM_{contrast} = \sum_{a=0}^{L-1} \sum_{b=0}^{L-1} (a-b)^2 P_{ab} \quad (2.22)$$

$$GLCM_{homogeneity} = \sum_{a=0}^{L-1} \sum_{b=0}^{L-1} \frac{P_{ab}}{1 + |a-b|} \quad (2.23)$$

Зведені результати для кожної з наведених метрик подано в табл. 2.3.

Таблиця 2.3 — Показники текстурних характеристик між оригінальними зображеннями та частотними картами²

Метрика	КТ- зображення (медіана по класах)	Частотна карта (медіана по класах)	Різниця, %	Медіана Δ	Типовий діапазон	Опис
Баланс покриття (між класами)	CV=0.737; max/min= 5.687	CV=0.447; max/min= 3.207	-	CV -39.38%; max/min= 43.60%	[0–100%]	Сильне вирівнювання представленості класів
Coverage (в межах класів)	0.00275421	0.00211334	-23.27%	- 0.00050354	[0, 1]	Жорсткіший відбір у межах класів
Energy concentration	0.00138542	0.00190052	+37.18%	0.0005151	[0, 1]	Більша частка маси на меншій кількості пікселів
Gini	0.00994549	0.125489	+1161.77%	0.115544	[0, 1]	Посилення концентрації ваги (нерівномірність розподілу)
Entropy	6.58189	6.28873	-4.45%	-0.29316	> 0	Розподіл став більш піковим

² Результати були відображені у матеріалах статті [95]

Продовження таблиці 2.3

Метрика	КТ- зображення (медіана по класах)	Частотна карта (медіана по класах)	Різниця, %	Медіана Δ	Типовий діапазон	Опис
Moran's I (маска)	0.937494	0.886506	-5.44%	-0.0389143	[-1, 1]	Менше суцільних плям
Boundary density	0.156724	0.573943	+266.21%	0.378327	≥ 0	Більше меж на площу
Largest component ratio	1	1	0.00%	0	[0, 1]	Домінантна компонента не зросла
GLCM energy	0.993428	0.993839	+0.04%	0.00041138 5	[0, 1]	Більше повторюваних локальних пар
GLCM contrast	0.41709	0.370998	-11.05%	-0.0456058	≥ 0	Менше різких локальних переходів
GLCM homogeneity	0.998582	0.997539	-0.10%	- 0.00104255	[0, 1]	Невелике зниження однорідності
Compactness	1.33834	1.27324	-4.86%	-0.0651025	≥ 1	Форми окремих осередків трохи компактніші

Аналіз отриманих результатів свідчить, що застосування фільтра призводить до суттєвого вирівнювання міжкласового покриття (зменшення коефіцієнта варіації на 39.38% та співвідношення max/min на 43.60%). Водночас відбувається концентрація ваги на відносно невеликій підмножині пікселів, що підтверджується зростанням показників Energy Concentration і коефіцієнта Джині, а також зниженням ентропії. Попри це, такі осередки збережених інтенсивностей залишаються дрібнодисперсними: зменшення показника просторової узгодженості Moran's I та збільшення щільності меж (Boundary Density) свідчать про підвищену фрагментованість без збільшення частки найбільшої зв'язної компоненти. Аналіз текстурних характеристик демонструє зростання повторюваності пар інтенсивностей, що відображається підвищенням показника GLCM-energy, а також зменшення кількості різких переходів, на що вказує зниження GLCM-contrast. Водночас спостерігається незначне зниження однорідності (GLCM-homogeneity). У підсумку можна зробити висновок, що запропонований фільтр досягає поставленої мети, а саме — виокремлює компактні «ядра» інтенсивностей без

формування великих суцільних сегментів. З огляду на це, подальший аналіз частотних карт доцільно зосередити на локальних і частотних патернах, які потенційно можуть виступати структурними ознаками, характерними для окремих класів.

Висновок до розділу 2

У даному розділі розглянуто етап попередньої підготовки КТ-зображень легень, спрямований на стандартизацію значень інтенсивності пікселів у межах виділених областей інтересу та забезпечення їхньої порівнюваності між пацієнтами й діагностичними групами. Встановлено, що навіть за наявності окресленої області інтересу зберігається суттєва неоднорідність розподілів інтенсивностей, зумовлена як індивідуальними особливостями зображень, так і міжкласовими відмінностями. Така неоднорідність ускладнює коректне зіставлення спостережень і створює передумови для систематичних зміщень на етапі статистичного аналізу та навчання моделей.

Для зниження цього впливу запропоновано та реалізовано процедуру відбору інтенсивностей, яка обмежує аналіз найбільш репрезентативною частиною сигналу та водночас враховує необхідність міжкласового балансування при формуванні правил відбору. Результатом виконання цього етапу є стандартизоване перетворення КТ-зображень у стандартизоване представлення зі збереженням інформативних компонентів інтенсивнісного профілю, а також формування просторових карт ваг, що відображають локальний внесок збережених інтенсивностей у межах області інтересу. Показано, що застосування фільтра забезпечує формування компактних інформативних областей та зменшує вплив фону, що підтверджено результатами кількісного оцінювання його впливу. Застосування фільтра супроводжується посиленням концентрації ваг збережених пікселів, що виражається у зростанні показника Gini з 0.00995 до 0.12549 та зниженні ентропії на 4.45%, що кількісно підтверджує зміщення вагового внеску на меншу частку пікселів (підвищення селективності відбору).

РОЗДІЛ 3

ВИДІЛЕННЯ ЛОКАЛЬНИХ СТРУКТУРНИХ ПАТЕРНІВ ТА ФОРМУВАННЯ ОЗНАК НА ОСНОВІ ЧАСТОТНИХ КАРТ ІНТЕНСИВНОСТЕЙ

Даний розділ присвячено побудові набору ознак з КТ-зображень на основі частотних карт інтенсивностей. Запропонований підхід реалізовано як послідовність процедур, що включає виділення локальних структур (патернів), приведення їх до стандартизованого кодування та подальше кількісне узагальнення у вигляді числових представлень.

У межах цього етапу первинні зображення переводяться з піксельного рівня у компактний опис, в якому кожна ознака відображає внесок відповідної локальної структури в загальний профіль зразка. Процедура зменшує надлишковість вихідних даних, знижує вплив випадкових варіацій та спрямовує аналіз на відтворювані закономірності локальної організації тканин, які можуть бути інформативними для розрізнення патологічних груп. Формування стандартизованого вектора ознак забезпечує порівнюваність зразків між собою та узгодженість подальших експериментів. Окрему увагу приділено процедурі пошуку клас-специфічних патернів та формуванню на їх основі компактного набору ознак для подальшої побудови на їх основі класифікаційних моделей.

3.1 Виокремлення локальних патернів патології

Отримані в результаті фільтрації частотні карти інтенсивностей пікселів надалі використовуються як вхідні дані для задачі виявлення локальних дискримінативних патернів. Під патернами у цьому контексті розуміються локальні просторові структури, що відповідають спільній появі певних груп пікселів і трапляються з визначеною частотою на картах у межах одного класу. Таким чином, такі патерни відображають узагальнені локальні закономірності, які кодуються частотними картами на етапі фільтрації та можуть слугувати

інформативними ознаками для подальшого аналізу і класифікації [104]. Загальний принцип процесу виявлення патернів показано на рисунку 3.1

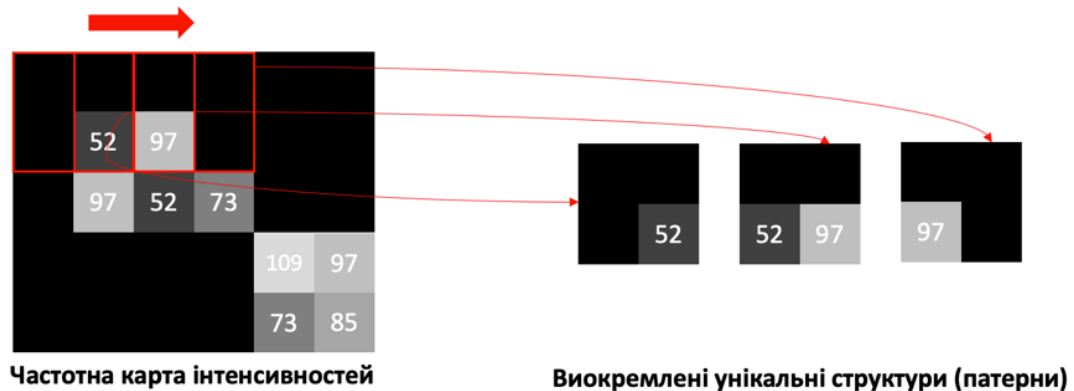


Рисунок 3.1 — Принцип вилучення структурних патернів

Формально, нехай задано множину карт $I_{c,n}(x, y)$, що позначає інтенсивність у пікселі з координатами (x, y) , індекс $c \in \{1, \dots, C\}$ відповідає діагностичному класу, індекс $n \in \{1, \dots, N_c\}$ нумерує зображення всередині класу c . H та W — висота і ширина зображення відповідно, $\mathbb{R}_{\geq 0}$ — множина невід’ємних дійсних чисел, що описують значення інтенсивності з частотної карти, тоді:

$$I_{c,n} : \{1, \dots, H\} \times \{1, \dots, W\} \rightarrow \mathbb{R}_{\geq 0} \quad (3.1)$$

На практиці, кожен $I_{c,n}$ зчитується з файлу як двовимірний цілочисельний масив. Наступний крок полягає у переході від цілого зображення до множини локальних фрагментів фіксованого розміру, які розглядатимуться як атомарні патерни. Здійснюється це за допомогою проходження по зображенню ковзним вікном розміром $p \in \mathbb{N}$. Загальний підхід дозволяє використовувати будь-яке значення $p \geq 2$. Для зображення $I_{c,n}$ та кожної позиції (i, j) , де $i \in \{1, \dots, H - p + 1\}$ та $j \in \{1, \dots, W - p + 1\}$, формується локальний патерн:

$$P_{i,j}^{(c,n)} = \begin{bmatrix} I_{c,n}(i, j) & I_{c,n}(i, j + 1) & \dots & I_{c,n}(i, j + p - 1) \\ I_{c,n}(i + 1, j) & I_{c,n}(i + 1, j + 1) & \dots & I_{c,n}(i + 1, j + p - 1) \\ \vdots & \vdots & \ddots & \vdots \\ I_{c,n}(i + p - 1, j) & I_{c,n}(i + p - 1, j + 1) & \dots & I_{c,n}(i + p - 1, j + p - 1) \end{bmatrix} \quad (3.2)$$

де $P_{i,j}^{(c,n)} \in \mathbb{R}^{p \times p}$ — фрагмент інтенсивностей у вікні розміром $p \times p$ з верхнім лівим кутом у точці (i, j) , а всі індекси i, j перебирають усі можливі положення вікна всередині зображення. Оскільки частотні карти містять значні фонові області у яких сигнал відсутній, такі ділянки відкидаються:

$$\sum_{u=0}^{p-1} \sum_{v=0}^{p-1} P_{i,j}^{(c,n)}(u, v) = 0, \quad (3.3)$$

де u та v — локальні координати всередині вікна розміром $p \times p$, не включаються в подальший аналіз. Таким чином алгоритм зосереджується на локальних фрагментах, де хоча б один піксель має ненульове значення інтенсивності. Збереження патернів, що вилучаються на кожному кроці проходження вікном із заданим розміром p вимагатиме великої кількості пам'яті. Для ефективного її використання вводиться поняття канонічного патерна, мета якого полягає в ототожненні локальних структур, які відрізняються лише елементарними геометричними перетвореннями, такими як обертання або горизонтальне віддзеркалення. Якщо два патерни відрізняються лише орієнтацією, їх доцільно вважати різними реалізаціями одного й того самого типу локальної структури (рисунок 3.2). Крім того, такий підхід інваріантний до обертання зображення та порядку слідування КТ-зрізів у серії.

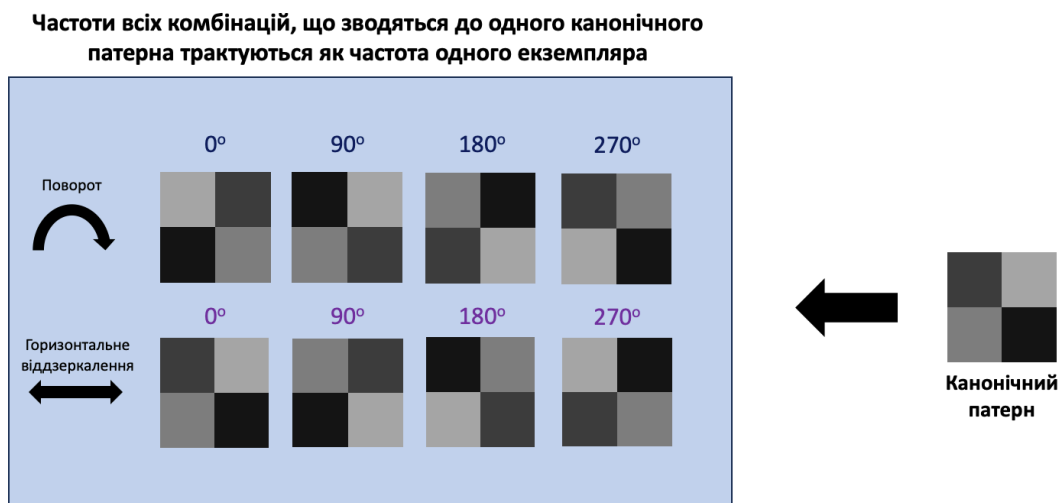


Рисунок 3.2 — Принцип виявлення канонічного патерна

Нехай $G = \{g_1, \dots, g_8\}$ — множина перетворень, що включає чотири обертання на кути $0^\circ, 90^\circ, 180^\circ, 270^\circ$ та комбінації горизонтальних віддзеркалень з обертанням. Кожен елемент $g \in G$ діє на фрагмент $P_{i,j}^{(c,n)}$, повертаючи значення такого ж розміру. Тоді для фіксованого патерна P розглядається множина всіх його образів:

$$\mathcal{V}(P) = \{g(P) \mid g \in G\}, \quad (3.4)$$

де $\mathcal{V}(P)$ — варіант патерну P за дією групи симетрій, а кожен елемент $g(P)$ — фрагмент розміру $p \times p$. Щоб вибрати єдиного представника з $\mathcal{V}(P)$, застосовується лексикографічне представлення. Нехай Q — матриця розміром $p \times p$, тоді визначається вектор $z(Q) \in \mathbb{R}^{p^2}$, що утворюється послідовним зчитуванням елементів матриці Q по рядках. Тоді лексикографічне представлення матриці Q це вектор $z(Q)$, а лексикографічний порядок між патернами Q_1 та Q_2 визначається як звичайний порядок векторів: Q_1 менший за Q_2 , якщо на першій позиції t , виконується $z(Q_1)_t < z(Q_2)_t$, де $z(Q_1)_t \neq z(Q_2)_t$. Канонічна форма патерна P визначається як той варіант елементу, що має мінімальне лексикографічне представлення:

$$\tilde{P} = \underbrace{\arg \min_{Q \in \mathcal{V}(P)} \text{lex}(z(Q))}, \quad (3.5)$$

де \tilde{P} це канонічний патерн, а $\text{lex}(z(Q))$ лексикографічний порядок вектора $z(Q)$. Таким чином вводиться відображення: функція канонізації ϕ , що кожному сирому патерну P ставить у відповідність його канонічну форму \tilde{P} :

$$\phi : P \rightarrow \tilde{P} \quad (3.6)$$

У реалізації, кожен \tilde{P} додатково кодується як HEX-значення (шістнадцятковий код) фіксованої довжини, який також відіграє роль унікального ідентифікатора цього фрагменту. Це зручно для зберігання знайдених структур у

вигляді словника для подальшого експорту отриманих даних та аналізу. Після того як для кожного вікна (i, j) обчислено канонічну форму патерна $\phi(P_{i,j}^{(c,n)})$, можна побудувати їх статистику появ. Множина K всіх патернів, які зустрічаються принаймні один раз у всьому наборі зображень $I_{c,n}(i, j)$ визначається функцією:

$$h_{c,n}(k) = \sum_{i,j} 1\{\phi(P_{i,j}^{(c,n)}) = k\}, \quad k \in K \quad (3.7)$$

де $h_{c,n}(k)$ — кількість появ канонічного патерна k у зображенні класу c з індексом n , функція $1\{\cdot\}$ набуває значення 1, якщо умова в дужках виконується і 0 в іншому випадку. Підсумовування здійснюється за всіма допустимими положеннями вікна (i, j) на зображенні. На рівні кожного окремого класу будується сумарна частота появ патерна k у всіх зображеннях цього класу:

$$H_c(k) = \sum_{n=1}^{N_c} h_{c,n}(k), \quad k \in K \quad (3.8)$$

де $H_c(k)$ це загальна кількість появ патерна k у класі c , для числа зображень N_c у цьому класі. Задача пошуку та підрахунку таких структур є обчислювально складною. Але оскільки обробка кожного окремого зображення не залежить від інших, їх аналіз можна здійснювати паралельно у вигляді множини незалежних підзадач. На практиці це реалізовано так, що для кожного зображення в межах класу окремий процес проходить ковзним вікном заданого розміру, обчислює канонічні патерни та формує локальні лічильники частот. Підзадачі розподіляються між кількома робочими процесами, кількість яких обмежена числом доступних ядер процесора робочої станції. Після завершення всі лічильники збираються в єдину структуру даних і підсумовуються на рівні класів. Така схема дає змогу істотно скоротити загальний час виконання аналізу порівняно з послідовним опрацюванням усіх зображень, оскільки основне навантаження переноситься на паралельне проходження по локальних вікнах, а централізовано

виконуються лише агрегування результатів та формування підсумкових значень. Тестування роботи алгоритму було здійснено на всьому наборі даних (див. підрозділ 2.1), який попередньо пройшов етап клас-балансної фільтрації, без застосування квантування і без фільтрації пікселів за масою. У результаті для класу зображень дрібноклітинного раку було вилучено 20243 унікальних структур, для аденокарциноми 20613, для плоскоклітинного раку з ороговінням 16478, без ороговіння 9926, для карциноїда 6015. Для зручного представлення було побудовано гістограму з відносною шкалою частоти появи таких структур. В усіх гістограмах використано єдиний глобальний простір слідування патернів, що дозволяє оцінити їх розподіл між різними класами. Для зменшення візуальної складності, було обрано відобразити 5000 найбільш представлених за частотою появи, додатково застосовано бінінг на рівні 10 (об'єднання сусідніх бінів частот). Результат продемонстровано на рисунку 3.3.

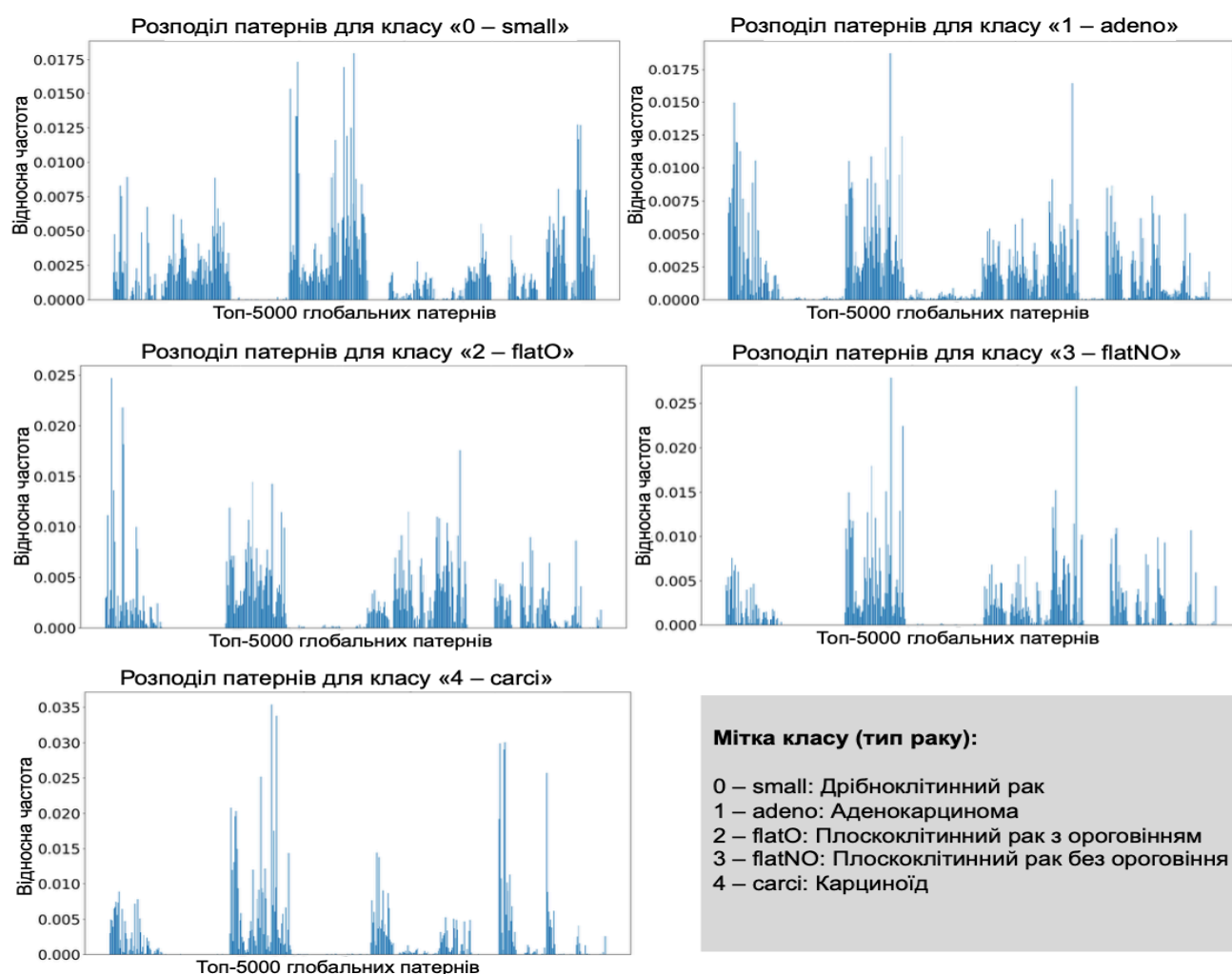


Рисунок 3.3 — Гістограма розподілу частот патернів для різних класів

Можна спостерігати чітку відмінність у профілі структур для дрібноклітинного раку та групи недрібноклітинних типів, при цьому відмінності у останніх присутні лише у вигляді різного типу розподілу. Метою подальшого аналізу є виокремлення частот тих структур, які домінують в межах конкретного класу. Для цього формується окремий набір вихідних даних, що описує розподіл знайдених патернів на рівні окремих зразків. Для кожного зразка дані містить тег класу, до якого він належить, код патерна та кількість його появ в даному зображенні.

3.2 Пошук клас-специфічних патернів

Для кожного зображення s з множини усіх зразків S , та кожного унікального патерна k з множини усіх канонічних патернів K визначається частота $f_{s,k}$, що дорівнює абсолютній кількості появ патерна k у зразку s . Дані збираються у матрицю $F = (f_{s,k})_{s \in S, k \in K}$ де кожен рядок відповідає одному зразку, а кожен стовпець одному патерну. У такій формі дані вже можуть безпосередньо використовуватися як ознаки, але поки що вони не враховують жодної інформації про відмінності між класами. Щоб сформуванати узагальнений профіль патернів, тобто наскільки часто кожен з них зустрічається в певному класі в цілому, на основі матриці F рахується сумарна кількість появ патерна k на зразку c :

$$F_c(k) = \sum_{s \in S_c} f_{s,k}, \quad k \in K \quad (3.9)$$

Оскільки різні класи можуть мати різну кількість зразків та розміри виділеної області інтересу, абсолютні частоти нормалізуються:

$$\tilde{F}_c(k) = \frac{F_c(k) - \min_{k \in K} F_c(k)}{\max_{k \in K} F_c(k) - \min_{k \in K} F_c(k) + \varepsilon}, \quad \varepsilon > 0, k \in K \quad (3.10)$$

Для кожного класу береться мінімальне та максимальне значення серед усіх патернів, а всі частоти лінійно відображаються у діапазон $[0, 1]$, що дозволяє трактувати нормалізовані значення як відносну активність певної структури в межах класу.

Наступний крок має на меті виділити патерни, які мають чітко виражену специфічність щодо одного з класів. Для кожного патерна знаходиться максимальне нормоване значення серед усіх класів та перевіряється чи є цей максимум унікальним. Формалізуємо даний крок: для фіксованого патерна k утворимо вектор $(\tilde{F}_c(k))_{c \in C}$, тобто нормалізовані частоти цього патерна в усіх агрегованих класах. Нехай $\tilde{F}^{max}(k) = \max_{c \in C} \tilde{F}_c(k)$ — найбільше значення, а $C^*(k)$ — множина класів в яких досягається цей максимум, тоді:

$$C^*(k) = \{c \in C \mid \tilde{F}_c(k) = \tilde{F}^{max}(k)\} \quad (3.11)$$

Розглядаються лише ті структури, для яких максимум є унікальним $|C^*(k)| = 1$. У цьому випадку $C^*(k) = \{c^*(k)\}$ де $c^*(k)$ можна трактувати як переважний клас для патерна k . Якщо максимум досягається в кількох класах одночасно, патерн вважається неоднозначним і не проходить критерій унікального максимуму.

Для кожної такої пари (c, k) у множині $C \times K$ вводиться тег (мітка):

$$T(c, k) = \begin{cases} \tilde{F}_c(k), & \text{якщо } c \in C, k \text{ має унікальний максимум, і } c = c^*(k) \\ 0, & \text{якщо } c \in C, k \text{ має унікальний максимум, але } c \neq c^*(k) \\ \emptyset, & \text{якщо } k \text{ не має унікальний максимум в } C \end{cases} \quad (3.12)$$

Для $T(c, k)$ формується подвійний зміст, з одного боку, це індикатор, що певний k пов'язаний з класом c , з іншого — числова міра сили цього зв'язку. Побудовані теги застосовуються назад до матриці F . На цьому кроці початкові частоти патернів у кожному зразку коригуються залежно від того, до якого класу належить зразок і як патерн поводить себе на рівні класів. Якщо k має унікальний максимум у певному класі, то його частоти зберігаються тільки для зразків цього класу, а для всіх інших жорстко зануляються.

$$\begin{aligned} \tilde{f}_{s,k} &= f_{s,k}, & \text{якщо } c(s) = c^*(k) \\ \tilde{f}_{s,k} &= 0, & \text{якщо } c(s) \neq c^*(k) \end{aligned} \quad (3.13)$$

Для патернів, які не пройшли критерій унікального максимуму, частоти залишаються без змін у всіх зразках. У результаті утворюється замаскована версія матриці частот, де внесок клас-специфічних патернів посилений, а їхня присутність у інших класах приглушена.

3.3 Формування ознак на основі клас-специфічних патернів

На основі скоригованої матриці \tilde{F} формуються кілька типів ознак, що будуть використані для подальшого моделювання. Найпростішою є бінарна репрезентація, у якій для кожного зразка s та патерну k розглядається лише факт появи, а не кількість повторів. Визначимо бінарну ознаку:

$$x_{s,k}^{binary} = \begin{cases} 1, & \text{якщо } \tilde{f}_{s,k} > 0 \\ 0, & \text{якщо } \tilde{f}_{s,k} = 0 \end{cases} \quad (3.14)$$

Інший тип представлений у вигляді відносних частот, який інваріантний до загальної кількості фрагментів у зразку. Для кожного зразка s обчислюється загальна кількість усіх патернів $N_s = \sum_{k \in K} \tilde{f}_{s,k}$, де N це загальне число локальних фрагментів, які не були відфільтровані маскуванням. Визначимо ознаку відносної частоти:

$$x_{s,k}^{relative} = \begin{cases} \frac{\tilde{f}_{s,k}}{N_s}, & \text{якщо } N_s > 0 \\ 0, & \text{якщо } N_s = 0 \end{cases} \quad (3.15)$$

Ще одним варіантом побудови ознак є застосування TF-IDF зважування поверх відносних частот. Ідея полягає в тому, щоб підсилити вагу тих структур, які відносно часто зустрічаються в конкретному зразку, але рідко в усьому наборі, і

навпаки, послабити внесок патернів, що є майже універсальними. Якщо $|S|$ загальна кількість зразків, а df_k це кількість зразків, у яких патерн k присутній хоча б один раз:

$$df_k = |\{s \in S | \tilde{f}_{s,k} > 0\}| \quad (3.16)$$

тоді можемо визначити зворотну частку:

$$idf_k = \log\left(\frac{|S| + 1}{df_k + 1}\right) + 1 \quad (3.17)$$

Доданки 1 у чисельнику та знаменнику запобігають діленню на нуль та логарифму нуля, а наприкінці idf_k завжди більший за нуль. Визначимо ознаку TF-IDF:

$$x_{s,k}^{tfidf} = x_{s,k}^{relative} \cdot idf_k \quad (3.18)$$

Додатково використовується логарифмічне перетворення та стандартизація по стовпцях, таке перетворення пом'якшує вплив дуже великих значень частот та наближує розподіл ознак до більш симетричного. Спершу для кожної пари (s, k) обчислюється $y_{s,k} = \log(1 + \tilde{f}_{s,k})$, доданок 1 гарантує коректність для нульових частот. Для кожного патерну k обчислюється середнє значення:

$$\mu_k = \frac{1}{|S|} \sum_{s \in S} y_{s,k} \quad (3.19)$$

та стандартне відхилення:

$$\sigma_k = \sqrt{\frac{1}{|S|} \sum_{s \in S} (y_{s,k} - \mu_k)^2} \quad (3.20)$$

Якщо σ_k виявляється надто малим або нульовим у випадках, коли патерн майже не варіює між зразками, у чисельних реалізаціях його замінюють на 1. Визначимо стандартизовані ознаки:

$$x_{s,k}^{logZ} = \frac{y_{s,k} - \mu_k}{\sigma_k} \quad (3.21)$$

Крім того були збережені й значення в абсолютних частотах.

Для узагальненої характеристики внеску окремих патернів у простір ознак було застосовано візуалізацію кумулятивної ваги ознак у вигляді гістограми, яка обчислюється як сума значень відповідної ознаки за всіма зразками вибірки та інтерпретували як інтегральну міру кумулятивної статистичної ваги патерна в наборі, що за змістом відповідає L1-мірі. З метою зменшення візуальної складності виконано дискретизацію шляхом агрегації 110 патернів у бін, а також використано спільний єдиний порядок ознак, що забезпечує коректну зіставність різних схем кодування частот.

Отримані профілі демонструють, що для представлень у відносних частотах та при TF-IDF зважуванні загальна конфігурація вектору ознак є подібною, тобто домінантні за кумулятивною вагою групи патернів зберігаються між цими двома способами подання (рисунок 3.4). На рівні міжкласових відмінностей спостерігається, що для класів дрібноклітинного раку, аденокарциноми та плоскоклітинного раку з ороговінням характерна наявність виражених вузлів, обмежених груп патернів, які формують непропорційно великий внесок у сумарний профіль відповідної патології. Натомість для плоскоклітинного раку без ороговіння та карциноїда профіль є менш контрастним і більш розрідженим — без домінування невеликої кількості бінів і з меншою концентрацією ваги в обмеженій підмножині патернів.

У сукупності це вказує на різний ступінь компресії сигналу отриманого з ознак, у частини класів розподіл визначається небагатьма повторюваними структурами, тоді як для інших внесок більш розподілений між ширшим спектром патернів.

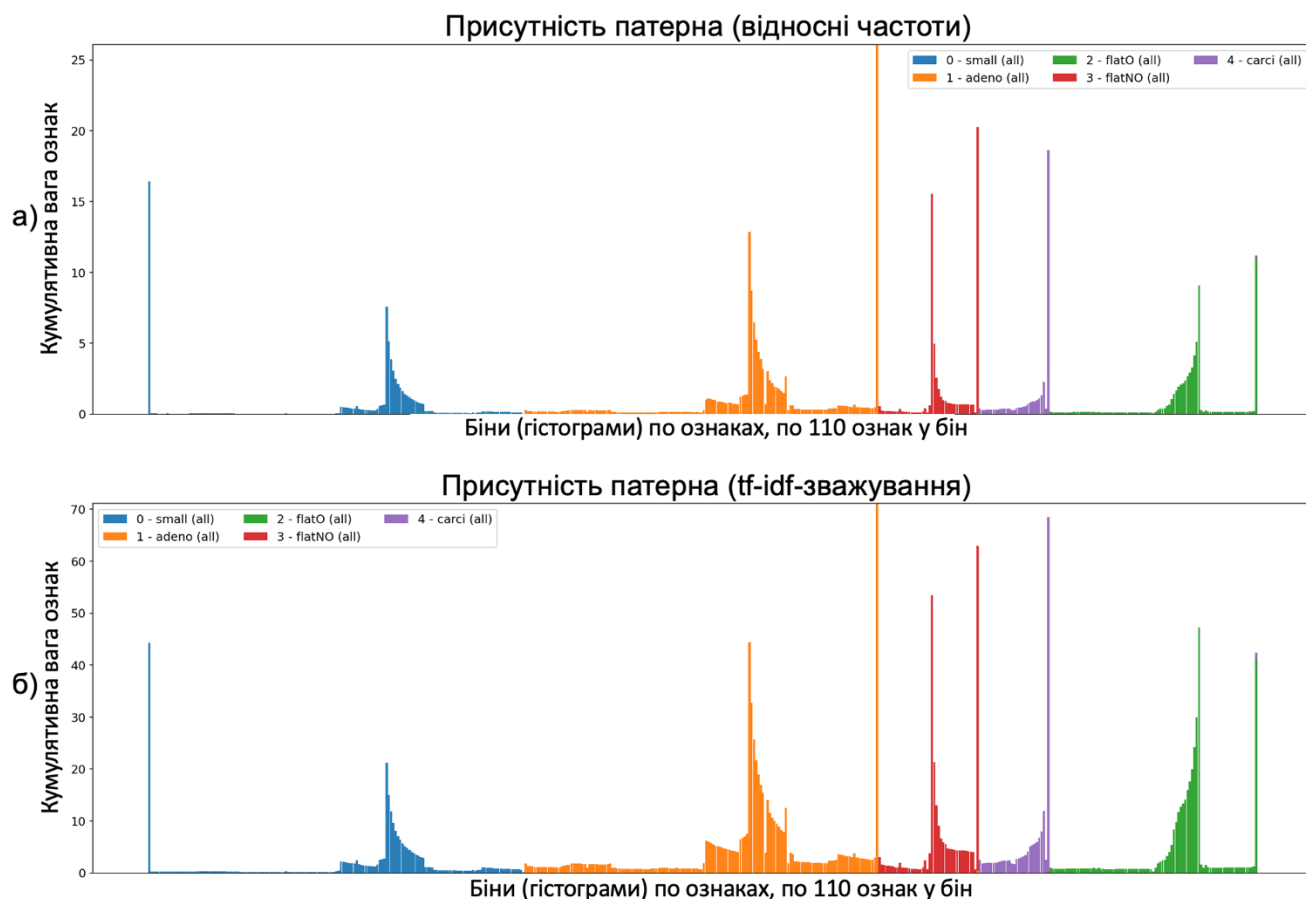


Рисунок 3.4 — Гістограма кумулятивної ваги ознак. а) представлення у відносних частотах; б) представлення TF-IDF зважування

У таких умовах TF-IDF є більш привабливим підходом до представлення ознак, оскільки переводить частотні значення в міру інформаційної цінності, зменшує внесок надто поширених, потенційно неспецифічних компонент і відносно підсилює рідкісні патерни, які можуть бути діагностично значущими, але залишатися малопомітними при аналізі лише абсолютних або відносних частот. Візуально це узгоджується з тим, що на TF-IDF-профілях зберігається структура основних піків, водночас зростає виразність окремих локальних внесків поза домінантними зонами.

Проста бінаризація зберігає інформацію лише про факт наявності патерна, але втрачає відомості про його інтенсивність, що зменшує чутливість представлення до градацій сигналу (рисунок 3.5). На рівні отриманих профілів це проявляється у згладжуванні слабких внесків і концентрації видимого сигналу навколо найбільш представлених структур, відповідно, виражені піки

інтерпретуються як патерни, притаманні значній частці зразків певного класу, тоді як зони малої сукупної ваги відповідають рідкісним та/або високоспецифічним структурам, представленим фрагментарно. Водночас така редукція може мати перевагу у вигляді меншої чутливості до випадкових флуктуацій частот, тобто до шуму.

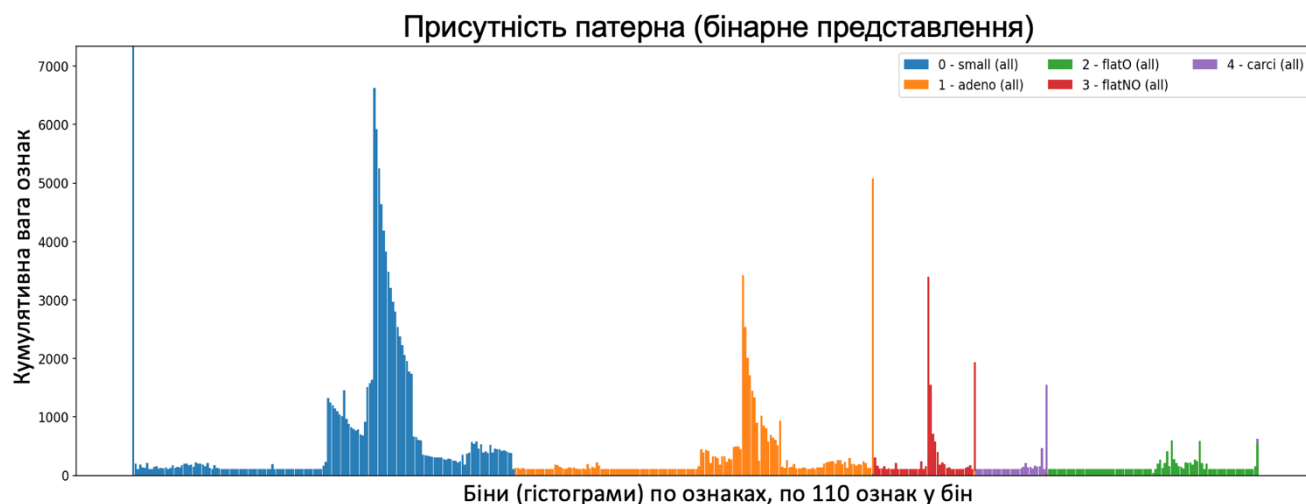


Рисунок 3.5 — Гістограма кумулятивної ваги ознак у бінарному представленні

Для аналізу сигналів із високою дисперсією застосовано логарифмічне перетворення та варіант логарифмічного представлення зі стандартизацією (рисунок 3.6а). Звичайне логарифмічне представлення стискає ваги домінантних піків, завдяки чому внески середнього та малого масштабу стають розрізняваними і не маскуються високими значеннями. Унаслідок цього з'являється можливість аналізувати структуру профілю поза найбільш вираженими зонами та оцінювати, наскільки периферійні ділянки простору ознак формують додатковий клас-специфічний сигнал.

Логарифмічне представлення зі стандартизацією (рисунок 3.6б) спрямоване на підкреслення відхилень від фонового рівня та підвищення співставності між класами. У такому поданні домінування окремих піків зменшується, а увага переноситься на відносні зміни профілю в межах широких ділянок простору ознак. Це забезпечує більш рівномірне порівняння класів і робить помітними закономірності, які у лінійному масштабі залишаються другорядними через перевагу кількох найвищих піків.

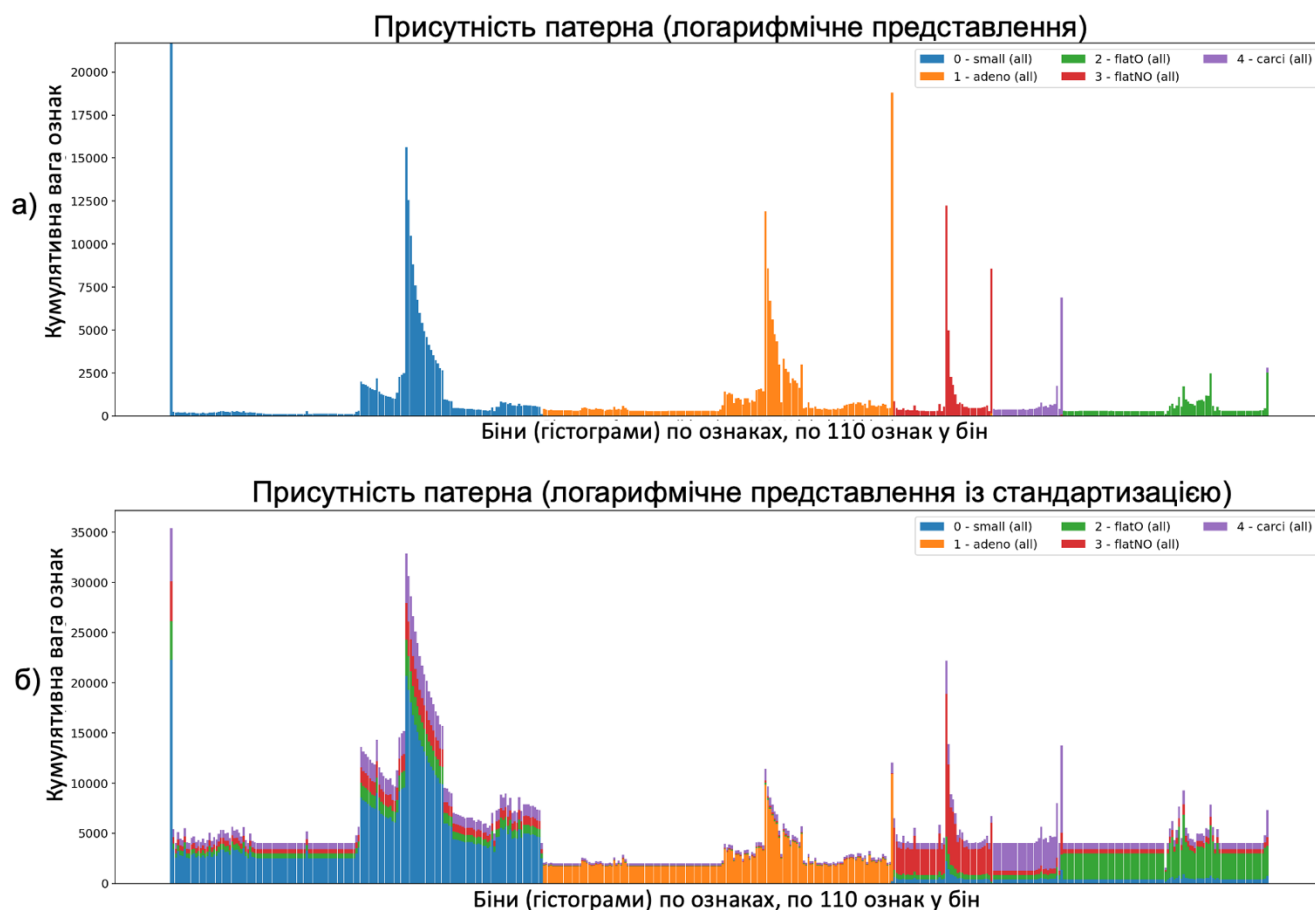


Рисунок 3.6 — Гістограма кумулятивної ваги ознак. а) Логарифмічне представлення; б) Логарифмічне представлення із стандартизацією

Для візуалізації цих ознак було застосовано процедуру локального нормування, що полягає у зміщенні значень всередині кожного біна відносно його мінімуму з подальшим перетворенням у частки, що дозволило зберегти та коректно відобразити локальну структуру варіативності між класами, уникнувши ефекту згладжування через значні глобальні зміщення від'ємних значень. Зазначене локальне нормування застосовано лише для візуалізації профілів і не змінює визначення ознак.

Висновок до розділу 3

У цьому розділі було запропоновано підхід до опису КТ-зображень через систему локальних патернів, що відображають повторювані структурні конфігурації специфічні для класу, та слугують основою для формування простору

ознак. На цій основі розроблено метод формування ознак за КТ-зображеннями, що ґрунтується на виділенні повторюваних локальних структурних патернів та їх кількісному поданні. Запропоновано декілька варіантів числового представлення частот патернів, які переводять спостережувані структури у зіставний простір ознак та дозволяють аналізувати внесок окремих компонент незалежно від вибраної схеми кодування.

Отримані результати демонструють, що розподіл внесків патернів є неоднорідним і має виражену концентрацію в обмежених ділянках простору ознак, а також виявлено міжкласові відмінності у характері такого розподілу. Порівняння альтернативних подань частот показало збереження загальної структури сигналу за одночасної зміни контрасту між домінантними та другорядними компонентами залежно від способу кодування.

Отримані способи представлення є методологічним підґрунтям для подальшого їх використання у ролі набору ознак, отриманих з КТ-зображень легень для класифікації типів легеневого раку.

РОЗДІЛ 4

СТРАТЕГІЯ «ОДИН ПРОТИ РЕШТИ» З ПОСЛІДОВНИМ ВІДОКРЕМЛЕННЯМ КЛАСІВ ДЛЯ ЗАДАЧІ БАГАТОКЛАСОВОЇ ДІАГНОСТИКИ

У цьому розділі розглядається стратегія побудови моделі, що передбачає використання набору базових класифікаторів та спеціалізованої схеми прийняття рішення у форматі «один проти решти» з послідовним видаленням класів. Дана схема дозволяє за рахунок послідовного відокремлення класів розв'язати задачу узгодженого багатокласового рішення в умовах дисбалансу класів і обмеженого обсягу даних, зменшуючи неоднорідність негативного класу на кожному кроці навчання. Реалізація цього етапу дає змогу отримати формалізований механізм прогнозування та порівнювані вихідні оцінки моделей, необхідні для подальшого зіставлення підходів і вибору робочої конфігурації класифікатора.

4.1 Багатокласова класифікація за стратегією «один проти решти»

Підхід «один проти решти» (One-vs-Rest, OvR) є одним з найпоширеніших способів розв'язання багатокласової задачі шляхом редукції до набору бінарних класифікацій. Якщо маємо K взаємовиключних класів $\{c_1, \dots, c_k\}$, тоді при OvR-схемі навчаються K бінарних моделей $f_k(x)$, де для k -го класифікатора позитивним є клас c_k , а негативним — об'єднання всіх інших класів $\{c_j : j \neq k\}$. На етапі прогнозування для зразка x обчислюють ймовірність для кожного k , після чого призначають клас із максимальним значенням $\arg \max_k f_k(x)$. Така редукція є загальною і використовується з багатьма базовими алгоритмами, зокрема, SVM або логістичною регресією, оскільки дозволяє зберегти простоту бінарного навчання та масштабуватися на довільну кількість класів [105,106].

Попри простоту, підхід «один проти решти» має характерну властивість: для кожного c_k проти решти негативний клас є неоднорідним, адже складається з суміші кількох класів із потенційно різною статистикою ознак. У задачах з малою

вибіркою та незбалансованим розподілом класів це може загострювати ще дві проблеми. По-перше, сильна асиметрія для рідкісних класів, позитивних прикладів мало, тоді як негативних — багато. Така ситуація є типовою причиною зміщення моделі в бік більшості та деградації чутливості до міноритарних класів [107]. По-друге, змішування різних класів у решту може ускладнювати формування стабільної розділювальної межі. Модель змушена апроксимувати все інше одним негативним розподілом, що може бути статистично некоректним, особливо за наявності близьких (важко розрізняваних) класів [105,108].

Додатково емпіричні дослідження редукційних схем показують, що «один проти решти» не завжди гарантує покращення якості та інколи може погіршувати результати порівняно з підходами, які природно підтримують багатокласовість або використовують інші на основі голосування/ансамблеві стратегії [109]. Це не заперечує доцільності підходу, але підкреслює потребу в адаптаціях, особливо коли вибірка даних мала й незбалансована.

4.2 Класифікація за стратегією каскадного «один проти решти»

У постановці цього дослідження робота здійснюється в умовах обмеженої кількості зразків та явної присутності дисбалансу між діагностичними групами. Природною є гіпотеза, що багатоступенева (каскадна) декомпозиція може частково нівелювати вищевказані ефекти. Це дозволить, по-перше, зменшити неоднорідність негативного класу на кожному кроці, по-друге, контролювати складність рішення через керований порядок відокремлення класів. Ідея послідовного прийняття рішень та організації набору бінарних класифікаторів у вигляді каскаду є відомою у літературі багатокласової класифікації, зокрема, у вигляді Decision DAG / DAGSVM, де підкреслюється ефективність та коректність багатоступеневого комбінування бінарних моделей [110].

Запропонована архітектура класифікатора реалізує послідовну елімінацію класів на основі логістичної регресії (рисунок 4.1). На відміну від звичайної "один проти решти", де всі K бінарних моделей навчаються паралельно на повній вибірці,

тут будується ланцюжок із $K-1$ моделей, а дані для навчання кожного наступного кроку звужуються до підмножини ще не відкинутих класів.

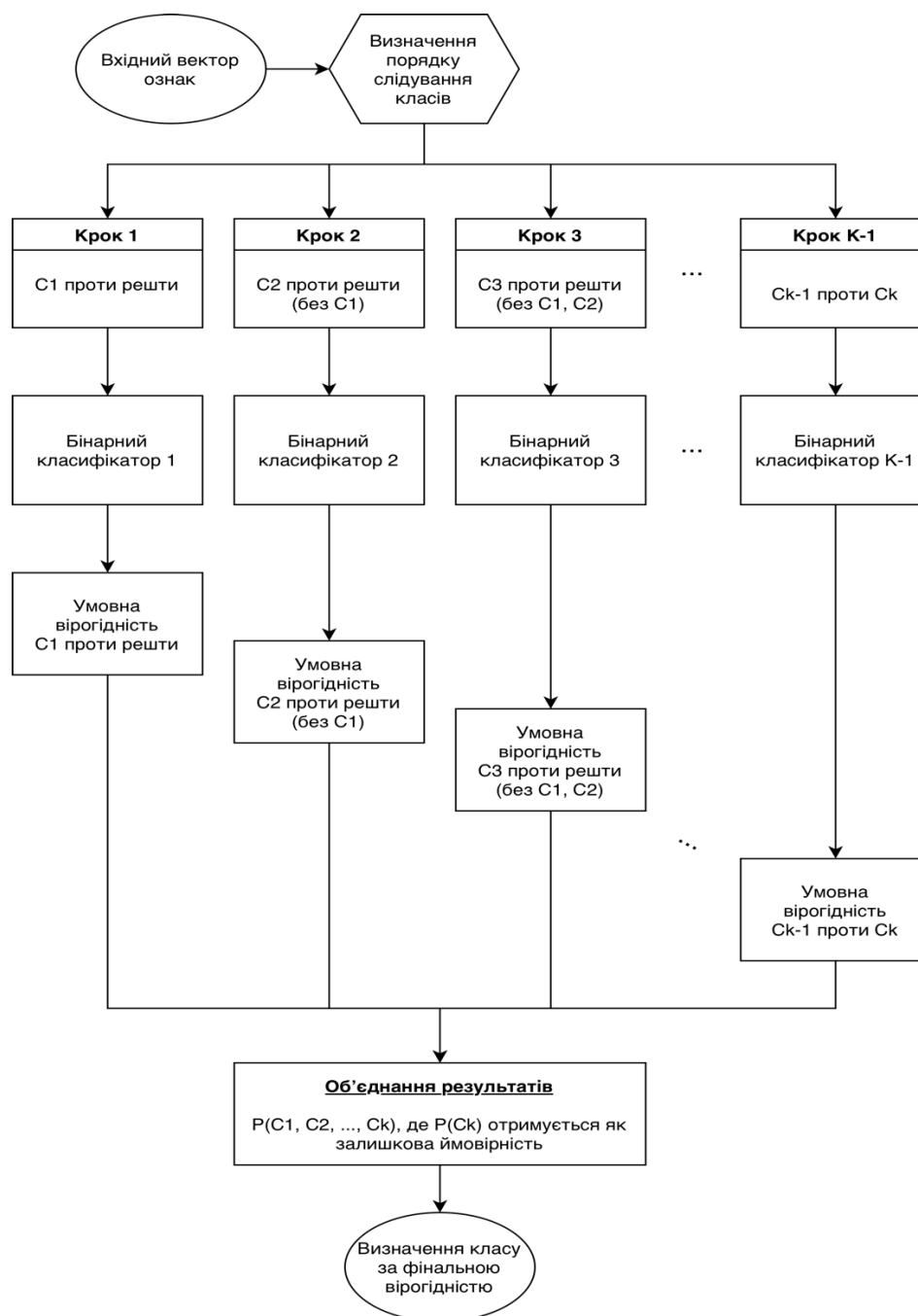


Рисунок 4.1 — Класифікації за стратегією каскадного «один проти решти»

Формально, нехай порядок відкидання класів визначено як $w = (c_1, c_2, \dots, c_K)$, тоді навчання відбувається:

Крок 1. Формується бінарна задача c_1 проти $\{c_2, \dots, c_K\}$ та навчається модель на основі логістичної регресії $h_1(x)$.

Крок 2. Із навчальної вибірки концептуально відкидаються всі приклади класу c_1 , після чого навчається $h_2(x)$ на задачі c_2 проти $\{c_3, \dots, c_k\}$

Крок $K - 1$. Навчається $h_{K-1}(x)$ для c_{K-1} проти $\{c_k\}$. Окремий класифікатор для c_K не потрібен, оскільки він є залишковим класом.

Ключовою відмінністю є те, що на кроці i модель бачить лише дані з класів $\{c_i, \dots, c_K\}$. Завдяки цьому негативний клас («решта») стає менш неоднорідним, а дисбаланс між цільовим класом і «рештою» для частини класів може бути менш різким, ніж у класичній OvR, де «решта» завжди об'єднує $K - 1$ інших класів. У контексті дисбалансу це узгоджується із загальними спостереженнями, що якість моделі чутливі не лише до кількості прикладів, а й до структури та складності негативного класу в навчанні [107].

При реалізації було передбачено керування порядком слідування класів. Порядок може бути фіксованим, у такому випадку обирається параметр "as is", або може задаватися вручну на основі клінічних чи дослідницьких пріоритетів, або визначатись автоматично шляхом багаторазової перестановки. Це важливо, оскільки каскадна декомпозиція не є симетричною. Ранні кроки отримують найбільш широкий негативний клас, а пізні — більш локальні рішення. Відповідно, порядок може підбиратися як елемент індуктивного упередження, наприклад, спочатку відокремлювати найбільш відмінний клас або найбільш представлений клас — залежно від мети експерименту. Схожі міркування про наслідки багатоступеневої організації класифікації (включно з ризиком накопичення помилки) широко обговорюються в контексті ієрархічних та DAG-подібних підходів (DAG, directed acyclic graph — орієнтований ациклічний граф) [111,112].

4.3 Виведення багатокласових ймовірностей

На кроці i логістична регресія повертає умовну ймовірність:

$$q_1(x) = P(y = c_i \mid y \in \{c_i, \dots, c_K\}, x) \quad (4.1)$$

Далі багатокласові (безумовні) ймовірності відновлюються каскадною формулою, яка відповідає послідовному відсіву класів:

$$\begin{aligned}
 P(y = c_1 | x) &= q_1(x) \\
 P(y = c_2 | x) &= (1 - q_1(x)) \cdot q_2(x) \\
 P(y = c_3 | x) &= (1 - q_1(x))(1 - q_2(x)) \cdot q_3(x) \\
 &\dots \\
 P(y = c_K | x) &= \prod_{j=1}^{K-1} (1 - q_j(x))
 \end{aligned} \tag{4.2}$$

Така конструкція має дві важливі властивості: вона гарантує нормування, сума ймовірностей дорівнює 1; вона робить модель інтерпретованою як послідовний стохастичний процес прийняття рішення, що концептуально близький до відомих DAG-побудов у багатокласовій класифікації, де остаточний клас визначається шляхом проходження маршруту рішень у графі [111].

Остаточне передбачення формується як клас із максимальною багатокласовою ймовірністю $\arg \max_k P(y = c_k | x)$ після приведення до стандартного порядку міток класів.

4.3.1 Додаткові механізми калібрування рішень

Оскільки в незбалансованих задачах важливо контролювати компроміс між precision (або точністю) та recall (або повнотою), у реалізації додатково використано підбір порогів для кожного класу за критерієм максимізації F1 на тренувальних даних. Для класу k обирається поріг t_k , що максимізує F1 уздовж кривої precision–recall (PR). Застосування PR-кривих є методологічно обґрунтованим саме для дисбалансних сценаріїв, де ROC-оцінки можуть бути надто оптимістичними, тоді як precision–recall є більш чутливою до помилок на міноритарному класі [113].

На етапі прогнозування рішення приймається так:

1. формується множина кандидатів $\{k: P_k(x) \geq t_k\}$;

2. якщо кандидати існують, обирається клас із найбільшим відносним запасом $P_k(x)/t_k$;
3. якщо жоден поріг не виконано, використовується стандартний $\arg \max_k P_k(x)$

У практичному сенсі це дозволяє задавати різні робочі точки для різних класів, що часто є необхідним при асиметричних втратах і при відмінностях у простоті розпізнавання окремих патологічних груп.

4.4 Автоматичний відбір порядку слідування класів

Оскільки каскадна декомпозиція не є симетричною відносно перестановок класів, порядок w фактично є окремим структурним параметром моделі. Ранні кроки каскаду формуються за умов найбільш неоднорідного негативного класу та максимального дисбалансу між позитивним та негативним результатом, тоді як пізні кроки працюють у звуженому просторі класів і, відповідно, реалізують локальні рішення. Унаслідок цього порядок w визначає як розподіл складності між кроками, так і характер накопичення помилок. Ранні помилки є незворотними в межах проходження каскаду, тоді як пізні рішення приймаються вже після послідовного відсування частини альтернатив. Отже, вибір w доцільно розглядати як задачу структурної оптимізації, спрямовану на узгодження загальної збалансованої якості багатокласового рішення та пріоритетів щодо окремих діагностичних груп.

Формально, нехай множина класів $\{c_1, \dots, c_K\}$ упорядкована перестановкою $w = (c_{\pi(1)}, \dots, c_{\pi(K)})$. Для фіксованого w каскад визначає відновлення безумовних багатокласових ймовірностей через умовні оцінки $q_i(x)$ за формулами (4.1)–(4.2).

Водночас практичне порівняння різних порядків потребує узгодженого правила дискретизації, оскільки навіть при однаковій архітектурі різні w можуть змінювати масштаб імовірностей окремих класів. Тому для кожного порядку додатково виконується добір клас-специфічних порогів t_k , що визначають робочий режим класифікатора, фіксуючи співвідношення між точністю та повнотою для

кожного класу. Зокрема, для кожного класу c_k поріг визначається як значення, що максимізує F1-міру у постановці « c_k проти решти» на валідаційних прогнозах:

$$t_k(w) = \arg \max_{t \in [0,1]} F1(\mathbb{I}[y = c_k], \mathbb{I}[P_w(y = c_k|x) \geq t]) \quad (4.3)$$

де $t_k(w)$ — оптимальний (за критерієм F1) поріг прийняття рішення для класу c_k за фіксованого порядку каскаду w , $t \in [0,1]$ — значення порогів для кандидата, F1-міра для бінарного розрізнення класу c_k проти решти, $\mathbb{I}[\cdot]$ — індикаторна функція, дорівнює 1, якщо умова в дужках істинна, і 0 — інакше, y — істинна мітка класу для спостереження x , $P_w(y = c_k|x)$ — оцінена моделлю з порядком w ймовірність належності x до класу c_k .

Після цього багатокласове рішення формується за пороговим правилом, описаним у п. 4.3.1, що забезпечує порівнюваність режимів роботи для різних класів і зменшує схильність моделі домінувати класами більшості.

Оцінювання кандидатних порядків здійснюється на навчальній вибірці в межах внутрішньої валідації із використанням позавибіркових (out-of-fold) прогнозів, що мінімізує оптимістичне зміщення при відборі конфігурації. Пошук оптимального порядку формалізується як задача максимізації (4.4) на множині кандидатних перестановок $W \subseteq S_K$:

$$w = \arg \max_{w \in W} MacroF1(w) \quad (4.4)$$

За малої кількості класів W можливе застосування повного перебору усіх перестановок S_K , що забезпечує глобальний відбір у межах заданого критерію. Однак зі зростанням K повний перебір стає обчислювально складною задачею через факторіальне зростання кількості варіантів, тому доцільним є обмежений пошук. У такому випадку W формується як поєднання детермінованих якірних порядків (наприклад, базовий порядок, упорядкування за частотністю класів, а також крайні режими із розміщенням пріоритетного класу на початку або наприкінці каскаду) та випадкових перестановок, що забезпечують стохастичне

покриття простору можливих структур. Оцінювання кожного $w \in W$ виконується за однаковою процедурою внутрішньої валідації та з однаковим правилом порогової калібровки за формулою (4.3).

Після визначення w порядок фіксується та використовується для навчання фінальної каскадної моделі на повній навчальній вибірці. Підсумкова оцінка якості здійснюється на незалежній тестовій вибірці, яка не залучалася ні до відбору порядку, ні до добору порогів, що забезпечує коректність висновків щодо внеску структурної оптимізації у підвищення узгодженості багатокласового рішення.

Висновок до розділу 4

У цьому розділі запропоновано метод побудови багатокласового класифікатора за стратегією каскадного «один проти решти», що має на меті підвищення узгодженості багатокласового рішення за умов неоднакової складності розпізнавання класів.

Метод дозволяє звести багатокласову задачу до послідовності бінарних підзадач, у межах яких на кожному кроці відокремлюється один клас від множини решти з поступовою елімінацією вже класифікованих спостережень та передаванням невіднесених прикладів на наступні етапи каскаду.

Було запропоновано механізм оптимізації структури каскаду, що полягає у відборі порядку перевірки класів на основі внутрішньої валідації з використанням позавибіркових прогнозів і максимізації інтегрального критерію якості, визначеного за макро-F1-мірою.

РОЗДІЛ 5

ДИФЕРЕНЦІАЦІЯ ТИПІВ РАКУ ЛЕГЕНЬ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

У даному розділі розглянуто застосування запропонованого підходу опису КТ-зображень легень через систему локальних структурних патернів до задачі диференціації гістологічних типів раку легень.

Розв'язання задачі багатокласової класифікації виконано на основі сформованого простору ознак частот патернів із порівнянням альтернативних схем подання та аналізом впливу ключових параметрів попередньої обробки на узагальнювальну здатність моделей. Також досліджено ефективність різних архітектур класифікаторів у порівнянні із запропонованою стратегією каскадного «один проти решти».

Наведено приклад практичного використання класифікаторів у прикладному сценарії та формат подання результатів для підтримки прийняття рішень.

5.1 Підготовка навчальної та тестової вибірок

Попередня обробка даних, а саме етапи фільтрації, екстракції патернів та формування ознак (включаючи їх нормалізацію та зважування) застосовувалися до повної сукупності зображень єдиним масивом. Формування тренувальної та тестової вибірок відбувалося безпосередньо перед етапом класифікації із застосуванням пацієнт-орієнтованої стратифікації. Ідентифікатор пацієнта брався з імені зразка, після чого всі зображення, що належали одному пацієнтові, розглядалися як неподільна група. Далі множина пацієнтів розбивалася на навчальну та тестову підмножини у пропорції 70/30 зі стратифікацією за класом на рівні пацієнта. Розбиття здійснювалося багаторазово з різними початковими ініціалізаціями генератора випадкових чисел. Серед отриманих варіантів обирався такий, де кожен клас представлений хоча б одним пацієнтом у кожній з підмножин. Якщо після серії спроб окремі класи все ж були відсутні у тренувальній або

тестовій підмножині, виконувалася мінімальна корекція розбиття: переносилась найменша необхідна кількість пацієнтів відповідних класів між тренувальним і тестовим наборами, по можливості не вилучаючи останнього представника класу з донорського набору. Ключовою особливістю даного підходу є суворе розмежування пацієнтів, що унеможливорює потрапляння зразків одного пацієнта одночасно до вибірки навчання та тестування, запобігаючи витoku інформації.

5.2 Обрані конфігурації для класифікації

Для оцінки впливу клас-балансного фільтра на результати класифікації, було систематично порівняно кілька груп сценаріїв формування частотних карт інтенсивностей, що відрізнялися як просторовим квантуванням, так і параметрами побудови. Було сформовано три групи сценаріїв: (1) без квантування; (2) квантування до рівня $L=128$ (під параметром L мається на увазі ступінь квантування); (3) квантування до рівня $L=64$.

Для кожного сценарію окремо сформовано по п'ять наборів даних КТ-зображень з різними рівнями фільтрації за глобальною масою пікселів: без фільтрації, де залишаються усі інтенсивності оригінальних зображень, з фільтрацією на рівнях 95% / 90% / 75% / 60% (під фільтрацією мається на увазі скільки відсотків маси пікселів на глобальному наборі інтенсивностей залишати для оригінальних зображень) Надалі рівень фільтрації буде визначатись через значення параметра α . Отримані 15 наборів із різними налаштуваннями пройшли етап пошуку та відбору патернів. У результаті отримано дані про розподіл частот знайдених патернів між усіма зразками, на основі яких сформовано шість варіантів кодування ознак: абсолютні значення частот, бінарні ознаки, де враховується лише факт наявності структури, відносні частоти, нормовані до інтервалу $[0,1]$, логарифмічне перетворення, логарифмічне перетворення із стандартизацією ознак та TF-IDF перетворення частот патернів.

Отримані 90 наборів даних КТ-зображень використовувалися для розв'язання задачі багатокласової класифікації із використанням 9 моделей, що

покривають як лінійні, так і нелінійні підходи, а також різні типи ансамблів. Додатково в ролі референтних моделей роботі було розглянуто два підходи до класифікації: класичні текстурні ознаки Гараліка та згорткові нейронні мережі. Ознаки Гараліка формувалися на основі матриць суміжності рівнів сірого (GLCM), з параметрами кількість рівнів 256, застосуванням симетричності, нормування, значення відстаней $\{1,2,3,4,5\}$ та кутів $0^\circ, 45^\circ, 90^\circ, 135^\circ$. Для зменшення впливу фону, пікселі із значенням 0 ігнорувались. Із кожної GLCM обчислювалися стандартні характеристики: контраст, відмінність, гомогенність, енергія, кореляція, кутовий момент, що в підсумку давало 120 числових ознак на зображення, які використовувалися як вхідний вектор ознак для класичних моделей класифікації.

Підхід на основі згорткових нейромереж (CNN) реалізовано у вигляді двох варіантів архітектури, що працюють із зображеннями, попередньо приведеними до фіксованого просторового розміру та нормалізованими за інтенсивністю до діапазону $[0;1]$. Для підвищення стійкості моделі до природних варіацій зображень під час навчання застосовано помірні перетворення вхідних даних (аугментація): горизонтальне віддзеркалення з імовірністю 0,20; випадкові повороти в межах $\pm 5^\circ$; а також афінні перетворення зі зміщенням до 3% від розміру зображення вздовж кожної осі та масштабуванням у межах 0,95–1,05. Додатково використано фотометричні варіації шляхом незначної зміни яскравості та контрасту (амплітуда 0,05 для кожного параметра).

Перший варіант архітектури є послідовною багатошаровою згортковою моделлю, що містить чотири блоки обробки з ядром 3×3 та кількістю каналів 32, 64, 128 і 256. Після кожного блоку застосовано нормалізацію та нелінійне перетворення за допомогою функції активації ReLU. Після перших трьох блоків виконувалося зменшення просторової роздільності за допомогою операції підвибірки 2×2 (max pooling). Узагальнене представлення ознак формувалося шляхом адаптивного усереднення (adaptive average pooling), після чого застосовано регуляризацію випадковим зануленням $p=0,35$ (dropout) та фінальне лінійне перетворення розмірності $256 \rightarrow K$, де K — кількість класів.

Другий варіант побудовано на основі стандартної архітектури ResNet18. Навчання обох моделей виконувалося методом адаптивної градієнтної оптимізації (AdamW) із використанням зваженої функції втрат багатокласової класифікації, де ваги класів задавалися обернено пропорційно їх представленості у навчальному наборі. Для запобігання перенавчанню застосовано процедуру раннього припинення навчання за критерієм тренувальної втрати.

У ролі основних моделей використовувалися:

1) Багатокласова (мультиноміальна) логістична регресія, навчання якої виконувалося ітераційним оптимізаційним методом (SAGA), що допускає застосування розріджувальної регуляризації (L1). Для компенсації дисбалансу між класами вводилися вагові коефіцієнти, обернено пропорційні частотам класів у навчальній вибірці (режим *balanced*). З метою забезпечення збіжності при високій розмірності простору ознак граничну кількість ітерацій збільшено до 5000.

2) До групи ансамблевих підходів на деревах рішень віднесено модель випадкового лісу (Random Forest), у якій формувалася сукупність 500 дерев, кожне з яких будувалося на бутстреп-підвибірці навчальних даних, а поділ у вузлах здійснювався за випадковою підмножиною ознак. Дисбаланс класів у цій моделі також враховувався за допомогою зважування (режим *balanced*).

3, 4) Окремо розглядалися методи послідовного нарощування ансамблю дерев (градієнтний бустинг), реалізовані у двох бібліотеках: (XGBoost) та (LightGBM). Обидві реалізації застосовувалися у багатокласовому режимі з отриманням оцінок належності до класів у вигляді ймовірностей та використанням багатокласової логарифмічної функції втрат (*multiclass log-loss*). Для (LightGBM) додатково фіксувалися початковий стан генератора випадкових чисел з метою відтворюваності.

5) Підхід на основі незалежного усереднення багатьох базових моделей (*bagging*), реалізований як ансамбль із 300 дерев рішень, навчальних на випадкових бутстреп-підвибірках. У цьому випадку стабілізація результатів досягалася переважно через усереднення великої кількості незалежних моделей, без додаткового посилення регуляризаційних обмежень на рівні окремого дерева.

6, 7) Два бустингові алгоритми: AdaBoost та Gradient Boosting. У AdaBoost як базову модель використовували неглибоке дерево рішень (максимальна глибина 2), а ансамбль складався з 300 таких дерев за фіксованої швидкості навчання ($\text{learning rate} = 0.5$). У Gradient Boosting формувався ансамбль із 300 дерев із глибиною до 3 рівнів за меншої швидкості навчання ($\text{learning rate} = 0.05$), що забезпечує більш поступове уточнення моделі та дає змогу описувати нелінійні залежності.

8) Багаторівневий ансамбль типу стекінгу (stacking). На першому рівні навчалися три базові моделі: випадковий ліс (400 дерев), XGBoost та LightGBM із параметрами, узгодженими з наведеними вище. На другому рівні використовувалася метамодель у вигляді логістичної регресії з регуляризацією L2, яка отримувала на вхід вихідні ознаки та додаткові метаознаки, сформовані з ймовірнісних прогнозів базових моделей. Як метаознаки використовувалися саме вектори ймовірностей для всіх класів, а не лише остаточні передбачені мітки.

9) Нарешті, використано власний підхід за стратегією каскадного «один проти решти», що будує послідовність бінарних моделей, де на кожному кроці формується задача «поточний клас проти решти», але навчання відбувається лише на підмножині зразків, які ще не були відсіяні попередніми кроками. Кожен такий бінарний класифікатор — це логістична регресія з L2-регуляризацією, балансуванням ваг класів та збільшеною максимальною кількістю ітерацій.

Для оцінювання якості класифікації застосовано сукупність загальних і клас-орієнтованих показників. Загальна правильність класифікації на тестовій вибірці (Assurasy) відображає частку коректно віднесених зразків, однак за нерівномірної представленості класів її інформативність є обмеженою, оскільки вона може не відображати зниження якості на малочисельних групах. З цієї причини додатково обчислювалися показники точності та повноти (на тестових даних) з однаковим внеском кожного класу (Macro Precision і Macro Recall), а також узагальнений показник збалансованої якості (Macro F1), який інтегрує точність і повноту на рівні класів. Окремо розраховувався зважений показник F1 (Weighted F1), що враховує частоти класів і характеризує середню якість на вибірці з урахуванням дисбалансу. Для контролю перенавчання використовували різницю між зваженим F1 на

навчальній і тестовій вибірках (Overfit Gap), яка інтерпретується як міра розбіжності між якістю на даних, що використовувалися для побудови моделі, та її узагальнювальною здатністю.

5.2.1 Підбір оптимальних параметрів фільтра

Першочергово було оцінено роботу клас-балансного фільтра та те, наскільки сильно можна стиснути простір інтенсивностей, не втрачаючи репрезентативності осередків для всіх класів. Варто уточнити, що подальші операції у цьому підрозділі здійснюються за першою стратегією розділення даних на навчальну та тестову вибірки, де всі етапи застосовувалися до повної сукупності зображень.

Якщо подивитися не лише на відсоток покриття класів, а й на те, які обсяги даних реально відкидає фільтр, можна спостерігати, що для жорсткого режиму $\alpha \approx 0.6$ у всіх трьох варіантах квантування до базового набору потрапляє лише 4–5% доступних рівнів інтенсивності (3 значення з 62 для $L=64$, 6/125 для $L=128$ і 11/247 без квантування), але вони вже несуть близько 60–62% сумарної маси пікселів (сумарної кількості пікселів у всіх зображеннях, що мають певні рівні інтенсивності). Понад 90% можливих рівнів просто не використовуються, хоча по масі це лише 38–40% даних. За такого стискання втрачається значна частина інформації тих класів, чий спектр інтенсивностей ширший або зміщений від глобальних піків — саме тому клас 0 - small та 3 - flatNo у цьому режимі найменш представлені.

При $\alpha \approx 0.9$ фільтр покриває вже 90–91% маси пікселів, використовується 13–15% усієї доступної шкали інтенсивностей, близько 85% рівнів відсіюються, оскільки несуть відносно невелику частку частот. Це добре узгоджується з природою КТ-даних, де більшість рідкісних інтенсивностей відповідає артефактам, тонким краям та поодиноким структурами.

При $\alpha \approx 0.95$ зберігається $\approx 95\%$ маси, залишається приблизно чверть усіх рівнів (15/62, 30/125, 59/247), відсікається близько 75% інтенсивності.

Зведені результати наведені у табл. 5.1

Таблиця 5.1 — Зведені результати відбору частки пікселів за різних рівнів квантування

Без квантування							
Глобальна маса (α)	К-ть базових бінів	Маса базових бінів, %	0 - small	1 - adeno	2- flatO	3 - flatNo	4 - carci
0.60	11	60.30	21.35	70.70	73.87	58.18	77.41
0.75	18	76.30	70.18	80.96	85.37	62.74	82.23
0.90	33	90.20	95.66	92.69	99.24	71.72	91.67
0.95	59	95.01	99.16	96.58	99.91	81.96	97.42
1.00	247	100.00	100.00	100.00	100.00	100.00	100.00
Квантування L = 128							
Глобальна маса (α)	К-ть базових бінів	Маса базових бінів, %	0 - small	1 - adeno	2- flatO	3 - flatNo	4 - carci
0.60	6	62.24	36.72	71.25	70.30	56.81	76.12
0.75	9	75.70	64.02	82.25	86.42	63.19	82.60
0.90	17	90.49	97.06	92.69	99.24	71.75	91.71
0.95	30	95.09	99.19	96.85	99.94	81.92	97.55
1.00	125	100.00	100.00	100.00	100.00	100.00	100.00
Квантування L = 64							
Глобальна маса (α)	К-ть базових бінів	Маса базових бінів, %	0 - small	1 - adeno	2- flatO	3 - flatNo	4 - carci
0.60	3	61.58	6.83	74.36	82.17	61.58	80.97
0.75	5	80.01	75.63	84.58	90.54	64.97	84.33
0.90	9	91.09	95.89	93.83	99.58	73.25	92.88
0.95	15	95.03	99.49	96.85	99.93	81.40	97.48
1.00	62	100.00	100.00	100.00	100.00	100.00	100.00

У результаті видаляється переважно малорепрезентативна частина даних, тоді як основна маса пікселів усіх класів зберігається. Поєднання цих двох ефектів — сильне скорочення простору інтенсивностей при мінімальній втраті маси, та стійкість результатів до вибору кількості рівнів квантування робить діапазон $\alpha \approx 0.9-0.95$ найбільш доцільним для подальшого аналізу та навчання класифікаторів.

5.2.2 Вплив квантування на результати класифікації

Порівняння режимів квантування демонструє, що перехід від неперервних інтенсивностей до дискретного представлення впливає на статистичні властивості

ознак і, як наслідок, на узагальнювальну здатність моделей. У табл. 5.2 наведено зведені середні результати класифікації за різних рівнів квантування.

Таблиця 5.2 — Зведені результати класифікації залежно від квантування інтенсивностей

Показники моделей	Режим квантування		
	L=128	L=64	Без квантування
Accuracy	0.9206 ± 0.0578	0.8923 ± 0.0739	0.8415 ± 0.1165
Macro Precision	0.9028 ± 0.0951	0.8371 ± 0.1244	0.8776 ± 0.1325
Macro Recall	0.8741 ± 0.0884	0.8137 ± 0.0916	0.7860 ± 0.1267
Macro F1	0.8722 ± 0.0911	0.8043 ± 0.1093	0.7794 ± 0.1450
Weighted F1 (Test)	0.9120 ± 0.0684	0.8694 ± 0.0909	0.8155 ± 0.1388
Weighted F1 (Train)	1.0000 ± 0.0002	0.9997 ± 0.0026	1.0000 ± 0.0000
Overfit Gap	0.0880 ± 0.0684	0.1303 ± 0.0907	0.1845 ± 0.1388

У режимі $L=128$ спостерігається найкраща узгодженість між високими тестовими метриками та низьким Overfit Gap. Це можна інтерпретувати як ефект контрольованого спрощення сигналу, квантування зменшує надмірну чутливість до дрібних коливань інтенсивності й тим самим знижує ризик навчання на випадкових деталях. Водночас за такого рівня залишається достатня кількість інформації на знімках та не втрачається ключова структурна інформація, яка потрібна для розрізнення класів.

Режим $L=64$ демонструє проміжну поведінку, сильне стиснення шкали інтенсивностей призводить до часткової втрати дискримінативних відмінностей. Це проявляється у зниженні макро-метрич і більшому розриві між результатами на навчальних та тестових даних. Тож можна зробити висновок, що надмірно грубе квантування може зменшити не лише шум, а й корисний сигнал, особливо у випадку класів, що відрізняються тонкими текстурними характеристиками.

Сценарій без квантування систематично характеризується найнижчою узагальнювальною якістю та найбільшою варіативністю метрик. Така поведінка узгоджується з тим, що неперервні інтенсивності несуть значно більше варіативності, через яку модель може підлаштовуватись під конкретні властивості тренувального набору. Додатковим практичним фактором є різко вищі витрати

часу на навчання у режимі без квантування. Це вказує на збільшення складності обчислень та простору ознак і опосередковано підтверджує, що даний режим формує менш компактне та менш стабільне представлення даних.

Найбільш доцільним для подальшого аналізу та навчання класифікаторів є сценарій із застосуванням рівня квантування $L=128$.

5.2.3 Вплив режиму квантування та жорсткості фільтрації інтенсивностей на результати класифікації

Параметр α фактично визначає ступінь насичення відбору інтенсивностей та безпосередньо впливає на те, скільки інформації з розподілу інтенсивностей потрапляє у фінальне представлення. Деталізація за сценаріями (табл. 5.3) показує, що залежність результатів від α не є лінійною. Натомість проявляється характерний компроміс між інформативністю отриманого опису та його стабільністю.

Таблиця 5.3 — Зведені результати класифікації для різних сценаріїв

Сценарій	Accuracy	Macro F1	Weighted F1	Overfit Gap
quant128_alpha60	0.9139 \pm 0.0875	0.8569 \pm 0.1225	0.8997 \pm 0.1055	0.1003 \pm 0.1055
quant128_alpha75	0.9181 \pm 0.0617	0.8636 \pm 0.0965	0.9095 \pm 0.0678	0.0904 \pm 0.0676
quant128_alpha90	0.9374 \pm 0.0337	0.8870 \pm 0.0781	0.9289 \pm 0.0446	0.0711 \pm 0.0446
quant128_alpha95	0.9235 \pm 0.0437	0.8805 \pm 0.0789	0.9167 \pm 0.0531	0.0833 \pm 0.0531
quant128_alpha100	0.9100 \pm 0.0445	0.8732 \pm 0.0699	0.9051 \pm 0.0522	0.0949 \pm 0.0522
quant64_alpha60	0.8251 \pm 0.1178	0.7272 \pm 0.1654	0.7943 \pm 0.1505	0.2057 \pm 0.1505
quant64_alpha75	0.8779 \pm 0.0796	0.7751 \pm 0.1123	0.8555 \pm 0.0948	0.1434 \pm 0.0946
quant64_alpha90	0.9226 \pm 0.0405	0.8489 \pm 0.0841	0.9074 \pm 0.0514	0.0926 \pm 0.0514
quant64_alpha95	0.9164 \pm 0.0461	0.8362 \pm 0.0914	0.8971 \pm 0.0587	0.1029 \pm 0.0588
quant64_alpha100	0.9076 \pm 0.0704	0.8338 \pm 0.1171	0.8886 \pm 0.0859	0.1114 \pm 0.0859
noquant_alpha60	0.7927 \pm 0.1200	0.7762 \pm 0.1577	0.8003 \pm 0.1485	0.1997 \pm 0.1485
noquant_alpha75	0.8052 \pm 0.1247	0.7334 \pm 0.1707	0.7805 \pm 0.1564	0.2195 \pm 0.1564
noquant_alpha90	0.8848 \pm 0.0940	0.8077 \pm 0.1191	0.8640 \pm 0.1084	0.1360 \pm 0.1084
noquant_alpha95	0.8580 \pm 0.1144	0.7796 \pm 0.1439	0.8339 \pm 0.1317	0.1661 \pm 0.1317
noquant_alpha100	0.8669 \pm 0.0927	0.8123 \pm 0.1167	0.8526 \pm 0.1027	0.1474 \pm 0.1027

Для рівня квантування $L=128$ сценарії з більшими значеннями α (близькі до 0.90–0.95) формують найбільш інформативне представлення. Водночас надмірне збільшення α до 1.00 не гарантує приросту, додавання додаткових значень

інтенсивностей може включати менш стабільні або менш значущі компоненти, які розмивають корисний сигнал. Для менших значень параметру α (0.60–0.75) спостерігається зниження якості, що логічно узгоджується з втратою частини релевантних структур, модель отримує більш очищені, але й більш збіднені дані.

У режимі $L=64$ сценарії з малими α виявилися найслабшими. Поєднання грубого квантування та жорсткого відбору інтенсивностей занадто сильно фільтрує корисний сигнал, тому ознаки стають недостатньо чутливими до відмінностей між класами. При переході до $\alpha \approx 0.90$ –1.00 ситуація покращується, однак загальний рівень якості зазвичай нижчий, ніж у відповідних сценаріях для $L=128$.

Для режиму без квантування варіативність результатів залишається високою майже для всіх α . З цього можна зробити проміжний висновок, що відбір маси інтенсивностей як механізм не може повністю компенсувати нестабільність, яку вносить неперервне інтенсивнісне представлення. У цьому випадку навіть при відносно оптимальних α модель залишається більш схильною до перенавчання, а приріст у макро-метриках є обмеженим.

Оптимальним для подальшого аналізу є використання пари рівня квантування $L=128$ при відборі 90-95% маси інтенсивностей після фільтрації.

5.2.4 Вплив способу представлення ознак на результати класифікації

Вибір представлення ознак визначає, які саме аспекти даних стають доступними моделі. Порівняння способів кодування ознак показує, що найкращу та найстабільнішу якість забезпечують відносні частоти і TF-IDF зважування. З огляду на однакові підсумкові показники для цих двох варіантів, можна дійти висновку, що вирішальна інформація для розмежування класів закладена передусім у співвідношеннях частот патернів у межах зразка, тобто в тому, які структури переважають і в яких пропорціях вони поєднуються. TF-IDF у цій ситуації виконує радше допоміжну роль. Він коригує внесок дуже поширених патернів і підсилює відносно рідкісні, однак у середньому не змінює картину настільки, щоб дати помітний виграш порівняно з простим нормуванням частот. Нижчі результати для бінарного подання є очікуваними, оскільки воно зводить опис до факту наявності або відсутності патерна у зразку і тим самим втрачає відомості про те, наскільки

часто він є представленим, а саме ця різниця виявилася важливою для класової відмінності. Логарифмічні перетворення також можуть послаблювати роздільну здатність, бо стискають діапазон значень і зменшують контраст між вираженими та слабо представленими компонентами. Найменш придатним є використання абсолютних частот, оскільки такі значення сильніше залежать від масштабу.

Зведені середні результати точності класифікаційних моделей для кожного з представлень наведені у табл. 5.4.

Таблиця 5.4 — Зведені результати класифікації залежно від методу формування ознак

Показники моделей	Подання					
	Relative	TF-IDF	Binary	Log	LogZ	ABS
Accuracy	0.9104 ± 0.0837	0.9104 ± 0.0837	0.7991 ± 0.0785	0.7883 ± 0.0842	0.7883 ± 0.0842	0.7481 ± 0.1046
Macro Precision	0.9083 ± 0.1092	0.9083 ± 0.1092	0.7841 ± 0.1131	0.7764 ± 0.1055	0.7764 ± 0.1055	0.7327 ± 0.1236
Macro Recall	0.8578 ± 0.1032	0.8578 ± 0.1032	0.7419 ± 0.0955	0.7315 ± 0.0977	0.7315 ± 0.0977	0.6733 ± 0.1031
Macro F1	0.8589 ± 0.1124	0.8589 ± 0.1124	0.7361 ± 0.1070	0.7240 ± 0.1101	0.7240 ± 0.1101	0.6994 ± 0.1223
Weighted F1 (Test)	0.8968 ± 0.0977	0.8968 ± 0.0977	0.7808 ± 0.0954	0.7692 ± 0.1022	0.7692 ± 0.1022	0.7443 ± 0.1137
Weighted F1 (Train)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9997 ± 0.0014
Overfit Gap	0.1032 ± 0.0977	0.1032 ± 0.0977	0.2192 ± 0.0954	0.2308 ± 0.1022	0.2308 ± 0.1022	0.2554 ± 0.1137

Оптимальними для подальшого застосування у класифікаторах є використання ознак відносної частоти або побудованих за принципом TF-IDF.

5.2.5 Вплив архітектури класифікатора

Результати порівняння моделей демонструють, що модель на основі випадкового лісу дерев рішень у середньому забезпечує найвищі значення якості за макро-метриками та найменший Overfit Gap. Це можна пояснити тим, що ансамблі дерев ефективно моделюють взаємодію ознак і локальні правила, не вимагаючи лінійності або гладкості розділювальної межі між класами. Для

структурних патернів це може бути важливо, оскільки їх комбінації можуть мати значення лише у сукупності, а не поодиноці. Матриці відповідності, для найкращої конфігурації, а саме, застосування квантування на рівні $L=128$ за відбору 90% маси відфільтрованих інтенсивностей на ознаках типу TF-IDF продемонстровані на рисунку 5.1 (кількість зразків для класу 0 - small — 25, 1 - adeno — 44, 2 - flatO — 39, 3 - flatNO — 14, 4 - carcі — 36).

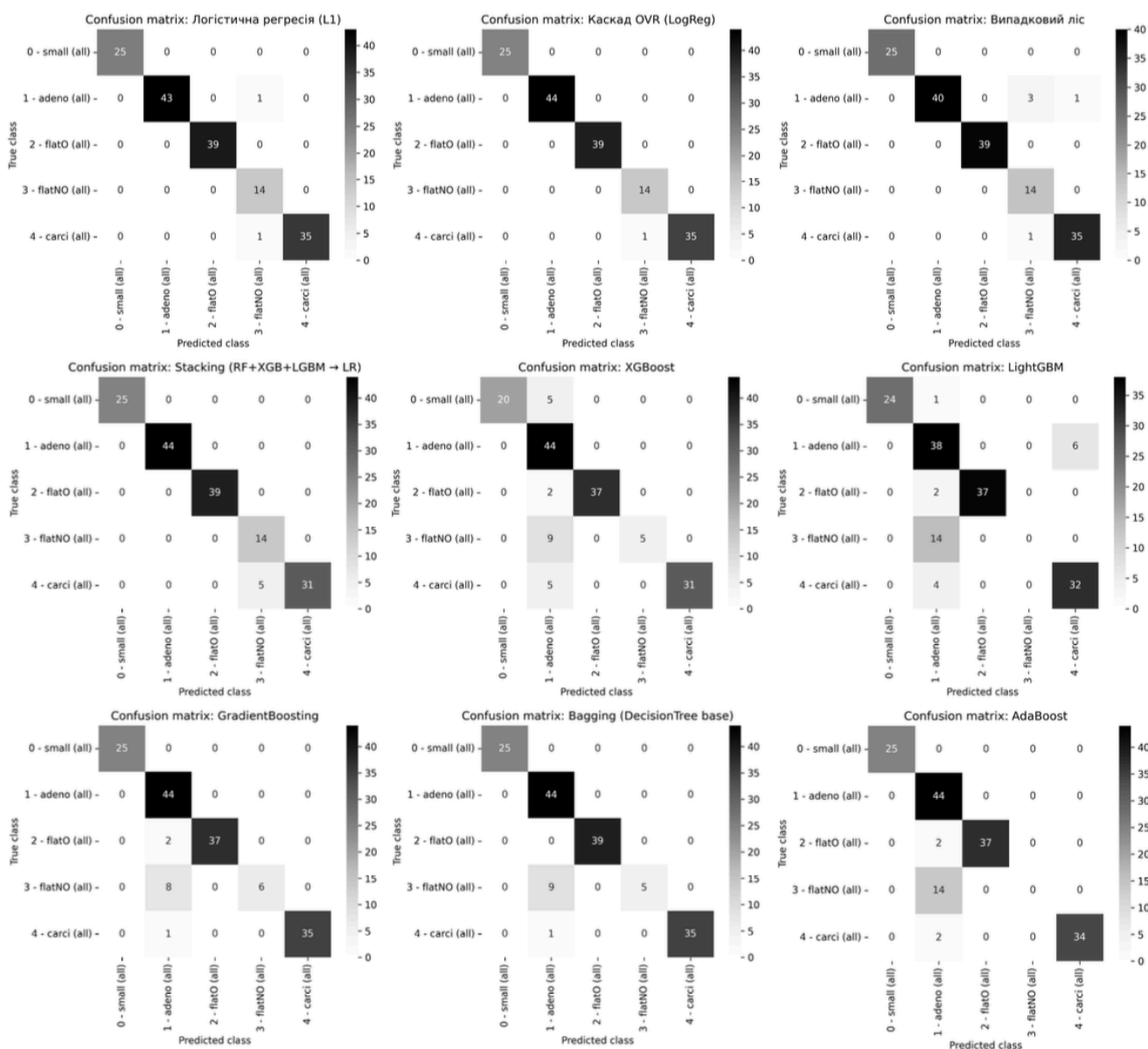


Рисунок 5.1 — Матриці відповідності для моделей

Стекінг демонструє високі результати, але в середньому поступається випадковому лісу. Імовірною причиною є те, що складні ансамблеві композиції є більш чутливими до якості калібрування й стабільності базових прогнозів.

Запропонована власна архітектура побудована за стратегією каскадного «один проти решти» на базі логістичної регресії також демонструє високі результати, вона здатна добре лінійно відокремлювати класи у сформованому просторі, особливо якщо виділяти їх послідовно. Це цікавий методологічний висновок, що значна частина роботи відбувається на рівні побудови ознак і сценарію, після чого навіть відносно проста лінійна модель, але з певним порядком слідування класів здатна працювати не гірше, ніж на рівні складних ансамблів.

GradientBoosting та Bagging дають проміжні результати. Bagging додатково характеризується значно більшим часом отримання прогнозу, що може бути критичним для практичного застосування. Для бустингових моделей XGBoost, LightGBM та AdaBoost у межах розглянутого простору налаштувань спостерігається нижча середня якість та більший Overfit Gap. Це може свідчити або про недостатньо оптимальні гіперпараметри для цих моделей у даній задачі, або про те, що їхній механізм послідовного підсилення залишкових помилок підвищує ризик підлаштування під шумові компоненти ознак у разі обмеженого обсягу даних і високої варіативності між пацієнтами.

Зведені результати для всіх моделей наведені у табл. 5.5.

Таблиця 5.5 — Зведені результати класифікації залежно від архітектури класифікатора

Модель	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1 (Test)	Weighted F1 (Train)	Overfit Gap
Випадковий ліс	0.9494 ± 0.0479	0.9494 ± 0.0625	0.9118 ± 0.0690	0.9161 ± 0.0738	0.9425 ± 0.0600	1.0000 ± 0.0000	0.0575 ± 0.0600
Стекінг	0.9150 ± 0.0782	0.9157 ± 0.0854	0.8791 ± 0.1054	0.8750 ± 0.1080	0.9058 ± 0.0922	1.0000 ± 0.0000	0.0942 ± 0.0922
Каскадний OvR	0.8952 ± 0.1196	0.8927 ± 0.1110	0.8637 ± 0.1476	0.8479 ± 0.1603	0.8800 ± 0.1416	0.9996 ± 0.0025	0.1196 ± 0.1414
Логістична регресія	0.8809 ± 0.1096	0.9108 ± 0.0860	0.8357 ± 0.1263	0.8321 ± 0.1389	0.8663 ± 0.1298	0.9994 ± 0.0039	0.1331 ± 0.1295
Gradient Boosting	0.8884 ± 0.0710	0.8950 ± 0.1011	0.8193 ± 0.0769	0.8209 ± 0.0914	0.8693 ± 0.0868	1.0000 ± 0.0000	0.1307 ± 0.0868

Продовження таблиці 5.5

Модель	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1 (Test)	Weighted F1 (Train)	Overfit Gap
Bagging	0.8955 ± 0.0989	0.8350 ± 0.1736	0.8167 ± 0.1117	0.8035 ± 0.1446	0.8690 ± 0.1280	1.0000 ± 0.0000	0.1310 ± 0.1280
XGBoost	0.8462 ± 0.0707	0.8594 ± 0.1083	0.7723 ± 0.0619	0.7707 ± 0.0840	0.8216 ± 0.0897	1.0000 ± 0.0000	0.1784 ± 0.0897
LightGBM	0.8577 ± 0.0671	0.7877 ± 0.1242	0.7699 ± 0.0617	0.7580 ± 0.0821	0.8276 ± 0.0839	1.0000 ± 0.0000	0.1724 ± 0.0839
AdaBoost	0.8364 ± 0.0877	0.7999 ± 0.1062	0.7506 ± 0.0746	0.7407 ± 0.0813	0.8091 ± 0.0908	1.0000 ± 0.0000	0.1909 ± 0.0908
Референтні моделі							
Модель	Accuracy	Macro F1		Weighted F1 (Test)		Overfit Gap	
Haralik	0.5814 ± 0.1204	0.3980 ± 0.0759		0.5252 ± 0.1077		0.4641 ± 0.1073	
CNN	0.3029	0.2633		0.2100		0.7869	
ResNet18	0.4400	0.3291		0.4330		0.5192	

Для порівняння з запропонованим підходом розглянуто референтні моделі, що охоплюють як класичні інженерні текстурні ознаки, так і підходи на основі нейромережевої обробки зображень. Результати моделей на ознаках Гараліка можна розглядати як нижній рівень порівняння. Вони відображають загальні статистичні характеристики текстури, однак не описують структуру локальних поєднань, на яких ґрунтується запропоноване подання. Унаслідок цього їхня якість є нижчою, а різниця між результатами на навчальній і тестовій вибірках є більш вираженою, що вказує на обмежену здатність таких ознак відтворювати міжкласові відмінності в умовах складної морфологічної мінливості. Такий характер результатів можна пояснити припущенням, що відмінності між класами зумовлені не стільки загальними текстурними характеристиками на кшталт неоднорідності, скільки специфічними просторовими конфігураціями та поєднаннями структурних елементів.

Базові згорткові нейромережеві моделі (CNN) у розглянутих налаштуваннях демонструють ще нижчу узагальнювальну якість. Для простішої архітектури CNN суттєва розбіжність між навчальними та тестовими показниками узгоджується з перенавчанням, коли модель добре відтворює навчальні приклади, але не формує ознак, достатньо стійких для перенесення на нові дані. Використання більш структурованої архітектури (ResNet18) у поєднанні з помірними перетвореннями навчальних зображень (аугментацією) частково покращує результати, однак не забезпечує зіставного рівня якості. Імовірним поясненням є поєднання обмеженого обсягу даних, значної варіативності між зразками пацієнтів та високої параметричної складності нейромережевих моделей, які за таких умов потребують або більшої кількості навчальних прикладів, або більш жорстких та цілеспрямованих процедур регуляризації й налаштування режиму навчання.

5.3 Оцінка стійкості класифікаційних моделей

Якість будь-якої моделі класифікації визначається не лише її узгодженістю з навчальними даними, але й здатністю до узагальнення, тобто прийняття коректних рішень на нових спостереженнях, які не використовувалися під час навчання. Це особливо критично в задачах медичної візуалізації, де присутня варіабельність у даних між пацієнтами, що можуть спричиняти нестійкість моделі та приховану адаптацію до специфіки навчальної вибірки.

З метою перевірки стійкості моделей в умовах наближених до реального застосування було реалізовано другу стратегію підготовки та розбиття вибірки КТ-зображень на навчальну і тестову, за якої розділення даних виконувалося до проведення будь-яких процедур обробки, а всі подальші операції виконувалися незалежно для кожного піднабору. За основу було взято результати попередніх досліджень (див. підрозділ 5.2), які показали, що найефективнішою конфігурацією є квантування інтенсивностей на рівні $L=128$, що показав зменшення варіативності, зумовлені дрібномасштабними флуктуаціями частот, а також зниження розмірності опису простору інтенсивностей, зберігаючи водночас достатньо інформативний

сигнал для подальшого розрізнення класів. Додатково застосовувався критерій відбору 90% маси інтенсивностей, який відсікає низькоінформативні ділянки гістограми розподілу інтенсивностей пікселів та зменшує внесок шумових компонент. Формування ознак виконувалося за принципом TF–IDF зважування, що переводить частоти у відносну шкалу та підсилює внесок рідкісних структур, які потенційно мають дискримінативну здатність для розділення класів. Загальний принцип цього процесу показано на рисунку 5.2.

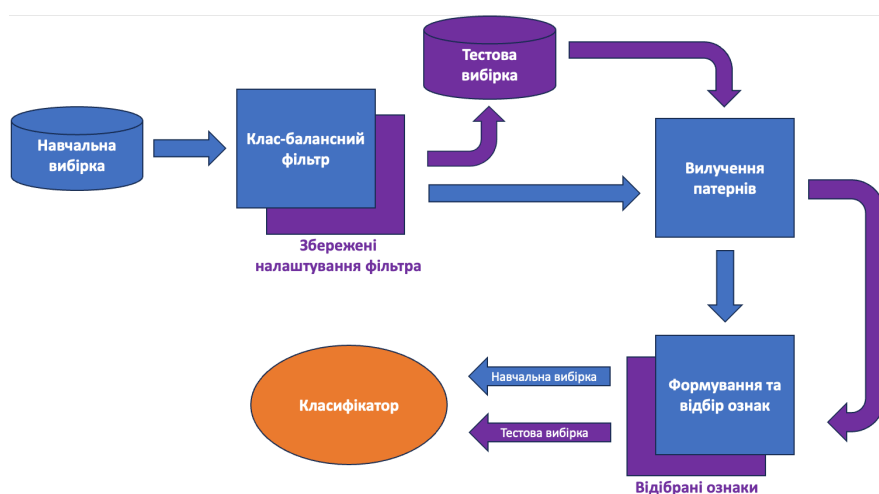


Рисунок 5.2 — Загальний принцип ізольованої обробки навчальної та тестової вибірки

Розбиття вибірки здійснювалося вручну на рівні незалежних піднаборів, при цьому контроль незалежності виконувався на рівні пацієнтів.

5.3.1 Етап 1. Фільтрація зображень

Етап фільтрації за допомогою клас-балансного фільтра складався з двох послідовних частин. На першому кроці навчальний набір подавався на фільтр, після чого виконувалася стандартна процедура, описана у розділі 2. Окрім збереження проміжних частотних представлень, окремо фіксувалися конфігураційні параметри, зокрема рівні квантування та порогові значення інтенсивності для відсіювання.

На рисунку 5.3 показано агреговані розподіли інтенсивностей на рівні класів для тестових даних після операції клас-балансної фільтрації у порівнянні з такими ж розподілами навчальному набору.

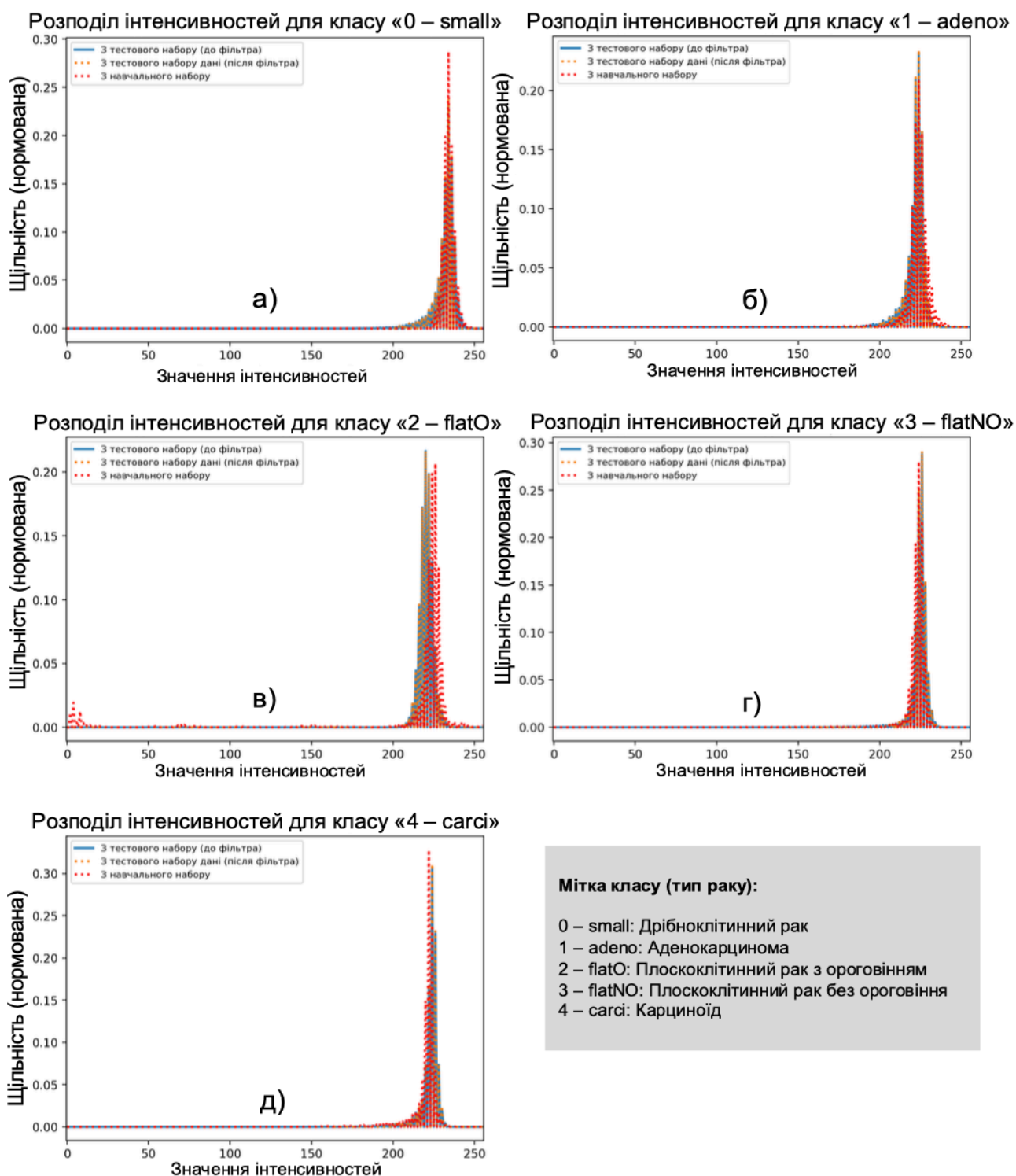


Рисунок 5.3 — Агрегований розподіл інтенсивностей для класів з тестової вибірки після фільтрації. а) для дрібноклітинного раку; б) для аденокарциноми; в) для плоскоклітинного раку з ороговінням; г) для плоскоклітинного раку без ороговіння; д) для карциноїда

У результаті роботи фільтра було отримано фіксований глобальний набір допустимих інтенсивностей у діапазоні 200–240 із кроком 2 (21 значення), причому

цей самий набір застосовувався однаково для всіх класів. Окрім цього, переносилися параметри побудови частотного представлення, що забезпечує відтворюваність формування вихідних частотних карт і узгодженість статистик між навчальною та тестовою вибірками.

Другий крок передбачав застосування тих самих, зафіксованих на навчальній вибірці параметрів до тестового набору без будь-якого додаткового підлаштування. У концептуальному сенсі це означає, що тестові дані проходять ідентичну за формою та суворо визначену попередню обробку, але без участі тестової статистики у виборі порогів чи шкали.

Можна спостерігати (рисунок 5.3), що фільтр приглушує частину сигналу, який не був представлений у навчальному наборі, при цьому загалом не змінює оригінальний розподіл частот. Для всіх класів тестового набору форма розподілу, а отже і частотний профіль збігаються із навчальним, крім плоскоклітинного раку з ороговінням (2 - flatO). Для останнього це говорить про те, що його розподіл значень пікселів незначною мірою відрізняється, ймовірно через потрапляння частин інших тканин паренхіми, проте це не завжди говорить про поганий результат, оскільки на цьому етапі ще не було здійснено пошук та аналіз унікальних структур.

5.3.2 Етап 2. Пошук структурних патернів

Пошук структурних патернів виконувався незалежно для навчальної та тестової вибірок. Для кожного набору зображень здійснювалося послідовне сканування локальним вікном фіксованого розміру 2x2 з формуванням множини виявлених патернів та підрахунком їхніх частот. Неінформативні області, зокрема фрагменти, що складаються виключно з нульових значень яскравості пікселів, виключалися з аналізу. Для кожного набору формувалася словник патернів та відповідні частотні характеристики. Ідентифікація патернів у даному підході здійснюється через канонічне представлення, тому структури однакової форми та з однаковими дискретизованими значеннями в обох наборах даних отримують ідентичний код. Це забезпечує уніфіковане зіставлення ознак між навчальною та тестовою вибірками навіть за умови незалежного виконання етапу пошуку, за

умови що до обох наборів застосовано однакову схему попередньої обробки інтенсивностей.

У результаті роботи алгоритму вилучення структурних патернів загальна їх кількість суттєво відрізнялася як між класами, так і між вибірками (табл. 5.6).

Таблиця 5.6 — Кількість вилічених патернів для кожного класу

Клас	Кількість унікальних патернів (шт.)	
	З навчальної вибірки	З тестової вибірки
0 - small	2265	5584
1 - adeno	5713	1275
2 - flatO	4058	2740
3 - flatNO	2321	2712
4 - carci	1841	1250

Такі відмінності узгоджуються з тим, що різноманіття патернів визначається не лише кількістю зображень у вибірці, а й структурою частотних карт інтенсивностей у межах області інтересу. Виявлені структури відображають локальні конфігурації частотних рівнів інтенсивності, відповідно, зростання просторової неоднорідності (наявність локальних переходів між рівнями частоти, контрастних зон та різної щільності представленості інтенсивностей) приводить до появи більшої кількості унікальних локальних конфігурацій навіть за меншої кількості зразків.

Для кожного класу було сформовано профіль розподілу відносних частот локальних структурних патернів, нормування виконувалося окремо в межах відповідного класу та окремо для кожної вибірки. Розрахунки проводилися у спільному просторі ознак, тобто з використанням єдиного упорядкування патернів для навчальної й тестової підвибірок. Така уніфікація забезпечує коректність безпосереднього порівняння та накладання профілів між вибірками, оскільки кожна позиція на осі відповідає тій самій структурі (агрегованій групі структур) незалежно від джерела даних. Для зменшення візуальної складності виконано дискретизацію шляхом агрегації по 5 патернів у один бін які відсортовані за частотою спадання, наведено перші 250 (рисунок 5.4). У всіх класів спостерігається різко асиметричний розподіл.

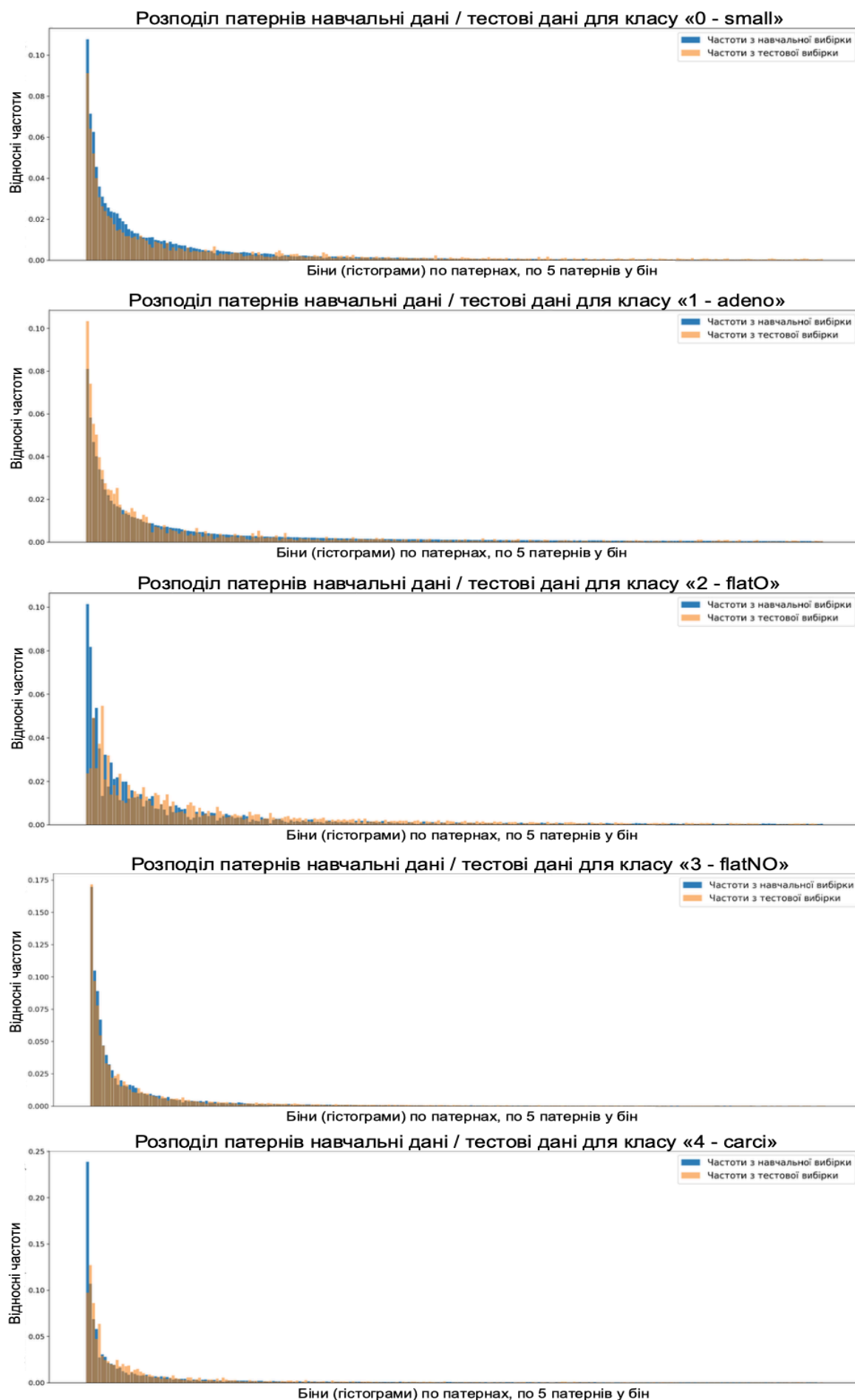


Рисунок 5.4 — Розподіли частот патернів у навчальній і тестовій вибірках

Невелика кількість найбільш частих бінів акумулює суттєву частку маси розподілу, тоді як велика кількість рідкісних з дуже малими відносними частотами формує довгий хвіст. Інтерпретація на рівні окремих рідкісних подій є нестійкою, тому в цьому аналізі вони не розглядаються

Ступінь узгодженості між вибірками дещо відрізняється між класами. Для класів дрібноклітинного раку (0 – small) та аденокарциноми (1 – adeno) загальна форма профілів між вибірками є близькою, що свідчить про подібність домінантних структур. Водночас для 0 – small у навчальній вибірці спостерігається відносно вищий початковий пік, що вказує на сильніше домінування обмеженої групи найчастіших патернів порівняно з тестовою підвибіркою. Для 1 – adeno тенденція є протилежною, оскільки більш виражена концентрація частоти у початку розподілу проявляється у тестових даних.

Для плоскоклітинного раку з ороговінням (2 - flatO) помітні найвиразніші відмінності, навчальний набір містить більш різкі максимуми, тоді як тестовий демонструє більш рівномірний розподіл у тій самій області. Втім словник патернів схожий, різниця лише в іншій концентрації на найчастіших структурах. Клас плоскоклітинного раку без ороговіння (3 - flatNO) показує загальну подібність між двома профілями.

Для карциноїда (4 - sarci) навчальні дані містять надмірно домінуючі структури які зустрічаються найчастіше, тоді як у тестових частоти розподіляються між ширшим набором.

Окремо було досліджено структурне різноманіття з урахуванням впливу розміру області інтересу на кількість унікальних патернів.

$$d = \frac{U}{M_{ROI}} \quad (5.1)$$

де d — щільність унікальних патернів на 1 мільйон пікселів, U — кількість унікальних патернів, M_{ROI} — площа інформативної області, виражена у мегапікселях,

Показник відображає щільність унікальних локальних патернів на одиницю інформативної площі, а саме на 1 мільйон пікселів ненульових значень. Аналіз чисел показує суттєву асиметрію між вибірками у деяких класах (табл. 5.7).

Таблиця 5.7 — Щільність патернів на одиницю інформативної площі

Клас	З навчальної вибірки	З тестової вибірки
0 - small	≈11458.6	≈81460.0
1 - adeno	≈81366.7	≈210326.6
2 - flatO	≈72496.7	≈60230.4
3 - flatNO	≈34076.7	≈48068.9
4 - carci	≈90333.7	≈124032.5

У класах 0 - small, 1 - adeno, 3 - flatNO та 4 - carci зафіксовано значне зростання різноманіття патернів у тестовій вибірці порівняно з навчальною. Найбільш виражений розрив спостерігається у 1 - adeno, що вказує на вищу локальну варіативність та структурну гетерогенність ознак у тестовій вибірці навіть за умови меншої кількості зображень і свідчить про розширення спектра ознак на одиницю площі. Натомість для класу 2 - flatO виявлено зворотну тенденцію. Загалом, виявлена розбіжність свідчить про переважання структурної складності у більшості сегментів тестової вибірки відносно навчальної.

Для оцінювання клас-специфічності локальних структур було проаналізовано відносну присутність патернів у межах різних класів. На відміну від підходу, в якому інформативність патерна визначають за його відотною частотою в межах одного класу, у цьому випадку акцент зроблено на оцінюванні умовної ймовірності його належності до класу. Така інтерпретація дозволяє відокремити ефект загальної поширеності структури від його діагностичної цінності. Патерни, для яких маса розподілу концентрується переважно в одному класі, можуть розглядатися як потенційні маркери, а ті що є близькими до рівномірного розподілу є неспецифічними і, відповідно, менш дискримінативними.

Для зменшення візуальної складності виконано дискретизацію шляхом агрегації по 10 патернів у бін. Для подальшого аналізу враховувалися лише перші 250 найбільш представлених бінів, тоді як решта розглядалася як сукупність

рідкісних конфігурацій з надто малими частотами появи, що обмежує їх інтерпретованість. Для навчальної та тестової вибірок використовувався єдиний узгоджений простір патернів (спільний порядок бінів), завдяки чому оцінені внески класів у межах кожного біна є безпосередньо співставними (рисунок 5.5).

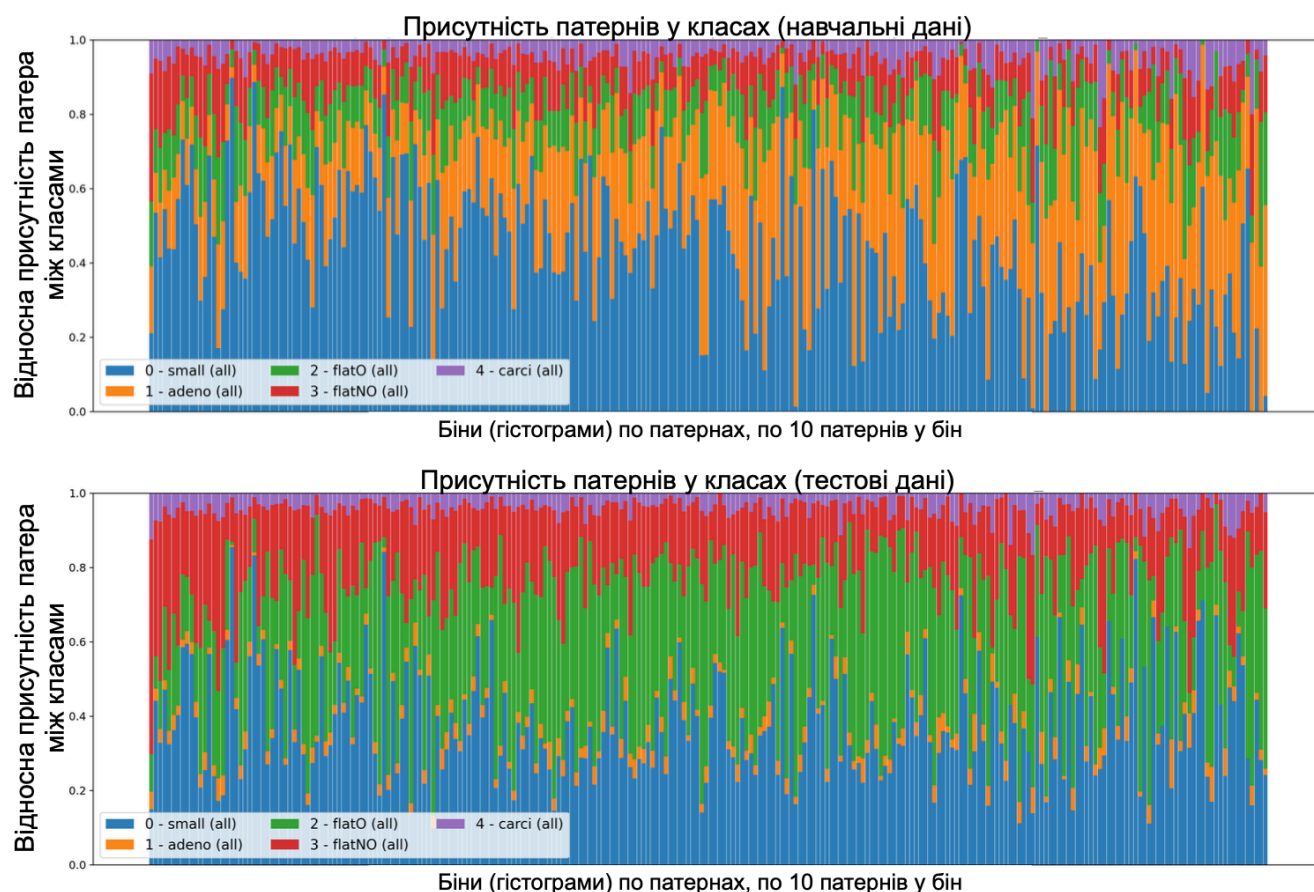


Рисунок 5.5 — Порівняння відносної присутності патернів у класах між навчальною та тестовою вибіркою

Отриманий результат показує, що простір локальних структур має змішану природу. Значна частина патернів є загальними і відображає типові конфігурації частотних рівнів на частотних картах інтенсивностей, які не прив'язані до одного класу, тоді як інша частина демонструє підвищену асоціацію з конкретним класом.

Порівняння вибірок додатково виявляє, що для окремих бінів співвідношення внесків класів змінюється. Один і той самий патерн може бути відносно ближчим до певного класу в навчальній вибірці і менш вираженим у тестовій, або навпаки. Це вказує, що частина патернів є чутливою до відмінностей у складі даних і

характеристиках області інтересу, тому дискримінативність таких ознак повинна оцінюватися не лише на навчальному наборі, а й підтверджуватися на незалежних даних. Наступним кроком є перевірка, чи дозволяє наявність і/або відносні частоти цих патернів сформуванню простору ознак, достатній для розділення класів.

5.3.3 Етап 3. Виокремлення клас-специфічних патернів

Механізм привласнення патернам домінантного класу та побудови масок (формула 3.12) $T_{(c,k)}$ є інформативним з точки зору інтерпретації, однак його пряме використання для формування ознак може призвести до так званого витoku інформації. Це трапляється тоді, коли під час побудови простору ознак для конкретного зразка використовується його істинна мітка класу, наприклад, для занулення нехарактерних патернів саме для цього класу. У такому випадку ознаки починають імпліцитно містити інформацію про цільову змінну, яка в реальному сценарії інференсу недоступна. Додатково витік виникає, якщо статистики для формування відбору патернів оцінюються на всьому наборі до розбиття на тренувальну й тестову частини, оскільки критерії відбору стають залежними від розподілу тестових даних.

Тому всі елементи, що визначають простір ознак і параметри перетворень, мають оцінюватися лише на тренувальній підмножині та застосовуватися до тестової як фіксована конфігурація без залучення її міток.

Щоб унеможливити описані джерела витoku, у роботі застосовано двоетапну процедуру відбору та перетворення ознак: 1) за навчальною вибіркою визначається підмножина інформативних патернів; 2) ці самі налаштування без змін застосовують до тестової вибірки.

5.3.3.1 Частотний критерій

Використовуючи навчальний набір зразків обчислюється сумарна частота патернів у межах класів:

$$F_{c,k} = \sum_{s \in S_{tr}: y_s = c} f_{s,k} \quad (5.2)$$

де $F_{c,k}$ — сумарна частота патерна k у класі c у наборі тренувальних даних, $c \in C$ — клас, $k \in K$ — індекс патерна, S_{tr} — множина навчальних зразків, s — індекс зразка, y_s — мітка класу для зразка s , $f_{s,k}$ — частота появи патерна k у зразку s .

За цим може бути визначене призначення домінантного патерна для певного класу:

$$c'(k) = \arg \max_{c \in C} F_{c,k} \quad (5.3)$$

де $c'(k)$ — домінантний клас для патерна k за агрегованою частотою.

Замість клас-специфічного занулення патернів у кожному зразку формується фіксований набір $K_{keep} \subset K$, який визначає спільний простір ознак для навчального й тестового набору. На практиці K_{keep} задається як об'єднання двох критеріїв, кожен з яких захоплює різний тип інформації про патерни. Критерії виконують дві взаємодоповнювальні функції: по-перше, вони зменшують розмірність простору ознак, відсікаючи рідкісні або випадкові патерни, що не несуть стабільної диференційної інформації; по-друге, вони забезпечують, що у фінальному наборі залишаються як патерни, які є типовими для кожного класу, так і патерни, які найкраще розрізняють класи між собою. Важливо зазначити, що результати відбору фіксуються на навчальному наборі і надалі не змінюються, тому тестові зразки не впливають на склад простору ознак, а лише проектуються у вже визначені координати K_{keep} .

Відбір за найбільшими $F_{c,k}$ націлений на патерни, які є типовими для класу в сенсі сумарної частоти по його присутності:

$$K_{top} = \bigcup_{c \in C} TopK_k(F_{c,k}) \quad (5.4)$$

де K_{top} — множина патернів, відібраних за частотністю, $TopK_k(\cdot)$ — оператор, що повертає K патернів k з найбільшими значеннями аргументів $F_{c,k}$ в межах фіксованого класу c . Критерій фіксує патерни, що формують ядро класу, вони або

часто з'являються у багатьох зразках класу, або мають високі частоти в його характерних зразках. Це особливо важливо для задачі з неоднорідними проявами патології, де частина клас-специфічних структур проявляється стабільно в більшості прикладів.

5.3.3.2 Дискримінативний критерій

Даний критерій відокремлює патерни, які найбільш нерівномірно розподілені між класами за самим фактом появи. На відміну від частотного критерію, він менш чутливий до абсолютних значень частоти й більше фокусується на факті появи патерна у класі. Така постановка є корисною, коли певні морфологічні або текстурні структури трапляються не дуже часто, але їхня поява є сильним маркером конкретного класу. Вводиться бінарна присутність:

$$b_{s,k} = \begin{cases} 1, & \text{якщо } f_{s,k} > 0 \\ 0, & \text{якщо } f_{s,k} = 0 \end{cases} \quad (5.5)$$

де $b_{s,k} \in \{0,1\}$ — ознака присутності патерна k у зразку s , $f_{s,k}$ — частота патерна k у зразку s . Далі для кожного патерна порівнюються спостережувані та очікувані кількості “присутній/відсутній” у класах через χ^2 -статистику:

$$\chi_k^2 = \sum_{c \in C} \frac{(o_{c,k} - e_{c,k})^2}{e_{c,k} + \varepsilon} + \sum_{c \in C} \frac{(\bar{o}_{c,k} - \bar{e}_{c,k})^2}{\bar{e}_{c,k} + \varepsilon} \quad (5.6)$$

де $\chi_k^2 - \chi^2$ — статистика для патерна k , $o_{c,k}$ — спостережувана кількість навчальних зразків класу c , у яких патерн k присутній, $e_{c,k}$ — очікувана кількість таких присутностей для класу c за гіпотезою незалежності, $\bar{o}_{c,k}$ — спостережувана кількість відсутностей, $\bar{e}_{c,k}$ — очікувана кількість відсутностей за гіпотезою незалежності. $\varepsilon > 0$ — константа для уникнення ділення на нуль.

Критерій відображає розрізнявальну здатність патерна. Це підсилює структури, що можуть бути не обов'язково найчастішими, проте з'являються переважно в певних класах і майже не з'являються в інших.

Фінальний набір визначається як:

$$K_{keep} = K_{top} \cup K_{\chi} \quad (5.7)$$

Принцип побудови ознак лишається без змін (формули 3.14-3.21). Відмінність лише в тому, що для кожного зразка s розглядаються частоти лише з K_{keep} .

5.3.3.3 Подібність до класів за прототипами

Крім локальних ознак по окремих патернах, також додається узагальнювальна група ознак, що відображає схожість зразка до характерного середнього профілю унікальних структур кожного класу. Мотивація цього кроку полягає в тому, що зразки можуть бути подібними не за окремими патернами, а за структурою їхнього розподілу у просторі K_{keep} . Такі ознаки стискають високовимірний вектор у невеликий набір значень (по одному на клас) та можуть підвищувати стійкість класифікації за розріджених або шумних патернових представлень, виступаючи узагальненим профільним індикатором належності зразка до класу.

Для цього на наборі навчальних даних для кожного класу c будується прототип як середній вектор відносних частот:

$$prot_c = \frac{1}{|S_{tr}^{(c)}|} \sum_{s \in S, y_s = c} P_s \quad (5.8)$$

де $prot_c$ – середній профіль класу c у просторі патернів, $S_{tr}^{(c)} = \{s \in S, y_s = c\}$ – підмножина навчальних зразків класу c , $|S_{tr}^{(c)}|$ – кількість таких зразків, P_s – вектор відносних частот зразка s з компонентами $p_{s,k}$ (формула відносних частот).

Далі для будь-якого зразка обчислюється косинусна подібність:

$$cossim(s, c) = \frac{P_s^T prot_c}{||P_s||_2 ||prot_c||_2 + \varepsilon} \quad (5.9)$$

де $\text{cossim}(s, c)$ – косинусна подібність зразка s до прототипу класу c , $P_s^T \text{prot}_c$ – скалярний добуток (транспонована T), $\|P_s\|_2 \| \text{prot}_c \|_2$ – евклідова норма вектора, $\varepsilon > 0$ – константа для запобігання ділення на нуль, P_s – вектор відносних частот зразка s . Величина $\text{cossim}(s, c)$ відображає, наскільки напрямок P_s узгоджується з прототипом класу c , і тому є інформативною величиною навіть за різних абсолютних масштабів частот.

На етапі налаштування (набір навчальних даних) формується конфігурація:

$$\Theta = (K_{\text{keep}}, \{idf_k\}, \{\mu_k, \sigma_k\}, \{\text{prot}_c\}) \quad (5.10)$$

де Θ – набір параметрів, що зберігається після етапу налаштування і застосовується на набір тестових даних, K_{keep} – відібраний простір патернів, $\{idf_k\}$ – множина (вектор) IDF-ваг для всіх $k \in K_{\text{keep}}$, $\{\mu_k, \sigma_k\}$ – параметри стандартизації лог-частот для всіх $k \in K_{\text{keep}}$, $\{\text{prot}_c\}$ – набір прототипів для всіх $c \in C$. Таким чином на етапі застосування тестові зразки відображаються у той самий простір ознак, а всі параметри перетворень обчислюються з тренувального набору, без залучення міток тестових даних.

5.3.4 Результати класифікації для незалежно оброблених вибірок

Перш за все було перевірено чи відтворюється класова спрямованість для структур у незалежних даних. Для обмеження розмірності простору ознак для кожного класу відбиралися 50 найбільш представлених патернів для кожного класу за частотним критерієм. За χ^2 -критерієм відбиралися 100 патернів з найвищою статистичною асоціацією з класом, обчисленим на бінарній ознаці присутності або відсутності патерна у зразку. Відібрані структури об'єднувалися в єдиний набір ознак, після чого для кожного зразка формувався вектор частот у фіксованому спільному просторі, відсутні кодувалися нульовими значеннями. У результаті відбору ознак було сформовано підмножину з 157 патернів та перевірено, наскільки їхня відносна асоціація з класами зберігається при переході від навчальних до тестових даних (рисунок 5.6).

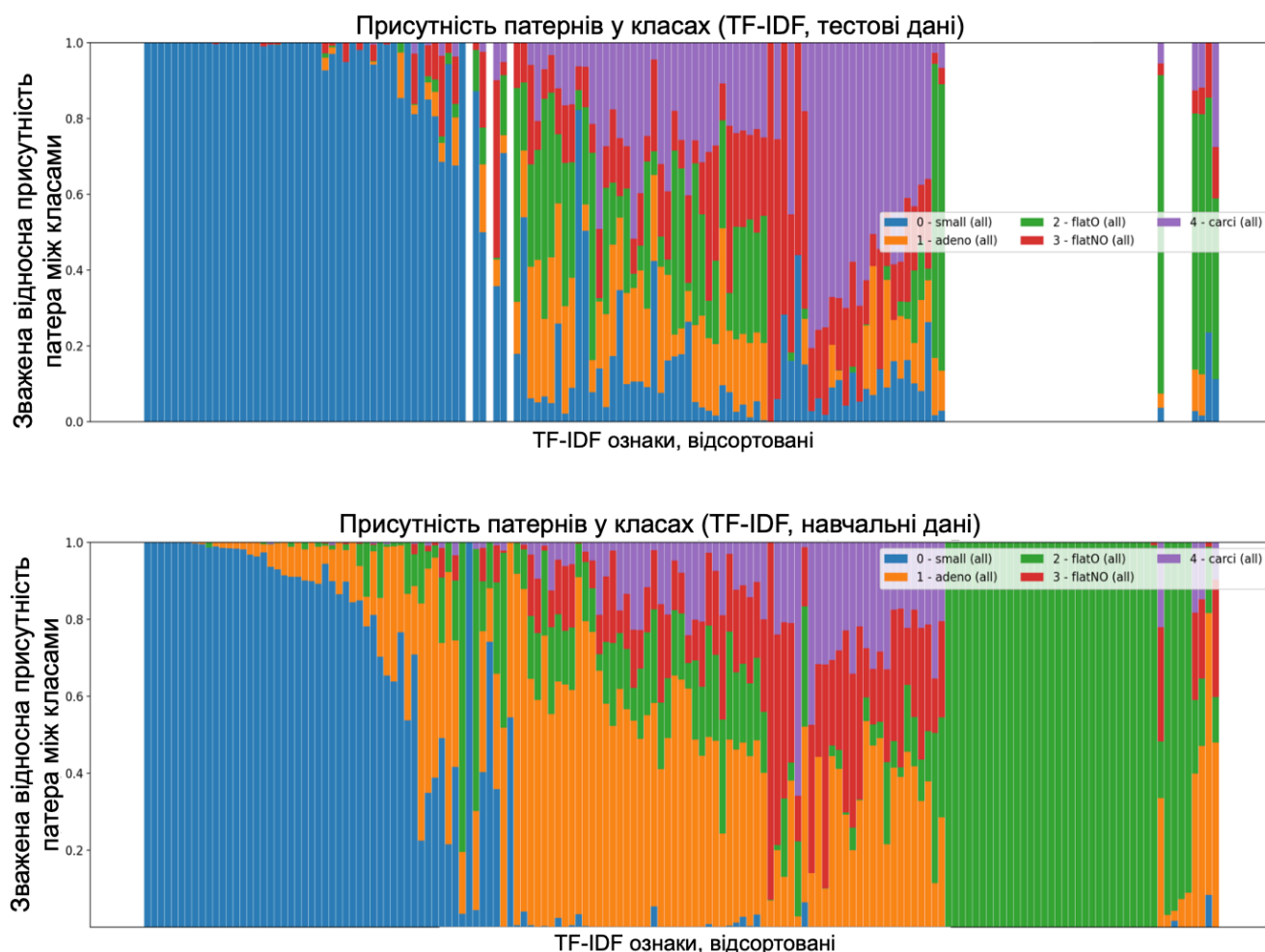


Рисунок 5.6 — Відносна присутність патернів у різних класах між навчальною та тестовою вибірками

Зіставлення вибірок показує, що частина патернів демонструє узгоджену поведінку між навчальною та тестовою вибірками. Їх класова диференційна вираженість відтворюється, що вказує на наявність відносно стійких морфологічних конфігурацій, притаманних певним типам патології, та підтримує припущення про їхню потенційну дискримінативну цінність. Водночас помітна частка ознак виявляє розбіжності між вибірками, причому ці відмінності мають кілька характерних проявів. Для окремих патернів знижується вираженість домінантної асоціації з певним класом: у тестовій вибірці їхній внесок розподіляється між кількома класами більш рівномірно, що зменшує специфічність. Також спостерігаються випадки, коли ті самі конфігурації в різних вибірках асоціюються з різними класами, що є ознакою чутливості до складу даних

та умов формування області інтересу. Відповідно, практично значущим висновком є те, що підмножина містить як стабільні, потенційно узагальнювані сигнали, так і ознаки, чия інформативність є залежною від вибірки.

Як класифікатори використано низку моделей аналогічний набір описаний у підрозділі 5.1. Найкращий загальний результат на тестовій вибірці продемонструвала мультиноміальна логістична регресія, яка забезпечила точність на тестових даних у 80.5% та макро-F1 76.6% при зваженій F1 (на тестових даних) 78.4%. У сукупності це забезпечує практично значущий баланс де модель не лише демонструє найвищі тестові агреговані метрики серед розглянутих підходів, але й зберігає відносно помірний розрив між показниками якості при навчанні та тесті, що узгоджується з більшою стійкістю до перенавчання (табл. 5.8).

Таблиця 5.8 — Результати класифікації на тестовій вибірці за умов ізольованої обробки даних

Модель	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1 (Test)	Weighted F1 (Train)	Overfit Gap
Логістична регресія	0.805	0.821	0.794	0.766	0.784	0.981	0.197
Стекінг	0.789	0.798	0.782	0.761	0.780	1.000	0.220
Каскадний OvR	0.789	0.831	0.775	0.743	0.762	0.992	0.230
LightGBM	0.545	0.681	0.570	0.543	0.544	1.000	0.456
Bagging	0.537	0.585	0.567	0.504	0.493	1.000	0.507
Gradient Boosting	0.504	0.714	0.534	0.453	0.446	1.000	0.554
AdaBoost	0.472	0.343	0.505	0.386	0.376	0.967	0.591
XGBoost	0.382	0.511	0.416	0.379	0.369	1.000	0.631
Випадковий ліс	0.447	0.370	0.485	0.377	0.368	1.000	0.632

Модель за стратегією каскадного «один проти решти» базується на послідовному виключенні класів, що робить результат чутливим до їхнього порядку. Для усунення суб'єктивності вибору порядку в роботі реалізовано автоматичний підбір найкращої послідовності. У підсумку модель досягає точності на тестових даних з результатом 78.9% й демонструє якість, близьку до найкращих результатів, хоча й поступається мультиноміальній логістичній регресії за

інтегральними показниками. Каскадна стратегія побудови моделі є методологічно значущою, оскільки демонструє конкурентний рівень узагальнення та надає механізм явного контролю за пріоритетами через оптимізацію порядку та порогів.

Результати інших моделей підтверджують, що для заданого представлення та обсягу незалежних даних перевага переважно залишається за лінійними або помірно складними підходами. Матриці відповідності для кожної з моделей продемонстровані на рисунку 5.7.

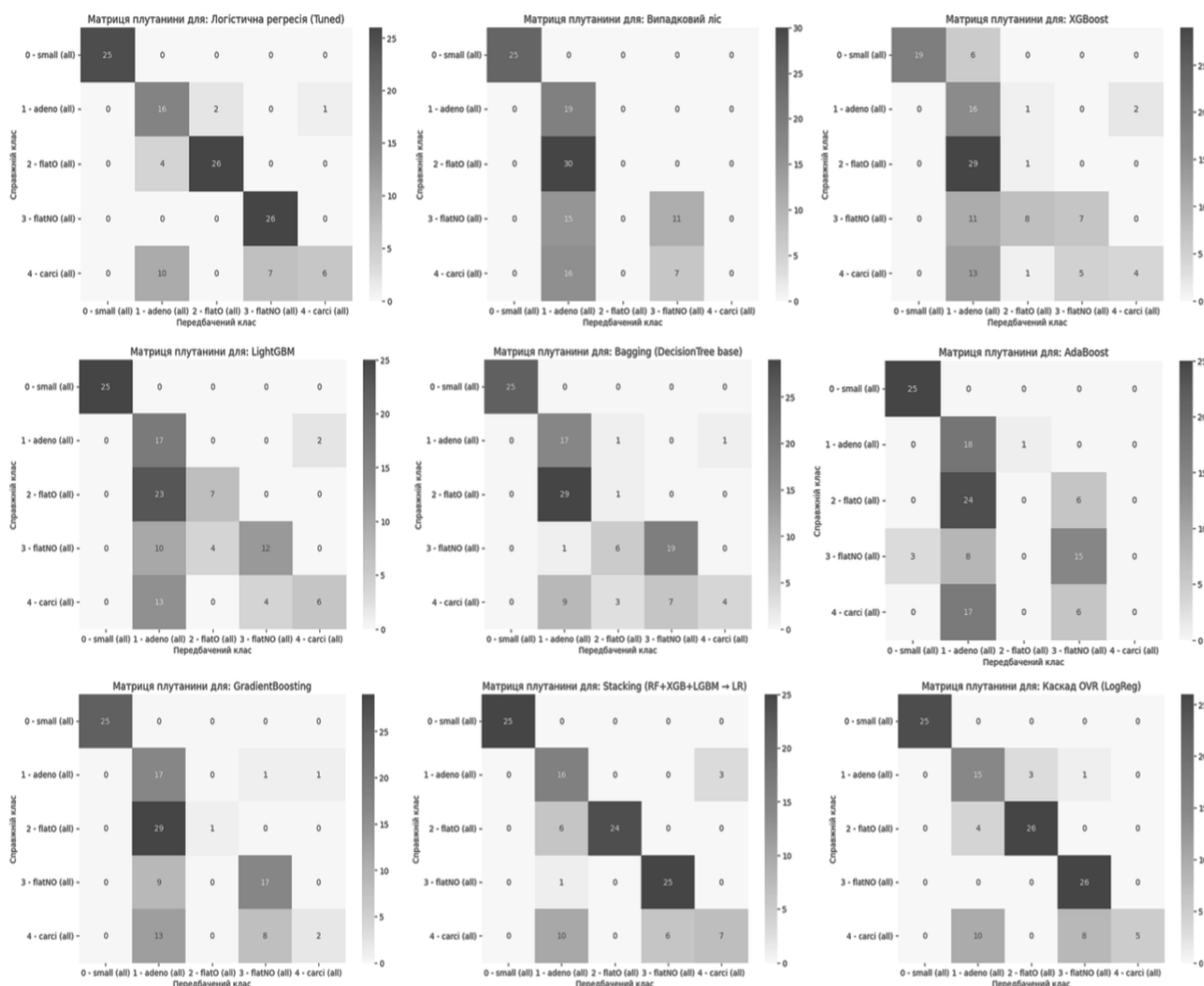


Рисунок 5.7 — Матриці відповідності для моделей

Стекінг забезпечив близьку якість до каскадного «один проти решти», однак демонстрував майже ідеальну якість на навчальній вибірці, що разом із підвищеним розривом вказує на ризик перенавчання в цій постановці. Серед

градієнтних й методів на основі дерев рішень найкращим з усіх виявився LightGBM, тоді як RandomForest, XGBoost, GradientBoosting, Bagging та AdaBoost показали суттєво нижчу якість на тестових даних при дуже високих навчальних метриках, що є типовим проявом перенавчання в умовах високої розмірності та обмеженої кількості даних. Сукупно це підкреслює, що в межах поточного дизайну експерименту та сформованого простору ознак найбільш обґрунтованим базовим вибором є логістична регресія з регуляризацією, тоді як стратегія каскадного «один проти решти» забезпечує конкурентну альтернативу з додатковою керованістю, придатну для задач, у яких важливо цілеспрямовано оптимізувати поведінку щодо окремих класів.

5.4 Практичне використання класифікаторів

У результаті виконання роботи побудовано та навчено 9 класифікаційних моделі, які дозволяють за набором вхідних даних пацієнта визначити його належність до класу дрібноклітинного раку, аденокарциноми, плоскоклітинного раку з ороговінням, плоскоклітинного раку без ороговіння, карциноїда.

Було реалізовано програмне забезпечення, що реалізує описані у роботі підходи до фільтрації зображень, пошуку патернів та автоматичної класифікації у вигляді послідовності кроків.

Програмну реалізацію виконано мовою Python із застосуванням фреймворку Streamlit для створення вебдодатка та бібліотеки pydicom для роботи з форматом медичних зображень (DICOM). Для застосування необхідно отримати дані комп'ютерної томографії легень пацієнтів та імпортувати їх до робочого середовища.

Підтримується медичний формат зображень DICOM, а також звичайні формати зображень, такі як PNG або TIFF. Підтримується завантаження як одного зображення так і серії. При завантаженні медичних зображень у форматі DICOM є можливість переглянути їх, налаштувати режим відображення легеневого вікна або задати значення ширини та рівня вікна вручну за HU шкалою.

Оскільки класифікатори можуть працювати лише за наявності виділеної області інтересу (ROI), програмне забезпечення надає можливість виділити зону інтересу вручну інструментами пензля або прямокутника. Виділені області інтересу конвертуються у бінарні маски, накладаються на оригінальні зображення та автоматично конвертуються у 8-бітні зображення придатні для подальших кроків аналізу.

Розмічені зображення у такому форматі можна одразу передати на етап фільтрації (рисунк 5.8).

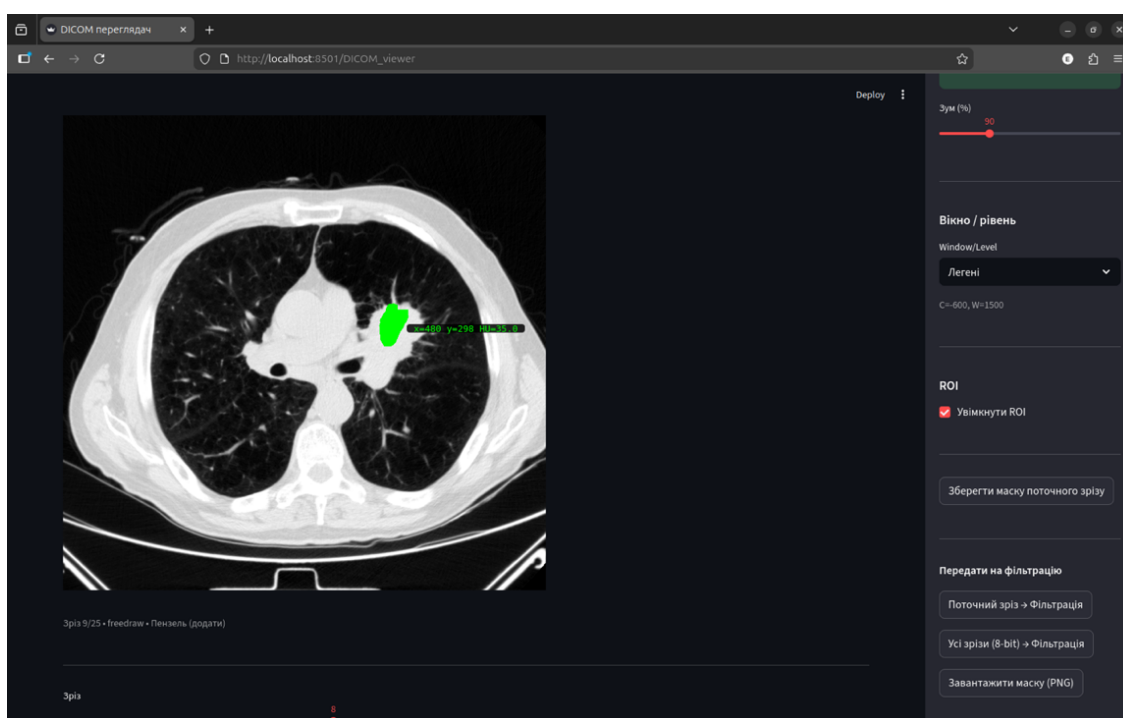


Рисунок 5.8 — Інтерфейс ручного виділення області інтересу на КТ-зображенні

Крок фільтрації можна застосувати як для одного зображення, так і для всієї серії. При цьому, якщо завантажуються серія, розподіл інтенсивностей пікселів усіх зразків автоматично нормуються та усереднюються за принципом, що був описаний у розділі 2. Для застосування фільтра слід завантажити конфігураційний файл, у якому зафіксовано параметри фільтрації, визначені за результатами аналізу навчальної вибірки.

Залежно від обраного файлу до завантажених знімків застосовуються відповідні параметри обробки. (рисунк 5.9).

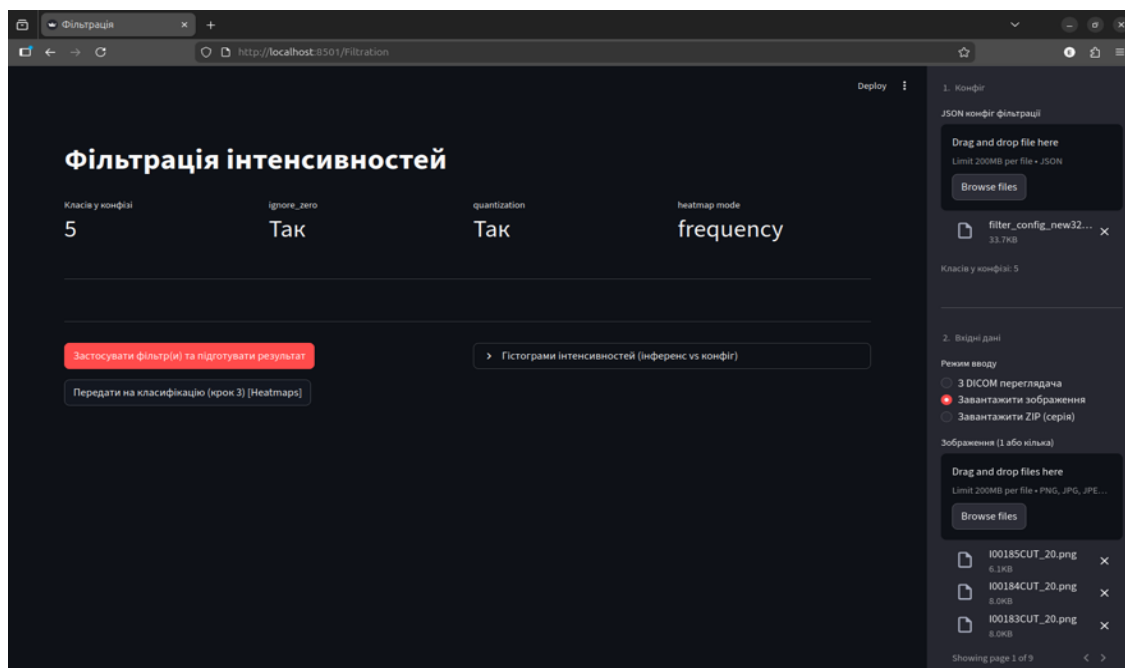


Рисунок 5.9 — Інтерфейс завантаження та застосування конфігураційного файлу клас-балансного фільтра

Додатково є можливість переглянути та порівняти розподіли частот інтенсивностей пікселів завантажених зображень з розподілами конфігураційного файлу (рисунок 5.10).

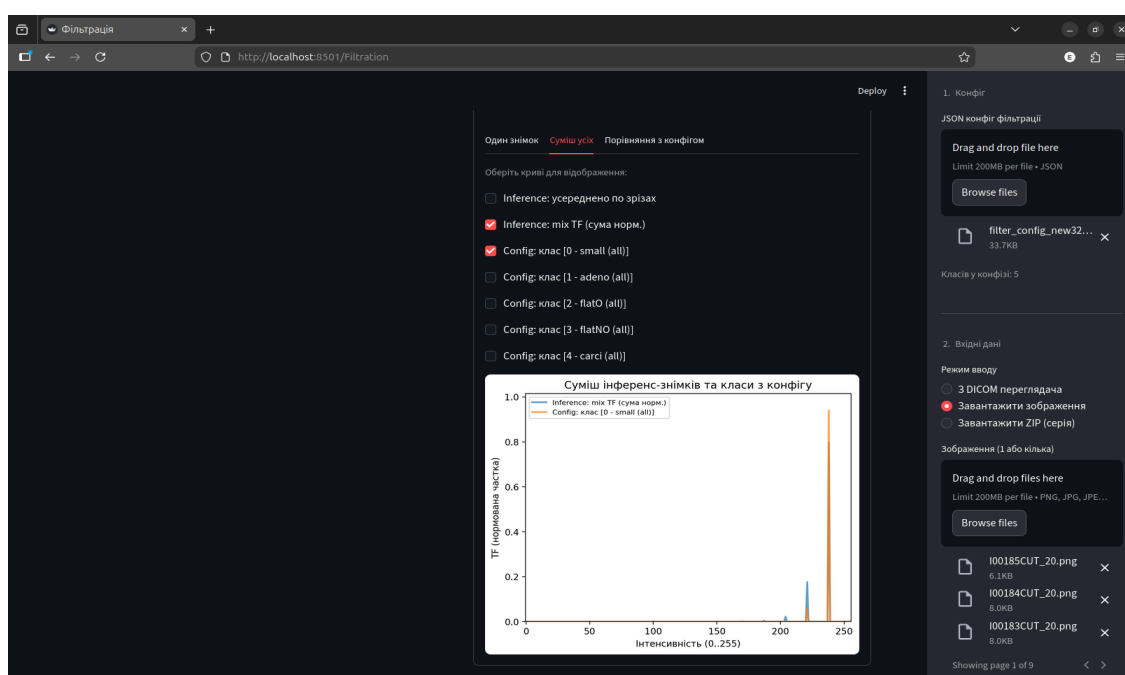


Рисунок 5.10 — Візуалізація порівняння завантажених даних із профілем конфігурації фільтра

За потреби надається можливість застосувати операцію квантування, та налаштування режимів відображення й накладання різних профілів розподілу інтенсивностей. Відфільтровані зображення за потреби можна одразу передати на наступний крок для класифікації, або завантажити їх окремо на відповідній вкладці.

На етапі класифікації, програмне забезпечення автоматично почне процедуру пошуку патернів на одному зображенні або серії (рисунок 5.11).

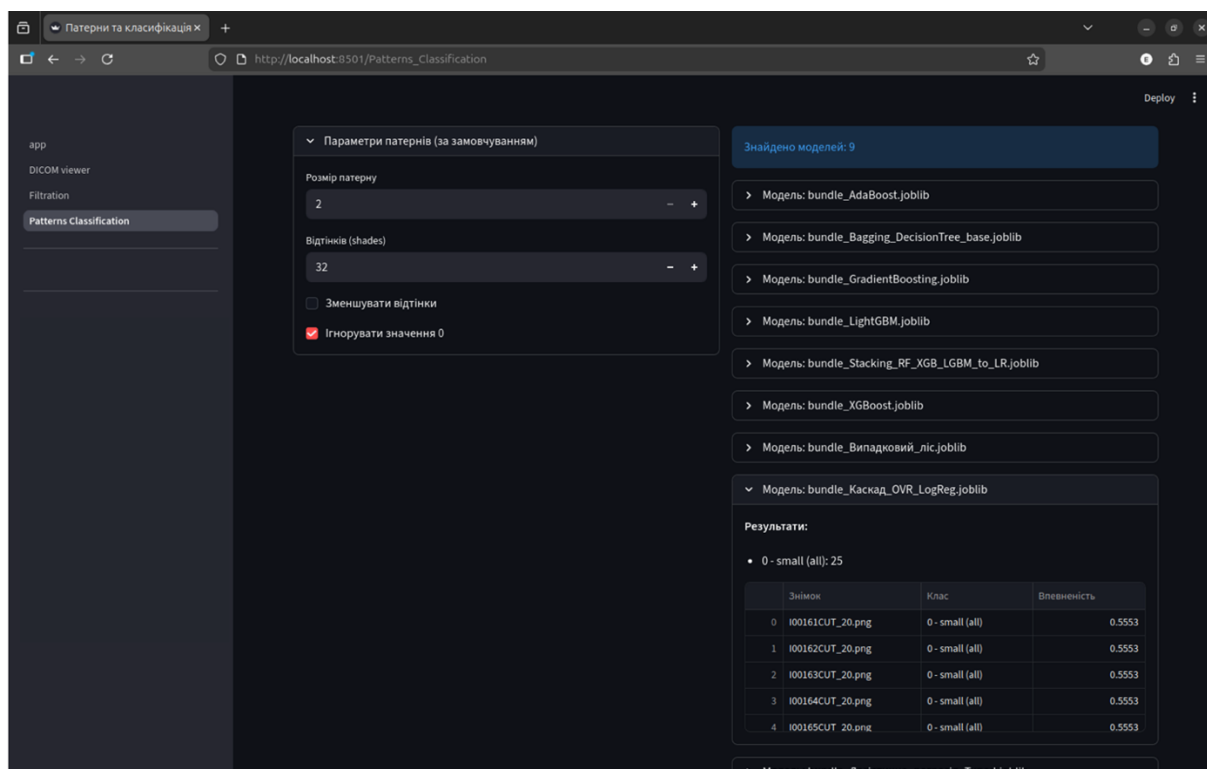


Рисунок 5.11 — Вікно результатів класифікації

За необхідності, надається можливість змінювати деякі параметри пошуку патернів, проте за замовчуванням встановлюються ті самі налаштування, що були використані на етапі обробки навчальної вибірки даних. Отриманий профіль виділених структур проходить етап перетворення на числові ознаки за тим самим принципом, що описано у підрозділі 3.3.

Усі попередньо навчені моделі класифікації автоматично завантажуються із заздалегідь збережених файлів. За потреби перелік моделей може бути розширено в подальшому шляхом додавання нових екземплярів. Результати класифікації

формуються окремо для кожного зразка всіма доступними моделями та подаються разом із оцінкою ймовірності (міри впевненості) відповідного рішення.

В якості подальшого розвитку розглядаються шляхи спрощеного режиму роботи програмного забезпечення із редукцією кількості параметрів налаштування для спрощеного користування. Додатково можлива реалізація відображення у зручній та інформативній формі розподілу знайдених на зразках інференсу патернів та їх співставного порівняння із бібліотекою структур, що були вилучені на етапі навчання.

Висновок до розділу 5

Застосування підходу опису КТ-зображень на основі системи локальних структурних патернів дозволяє отримати високу якість багатокласової класифікації гістологічних типів раку легень. При цьому надійність отриманих результатів суттєво залежить від узгодженості етапів попередньої обробки та формування ознак, зокрема від вибору параметрів квантування й фільтрації інтенсивностей, які визначають співвідношення між стійкістю ознак і збереженням дискримінативної інформації.

Встановлено вплив параметрів фільтрації та режимів квантування на якість класифікації, порівняно альтернативні способи подання ознак і архітектури моделей та визначено найбільш придатну конфігурацію, а також підтверджено стійкість отриманих результатів за різних умов обробки даних.

Порівняння альтернативних способів подання частот патернів показало, що TF-IDF забезпечує більш збалансовану якість розпізнавання між класами за рахунок пригнічення надмірно поширених компонент і підсилення рідкісних, потенційно більш специфічних ознак. Разом із тим, відносні частоти демонструють близькі результати, що свідчить про те, що основна дискримінативна інформація зосереджується у композиційних співвідношеннях патернів, а не лише у їх глобальній рідкісності.

Застосування класичних статистичних класифікаторів забезпечує інтерпретованість рішення через внесок окремих компонент простору ознак, що є важливим для подальшого аналізу причин помилок та характеристик міжкласового розмежування. Запропонована каскадна стратегія «один проти решти» додатково підвищує контроль над процесом прийняття рішення в умовах неоднакової складності розпізнавання класів, оскільки дозволяє адаптувати порядок перевірки та порогові значення для окремих груп. За результатами порівняльних експериментів запропонований підхід за обраних параметрів попередньої обробки забезпечив точність на рівні 0.949 ± 0.048 та макро-F1-міри 0.916 ± 0.074 . Додаткова перевірка в умовах ізольованої обробки навчальної та тестової вибірок підтвердила відтворюваність отриманих результатів і практичну придатність для сценарію застосування на нових даних. У додатковому, більш консервативному сценарії оцінювання отримано точність на рівні 0.805, макро-F1-міра 0.766.

Розроблено систему диференційної діагностики, що інтегрує попередню обробку КТ-зображень, формування патерн-орієнтованих ознак і модуль багатокласової класифікації.

ВИСНОВКИ

Метою роботи було підвищення точності та інтерпретованості автоматизованої диференційної діагностики гістологічних типів раку легень за даними комп'ютерної томографії шляхом розроблення методу багатокласової класифікації на основі локальних структурних патернів.

1. Виконано аналіз сучасного стану досліджень у галузі автоматизованого аналізу КТ-зображень легень методами машинного навчання, визначено ключові обмеження існуючих підходів у задачах диференційної діагностики онкологічних патологій, зокрема обмежену інтерпретованість моделей та нестачу даних для їх надійного навчання, обґрунтовано доцільність використання описів локальних структурних конфігурацій як компактної та інформативної основи для побудови класифікаційних моделей.

2. Досліджено особливості КТ-зображень і процедур їх попередньої обробки, що впливають на відтворюваність і точність автоматизованої класифікації, сформовано єдиний підхід до стандартизації інтенсивностей пікселів та забезпечення порівнюваності зображень між пацієнтами і діагностичними групами, а також визначено принципи ізольованої підготовки навчальної і тестової вибірок для уникнення упередженості оцінювання. У межах запропонованого підходу зображення проходить процедуру фільтрації й перетворюється у зважене представлення, у якому кожному відібраному пікселю призначається скалярна вага, що відображає частотну значущість його рівня інтенсивності у вибірці. Кількісна оцінка підтверджує посилення вибіркості фільтрації, зокрема середня частка збережених пікселів зменшилась на 23.27%, коефіцієнт Gini зріс з 0.00995 до 0.12549, а ентропійна міра знизилась на 4.45%. Це свідчить, що більша частка суми ваг припадає на меншу кількість пікселів, тому зберігаються переважно пікселі з найбільшими значеннями.

3. Розроблено метод формування ознак на основі локальних структурних характеристик КТ-зображень, що включає виділення повторюваних патернів та

побудову числового представлення спостережуваних структур у вигляді зіставного простору ознак, придатного для статистичного моделювання.

4. Проведено дослідження впливу різних методів формування комплексних ознак на їх здатність розділяти класи, здійснено порівняльний аналіз альтернативних способів подання частот патернів та показано, що вибір схеми кодування істотно визначає баланс між внеском домінантних і другорядних компонент, а також стабільність міжкласового розмежування за умов неоднорідності даних. За результатами порівняльного оцінювання встановлено, що подання ознак у відносній формі та при TF-IDF-зважуванні, забезпечує найкращу узагальнювальну якість моделі. У середньому за сукупністю проведених експериментальних сценаріїв оцінювання, отримано загальну точність 0.910 ± 0.084 та макро-F1-міру 0.859 ± 0.112 .

5. Розроблено метод багатокласової класифікації, що враховує різну складність розмежування діагностичних класів за обмеженого обсягу вибірки. Запропоновано стратегію каскадного «один проти решти» з послідовним відокремленням класів та механізмами керованості прийняття рішень, зокрема за рахунок автоматизованого добору порядку проходження класів та налаштування клас-специфічних порогів, що дає змогу оптимізувати загальну якість моделі з урахуванням пріоритетів щодо окремих класів. Експериментально підтверджено результативність запропонованої стратегії, для якої досягнуто загальної точності на рівні 0.895 ± 0.120 та значення макро-F1-міри 0.848 ± 0.160 .

6. На основі розроблених методів створено систему диференційної діагностики гістологічних типів раку легень за даними КТ, яка забезпечує застосування попередньо навчених моделей для автоматизованої класифікації та подання результатів у формі, придатній для підтримки прийняття рішень у практичних сценаріях. У базовому експериментальному сценарії як основному режимі застосування системи отримано загальну точність 0.949 ± 0.048 та макро-F1-міру 0.916 ± 0.074 , а додаткова контрольна перевірка у більш консервативному сценарії оцінювання показала загальну точність 0.805 та макро-F1-міру 0.766, що демонструє відтворюваність висновків за зміни умов експериментальної перевірки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Zhou J, Xu Y, Liu J, Feng L, Yu J, Chen D. Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN. *Cancer Epidemiology*. 2024; 93: 102693. doi: 10.1016/j.canep.2024.102693
2. Vynnychenko OI, Moskalenko RA, Moskalenko YV. LUNG CANCER IN UKRAINE. *Clinical and Preventive Medicine*. 2025; (2): 6–13. doi: 10.31612/2616-4868.2.2025.01
3. Boubnovski Martell M, Linton-Reid K, Chen M, Aboagye EO. Radiomics for lung cancer diagnosis, management, and future prospects. *Clinical Radiology*. 2025; 86: 106926. doi: 10.1016/j.crad.2025.106926
4. Tahir I, Marquardt JP, Mercaldo ND, Bourgouin PP, Wrobel MM, Mrah S, et al. Utility of Noncancerous Chest CT Features for Predicting Overall Survival and Noncancer Death in Patients With Stage I Lung Cancer Treated With Stereotactic Body Radiotherapy. *American Journal of Roentgenology*. 2022; 219(4): 579–589. doi: 10.2214/AJR.22.27484
5. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*. 2019; 25(6): 954–961. doi: 10.1038/s41591-019-0447-x
6. Retico A. Computer-aided detection for pulmonary nodule identification: improving the radiologist's performance? *Imaging in Medicine*. 2013; 5(3): 249–263. doi: 10.2217/iim.13.24
7. Hwang KH, Lee H, Choi D. Medical Image Retrieval: Past and Present. *Healthcare Informatics Research*. 2012; 18(1): 3. doi: 10.4258/hir.2012.18.1.3
8. Hounsfield GN. Computerized transverse axial scanning (tomography): Part 1. Description of system. *The British Journal of Radiology*. 1973; 46(552): 1016–1022. doi: 10.1259/0007-1285-46-552-1016
9. Glide-Hurst C, Chen D, Zhong H, Chetty IJ. Changes realized from extended bit-depth and metal artifact reduction in CT. *Medical Physics*. 2013; 40(6Part1): 061711. doi: 10.1118/1.4805102

10. Hounsfield GN. Computed Medical Imaging: Journal of Computer Assisted Tomography. 1980; 4(5): 665–674. doi: 10.1097/00004728-198010000-00017
11. Cui X, Fan S, Heuvelmans MA, Han D, Zhao Y, Groen HJM, et al. Optimization of CT windowing for diagnosing invasiveness of adenocarcinoma presenting as sub-solid nodules. European Journal of Radiology. 2020; 128: 108981. doi: 10.1016/j.ejrad.2020.108981
12. National Lung Screening Trial [An official website of the United States government] [Интернет]. USA: National Cancer Institute; 2014. National Lung Screening Trial. Доступный у: <https://www.cancer.gov/types/lung/research/nlst>
13. De Koning HJ, Van Der Aalst CM, De Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. New England Journal of Medicine. 2020; 382(6): 503–513. doi: 10.1056/NEJMoa1911793
14. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology. 2017; 284(1): 228–243. doi: 10.1148/radiol.2017161659
15. Owens C, Hindocha S, Lee R, Millard T, Sharma B. The lung cancers: staging and response, CT,¹⁸ F-FDG PET/CT, MRI, DWI: review and new perspectives. The British Journal of Radiology. 2023; 96(1148): 20220339. doi: 10.1259/bjr.20220339
16. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer. 2009; 45(2): 228–247. doi: 10.1016/j.ejca.2008.10.026
17. Willemink MJ, Noël PB. The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence. European Radiology. 2019; 29(5): 2185–2195. doi: 10.1007/s00330-018-5810-7
18. Đức TM, Hương NT. Improvement of Vehicle detection and classification performance with Region of Interest. ITEJ (Information Technology Engineering Journals). 2023; 8(1): 34–41. doi: 10.24235/itej.v8i1.116

19. Xue C, Yuan J, Lo GG, Chang ATY, Poon DMC, Wong OL, et al. Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review. *Quantitative Imaging in Medicine and Surgery*. 2021; 11(10): 4431–4460. doi: 10.21037/qims-21-86
20. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics*. 2018; 102(4): 1143–1158. doi: 10.1016/j.ijrobp.2018.05.053
21. Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Medical Physics*. 2017; 44(3): 1050–1062. doi: 10.1002/mp.12123
22. Pham DL, Xu C, Prince JL. Current Methods in Medical Image Segmentation. *Annual Review of Biomedical Engineering*. 2000; 2(1): 315–337. doi: 10.1146/annurev.bioeng.2.1.315
23. Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*. 2004; 13(1): 146. doi: 10.1117/1.1631315
24. Niblack W. An introduction to digital image processing. Nachdr. Englewood Cliffs, NJ: Prentice Hall International; 1988. 215 c.
25. Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*. 1979; 9(1): 62–66. doi: 10.1109/TSMC.1979.4310076
26. Hao R, Qiang Y, Yan X. Juxta-Vascular Pulmonary Nodule Segmentation in PET-CT Imaging Based on an LBF Active Contour Model with Information Entropy and Joint Vector. *Computational and Mathematical Methods in Medicine*. 2018; 2018: 1–10. doi: 10.1155/2018/2183847
27. Scholten ETh, Jacobs C, Van Ginneken B, Van Riel S, Vliegenthart R, Oudkerk M, et al. Detection and quantification of the solid component in pulmonary subsolid nodules by semiautomatic segmentation. *European Radiology*. 2015; 25(2): 488–496. doi: 10.1007/s00330-014-3427-z

28. Adams R, Bischof L. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1994; 16(6): 641–647. doi: 10.1109/34.295913
29. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. *International Journal of Computer Vision*. 1988; 1(4): 321–331. doi: 10.1007/BF00133570
30. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*. 2006; 31(3): 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015
31. Pierre F, Amendola M, Bigeard C, Ruel T, Villard PF. Segmentation with Active Contours. *Image Processing On Line*. 2021; 11: 120–141. doi: 10.5201/ipol.2021.298
32. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017; 42: 60–88. doi: 10.1016/j.media.2017.07.005
33. Carmo D, Ribeiro J, Dertkigil S, Appenzeller S, Lotufo R, Rittner L. A Systematic Review of Automated Segmentation Methods and Public Datasets for the Lung and its Lobes and Findings on Computed Tomography Images. *Yearbook of Medical Informatics*. 2022; 31(01): 277–295. doi: 10.1055/s-0042-1742517
34. Bao N, Yuan Y, Luo Q, Li Q, Zhang LB. Edge-enhancement cascaded network for lung lobe segmentation based on CT images. *Frontiers in Physics*. 2023; 11: 1098756. doi: 10.3389/fphy.2023.1098756
35. Gao C, Wu L, Wu W, Huang Y, Wang X, Sun Z, et al. Deep learning in pulmonary nodule detection and segmentation: a systematic review. *European Radiology*. 2024; 35(1): 255–266. doi: 10.1007/s00330-024-10907-0
36. Akinci D’Antonoli T, Farchione A, Lenkiewicz J, Chiappetta M, Cicchetti G, Martino A, et al. CT Radiomics Signature of Tumor and Peritumoral Lung Parenchyma to Predict Nonsmall Cell Lung Cancer Postsurgical Recurrence Risk. *Academic Radiology*. 2020; 27(4): 497–507. doi: 10.1016/j.acra.2019.05.019
37. Tang X, Huang H, Du P, Wang L, Yin H, Xu X. Intratumoral and peritumoral CT-based radiomics strategy reveals distinct subtypes of non-small-cell lung cancer. *Journal*

- of Cancer Research and Clinical Oncology. 2022; 148(9): 2247–2260. doi: 10.1007/s00432-022-04015-z
38. Zhang X, Zhang G, Qiu X, Yin J, Tan W, Yin X, et al. Optimizing the Size of Peritumoral Region for Assessing Non-Small Cell Lung Cancer Heterogeneity Using Radiomics. В: Li Y, Huang Z, Sharma M, Chen L, Zhou R, за ред. Health Information Science [Інтернет]. Singapore: Springer Nature Singapore; 2023 [цит. за 19, Січень 2026]. с. 309–320. (Lecture Notes in Computer Science). Доступний у: https://link.springer.com/10.1007/978-981-99-7108-4_26 doi: 10.1007/978-981-99-7108-4_26
39. Su Y, Tao J, Lan X, Liang C, Huang X, Zhang J, et al. CT-based intratumoral and peritumoral radiomics nomogram to predict spread through air spaces in lung adenocarcinoma with diameter ≤ 3 cm: A multicenter study. European Journal of Radiology Open. 2025; 14: 100630. doi: 10.1016/j.ejro.2024.100630
40. Tang H, Zhang C, Xie X. NoduleNet: Decoupled False Positive Reduction for Pulmonary Nodule Detection and Segmentation [Інтернет]. arXiv; 2019 [цит. за 19, Січень 2026]. Доступний у: <https://arxiv.org/abs/1907.11320> doi: 10.48550/ARXIV.1907.11320
41. Marinakis I, Karampidis K, Papadourakis G. Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review. BioMedInformatics. 2024; 4(3): 2043–2106. doi: 10.3390/biomedinformatics4030111
42. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. В: Navab N, Hornegger J, Wells WM, Frangi AF, за ред. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 [Інтернет]. Cham: Springer International Publishing; 2015 [цит. за 19, Січень 2026]. с. 234–241. (Lecture Notes in Computer Science). Доступний у: http://link.springer.com/10.1007/978-3-319-24574-4_28 doi: 10.1007/978-3-319-24574-4_28
43. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology

problem. *European Radiology Experimental*. 2020; 4(1): 50. doi: 10.1186/s41747-020-00173-2

44. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*. 2017; 14(12): 749–762. doi: 10.1038/nrclinonc.2017.141

45. Давидович І, Бабенко В. Методології аналізу зображень для діагностичних застосувань (огляд). *Біомедична інженерія і технологія*. 2024; (14): 54–63. doi: 10.20535/2617-8974.2024.14.302135

46. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*. 2012; 48(4): 441–446. doi: 10.1016/j.ejca.2011.11.036

47. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016; 278(2): 563–577. doi: 10.1148/radiol.2015151169

48. Yin J, Xi J, Liang J, Zhan C, Jiang W, Lin Z, et al. Solid Components in the Mediastinal Window of Computed Tomography Define a Distinct Subtype of Subsolid Nodules in Clinical Stage I Lung Cancers. *Clinical Lung Cancer*. 2021; 22(4): 324–331. doi: 10.1016/j.clcc.2021.02.015

49. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*. 2012; 30(9): 1234–1248. doi: 10.1016/j.mri.2012.06.010

50. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020; 295(2): 328–338. doi: 10.1148/radiol.2020191145

51. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*. 2014; 5(1): 4006. doi: 10.1038/ncomms5006

52. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. Tian J, за ред. PLOS ONE. 2016; 11(12): e0166550. doi: 10.1371/journal.pone.0166550
53. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*. 2017; 77(21): e104–e107. doi: 10.1158/0008-5472.CAN-17-0339
54. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973; SMC-3(6): 610–621. doi: 10.1109/TSMC.1973.4309314
55. Oberholzer M, Östreicher M, Christen H, Brühlmann M. Methods in quantitative image analysis. *Histochemistry and Cell Biology*. 1996; 105(5): 333–355. doi: 10.1007/BF01463655
56. Liu X, Aldrich C. Deep Learning Approaches to Image Texture Analysis in Material Processing. *Metals*. 2022; 12(2): 355. doi: 10.3390/met12020355
57. Singh A, Regenauer-Lieb K, Walsh SDC, Armstrong RT, Van Griethuysen JJM, Mostaghimi P. On Representative Elementary Volumes of Grayscale Micro-CT Images of Porous Media. *Geophysical Research Letters*. 2020; 47(15): e2020GL088594. doi: 10.1029/2020GL088594
58. Sathish D, Kamath S, Prasad K, Kadavigere R. Texture analysis of breast thermograms using neighbourhood grey tone difference matrix. *International Journal of Bioinformatics Research and Applications*. 2018; 14(1/2): 104. doi: 10.1504/IJBRA.2018.089228
59. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006; 7(1): 91. doi: 10.1186/1471-2105-7-91
60. Yao Y, Wang H. Optimal subsampling for softmax regression. *Statistical Papers*. 2019; 60(2): 585–599. doi: 10.1007/s00362-018-01068-6
61. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1996; 58(1): 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

62. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970; 12(1): 55–67. doi: 10.1080/00401706.1970.10488634
63. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2005; 67(2): 301–320. doi: 10.1111/j.1467-9868.2005.00503.x
64. Das SK, Fang KW, Xu L, Li B, Zhang X, Yang HF. Integrative nomogram of intratumoral, peritumoral, and lymph node radiomic features for prediction of lymph node metastasis in cT1N0M0 lung adenocarcinomas. *Scientific Reports*. 2021; 11(1): 10829. doi: 10.1038/s41598-021-90367-4
65. Wu S, Shen G, Mao J, Gao B. CT Radiomics in Predicting EGFR Mutation in Non-small Cell Lung Cancer: A Single Institutional Study. *Frontiers in Oncology*. 2020; 10: 542957. doi: 10.3389/fonc.2020.542957
66. Warkentin MT, Al-Sawaihey H, Lam S, Liu G, Diergaarde B, Yuan JM, et al. Radiomics analysis to predict pulmonary nodule malignancy using machine learning approaches. *Thorax*. 2024; 79(4): 307–315. doi: 10.1136/thorax-2023-220226
67. Selvam M, Sadanandan A, Chandrasekharan A, Ramesh S, Murali A, Krishnamurthi G. Radiomics for differentiating adenocarcinoma and squamous cell carcinoma in non-small cell lung cancer beyond nodule morphology in chest CT. *Scientific Reports*. 2024; 14(1): 32088. doi: 10.1038/s41598-024-83786-6
68. Breiman L. Random Forests. *Machine Learning*. 2001; 45(1): 5–32. doi: 10.1023/A:1010933404324
69. Jia TY, Xiong JF, Li XY, Yu W, Xu ZY, Cai XW, et al. Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. *European Radiology*. 2019; 29(9): 4742–4750. doi: 10.1007/s00330-019-06024-y
70. Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 2001; 29(5): 1189–1232. doi: 10.1214/aos/1013203451
71. Nastencko I, Linnik M, Honcharuk M, Davydovych I, Lutchenko V, Babenko V, et al. Diagnosis of COVID-19-associated Cardiopulmonary Pathology From CT Data Using

- Artificial Intelligence: a Review of Methods and Future Research Directions. *Innovative Biosystems and Bioengineering*. 2025; 9(4): 16–27. doi: 10.20535/ibb.2025.9.4.325335
72. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. B: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Інтернет]. San Francisco California USA: ACM; 2016 [цит. за 21, Січень 2026]. с. 785–794. Доступний у: <https://dl.acm.org/doi/10.1145/2939672.2939785> doi: 10.1145/2939672.2939785
 73. Юхимюк Р, Шкепаст М, Настенко Є, Лінник М, Давидович І, Бабенко В. Використання глибоких нейронних мереж для порівняльного аналізу норми, пневмонії і COVID-19. *Біомедична інженерія і технологія*. 2023; (12): 56–64. doi: 10.20535/2617-8974.2023.12.292729
 74. Perepeka E, Lazoryshynets V, Babenko V, Davydovych I, Nastenko I. Cardiomyopathy prediction in patients with permanent ventricular pacing using machine learning methods. *System research and information technologies*. 2024; (1): 33–41. doi: 10.20535/SRIT.2308-8893.2024.1.03
 75. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521(7553): 436–444. doi: 10.1038/nature14539
 76. Goodfellow I, Bengio Y, Courville A. Deep Learning [Інтернет]. MIT Press; 2016. Доступний у: <http://www.deeplearningbook.org>
 77. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Інтернет]. arXiv; 2015 [цит. за 21, Січень 2026]. Доступний у: <https://arxiv.org/abs/1502.03167> doi: 10.48550/ARXIV.1502.03167
 78. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks [Інтернет]. arXiv; 2015 [цит. за 21, Січень 2026]. Доступний у: <https://arxiv.org/abs/1506.02025> doi: 10.48550/ARXIV.1506.02025
 79. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*. 2019; 6(1): 60. doi: 10.1186/s40537-019-0197-0
 80. Thanoon MA, Zulkifley MA, Mohd Zainuri MAA, Abdani SR. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics*. 2023; 13(16): 2617. doi: 10.3390/diagnostics13162617

81. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*. 2021; 109(5): 820–838. doi: 10.1109/JPROC.2021.3054390
82. Mienye ID, Swart TG, Obaido G, Jordan M, Ilono P. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*. 2025; 16(3): 195. doi: 10.3390/info16030195
83. Dhar T, Dey N, Borra S, Sherratt RS. Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*. 2023; 4(1): 68–75. doi: 10.1109/TTS.2023.3234203
84. Sivic, Zisserman. Video Google: a text retrieval approach to object matching in videos. B: *Proceedings Ninth IEEE International Conference on Computer Vision* [Інтернет]. Nice, France: IEEE; 2003 [цит. за 19, Січень 2026]. с. 1470–1477 вип.2. Доступний у: <http://ieeexplore.ieee.org/document/1238663/> doi: 10.1109/ICCV.2003.1238663
85. Xu R, Hirano Y, Tachibana R, Kido S. A Bag-of-Features Approach to Classify Six Types of Pulmonary Textures on High-Resolution Computed Tomography. *IEICE Transactions on Information and Systems*. 2013; E96.D(4): 845–855. doi: 10.1587/transinf.E96.D.845
86. Yang J, Feng X, Angelini ED, Laine AF. Texton and sparse representation based texture classification of lung parenchyma in CT images. B: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* [Інтернет]. Orlando, FL, USA: IEEE; 2016 [цит. за 19, Січень 2026]. с. 1276–1279. Доступний у: <http://ieeexplore.ieee.org/document/7590939/> doi: 10.1109/EMBC.2016.7590939
87. Tsai CF. Bag-of-Words Representation in Image Annotation: A Review. *ISRN Artificial Intelligence*. 2012; 2012: 1–19. doi: 10.5402/2012/376804
88. Caplin ME, Baudin E, Ferolla P, Filosso P, Garcia-Yuste M, Lim E, et al. Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical

- pulmonary carcinoids. *Annals of Oncology*. 2015; 26(8): 1604–1620. doi: 10.1093/annonc/mdv041
89. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors. *Journal of Thoracic Oncology*. 2015; 10(9): 1243–1260. doi: 10.1097/JTO.0000000000000630
90. Haibo He, Garcia EA. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. 2009; 21(9): 1263–1284. doi: 10.1109/TKDE.2008.239
91. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*. 2018; 106: 249–259. doi: 10.1016/j.neunet.2018.07.011
92. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020; 295(2): 328–338. doi: 10.1148/radiol.2020191145
93. Larue RTHM, Van Timmeren JE, De Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncologica*. 2017; 56(11): 1544–1553. doi: 10.1080/0284186X.2017.1351624
94. Shafiq-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Scientific Reports*. 2018; 8(1): 10545. doi: 10.1038/s41598-018-28895-9
95. Давидович І, Настенко Є. Клас-балансний відбір інтенсивностей у КТ-зображеннях легень. *Біомедична інженерія і технологія*. 2025; 5(20): 15–25. doi: 10.20535/.2025.20.344360
96. Riitters KH, O'Neill RV, Hunsaker CT, Wickham JD, Yankee DH, Timmins SP, et al. A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*. 1995; 10(1): 23–39. doi: 10.1007/BF00158551

97. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*. 2024; 57(10): 273. doi: 10.1007/s10462-024-10884-2
98. Guzmán C, Bagga M, Kaur A, Westermarck J, Abankwa D. ColonyArea: An ImageJ Plugin to Automatically Quantify Colony Formation in Clonogenic Assays. Rota R, за ред. PLoS ONE. 2014; 9(3): e92444. doi: 10.1371/journal.pone.0092444
99. Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948; 27(3): 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
100. Frank A. Farris. The Gini Index and Measures of Inequality. *The American Mathematical Monthly*. 2010; 117(10): 851. doi: 10.4169/000298910x523344
101. Li X, He HS, Bu R, Wen Q, Chang Y, Hu Y, et al. The adequacy of different landscape metrics for various landscape patterns. *Pattern Recognition*. 2005; 38(12): 2626–2638. doi: 10.1016/j.patcog.2005.05.009
102. Chen Z gui, Kong L jun, Wang M, Liu H kai, Xiao D kai, Wu W ping. The Effect of Urban Spatial Form on Energy Efficiency: A Cross-Sectional Study in China. *Frontiers in Energy Research*. 2021; 9: 792199. doi: 10.3389/fenrg.2021.792199
103. Moran PAP. NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika*. 1950; 37(1–2): 17–23. doi: 10.1093/biomet/37.1-2.17
104. Давидович І, Настенко Є. Визначення специфічних структурних патернів гістологічно підтверджених форм раку легень за даними КТ-зображень. *Наука і техніка сьогодні*. 2025; (11(52)): 2036–2057. doi: 10.52058/2786-6025-2025-11(52)-2036-2057
105. Allwein E, Schapire R, Singer Y. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research*. 2001; 1: 113–141. doi: 10.1162/15324430152733133
106. Rifkin R, Klautau A. In Defense of One-Vs-All Classification. *Machine Learning Research (Google)*. 2004; 5: 101–141.
107. Haibo He, Garcia EA. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. 2009; 21(9): 1263–1284. doi: 10.1109/TKDE.2008.239

108. Chih-Wei Hsu, Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*. 2002; 13(2): 415–425. doi: 10.1109/72.991427
109. Eichelberger RK, Sheng VS. An Empirical Study of Reducing Multiclass Classification Methodologies. В: Perner P, за ред. *Machine Learning and Data Mining in Pattern Recognition* [Інтернет]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013 [цит. за 19, Січень 2026]. с. 505–519. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., eds. *Lecture Notes in Computer Science*). Доступний у: http://link.springer.com/10.1007/978-3-642-39712-7_39 doi: 10.1007/978-3-642-39712-7_39
110. Park SH, Fürnkranz J. Efficient Pairwise Classification. В: Kok JN, Koronacki J, Mantaras RLD, Matwin S, Mladenič D, Skowron A, за ред. *Machine Learning: ECML 2007* [Інтернет]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [цит. за 19, Січень 2026]. с. 658–665. (*Lecture Notes in Computer Science*). Доступний у: http://link.springer.com/10.1007/978-3-540-74958-5_65 doi: 10.1007/978-3-540-74958-5_65
111. Platt J, Cristianini N, Shawe-Taylor J. Large Margin DAGs for Multiclass Classification. В: Вип. 12. MIT Press; 2000. с. 547–553.
112. Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*. 2011; 22(1–2): 31–72. doi: 10.1007/s10618-010-0175-9
113. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. В: *Proceedings of the 23rd international conference on Machine learning - ICML '06* [Інтернет]. Pittsburgh, Pennsylvania: ACM Press; 2006 [цит. за 19, Січень 2026]. с. 233–240. Доступний у: <http://portal.acm.org/citation.cfm?doid=1143844.1143874> doi: 10.1145/1143844.1143874

ДОДАТОК А

СПИСОК ПУБЛІКАЦІЙ ТА ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

Перелік публікацій:

1. Юхимюк Р, Шкепаст М, Настенко Є, Лінник М, Давидович І, Бабенко В. Використання глибоких нейронних мереж для порівняльного аналізу норми, пневмонії і COVID-19. Біомедична інженерія і технологія. 2023; (12): 56-64. doi:10.20535/2617-8974.2023.12.292729

Внесок кожного з авторів: Юхимюк Р. - проектування та реалізація автоенкодеру, написання статті, Шкепаст М. - проектування та реалізація CNN, написання статті, Настенко Є. - остаточне схвалення роботи, Лінник М. - експертна думка, **Давидович І. - концепція роботи та дизайн**, Бабенко В. - критичний огляд роботи.

2. Давидович І, Бабенко В. Методології аналізу зображень для діагностичних застосувань (огляд). Біомедична інженерія і технологія. 2024; (14): 54-63. doi:10.20535/2617-8974.2024.14.302135

Внесок кожного з авторів: **Давидович І. - концепція та дизайн роботи, аналіз матеріалів, написання статті**, Бабенко В. - критичний огляд, написання статті.

3. Perepeka E, Lazoryshynets V, Babenko V, et al. Cardiomyopathy Prediction in Patients with Permanent Ventricular Pacing Using Machine Learning Methods. System Research and Information Technologies. 2024; 1: 33-41. doi:10.20535/SRIT.2308-8893.2024.1.03

Внесок кожного з авторів: Перепека Є. - концепція роботи та дизайн, написання статті, Лазоришинець В. - написання статті, Бабенко В. - критичний огляд, **Давидович І. - критичний огляд**, Настенко Є. - остаточне схвалення статті.

4. Nastenko I, Linnik M, Honcharuk M, Davydovych I, Lutchenko V, Babenko V, et al. Diagnosis of COVID-19-associated Cardiopulmonary Pathology From CT Data Using Artificial Intelligence: a Review of Methods and Future Research Directions.

Innovative Biosystems and Bioengineering. 2025; 9(4): 16-27.
doi:10.20535/ibb.2025.9.4.325335

Внесок кожного з авторів: Настенко Є. - остаточне схвалення статті, Линник М. - експертна думка, Гончарук М. - проектування та реалізація алгоритму виділення ознак, написання статті, **Давидович І. - проектування та реалізація алгоритму виділення ознак, написання статті**, Лутченко В. - проектування та реалізація класифікаторів, Бабенко В. - написання статті, Долінчук Л. - критичний огляд.

5. Давидович І, Настенко Є. Клас-балансний відбір інтенсивностей у КТ-зображеннях легень. Біомедична інженерія і технологія. 2025; (20): 15-25.
doi:10.20535/.2025.20.344360

Внесок кожного з авторів: **Давидович І. - концепція роботи та дизайн, проектування та реалізація алгоритму фільтрації, написання статті**, Настенко Є. - критичний огляд, остаточне схвалення статті.

6. Давидович І, Настенко Є. Визначення специфічних структурних патернів гістологічно підтверджених форм раку легень за даними КТ-зображень. Наука і техніка сьогодні. 2025; (11(52)): 2036-2057. doi:10.52058/2786-6025-2025-11(52)-2036-2057

Давидович І. - концепція роботи та дизайн, проектування та реалізація способу вилучення ознак, проектування та реалізація моделей класифікації, написання статті, Настенко Є. - критичний огляд, остаточне схвалення статті.

Апробація матеріалів дисертації:

7. Давидович І, Настенко Є. Клас-балансний підхід до фільтрації інтенсивностей у КТ-зображеннях легень. In: Scientific Innovation: Theoretical Insights and Practical Impacts: Collection of Scientific Papers with the Proceedings of the 4th International Scientific and Practical Conference; 2025 Dec 8-10; Naples, Italy (online). European Open Science Space; 2025. p. 141-144. doi:10.70286/EOSS-08.12.2025. ISBN: 979-8-89704-952-3

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

8. Давидович І, Настенко Є. Локальні текстурні патерни КТ-зображень легень для розмежування гістологічних форм раку. In: Research in Science, Technology and Economics: Collection of Scientific Papers with Proceedings of the 5th International Scientific and Practical Conference; 2025 Dec 10-12; Luxembourg, Luxembourg (online). International Scientific Unity; 2025. p. 257-260. doi:10.70286/ISU-10.12.2025. ISBN: 979-8-89704-985-1

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

9. Давидович І, Настенко Є. Вилучення ознак із локальних структурних патернів частотних карт КТ-зображень легень у багатокласовій класифікації. In: Modern Perspectives on Global Scientific Solutions: Collection of Scientific Papers with the Proceedings of the 7th International Scientific and Practical Conference; 2026 Jan 26-28; Bergen, Norway (online). European Open Science Space; 2026. p. 60-64. doi:10.70286/EOSS-26.01.2026. ISBN: 979-8-89704-959-2

Внесок кожного з авторів: **Давидович І. - написання тез, доповідь,**
Настенко Є. - остаточне схвалення матеріалів.

ДОДАТОК Б

АКТИ ВПРОВАДЖЕННЯ

№ 6/7
від 09.12.23р.

«ЗАТВЕРДЖУЮ»

Заступник генерального директора
з наукової роботи ДУ «Національний
науковий центр фтизіатрії, пульмонології
та алергології ім. Ф. Г. Яновського»
НАМН України
д-р мед. н., професор Калабуха І.А.

2025 р.

АКТ ВПРОВАДЖЕННЯ

- Назва запропонованого впровадження:** «Технологія визначення специфічних структурних патернів підтверджених форм раку легень за даними КТ-зображень» на основі алгоритмів машинного навчання.
- Установа-розробник:** Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» МОН України.
- Джерела інформації:**
 - Юхимюк Р., Шкепаст М., Настенко С., Линник М., Давидович І., Бабенко В. Використання глибоких нейронних мереж для порівняльного аналізу норми, пневмонії і COVID-19 / Biomedical Engineering and Technology. Issue 12, 2023, p. 56–64. // Юхимюк Р., Шкепаст М., Настенко С., Линник М., Давидович І., Бабенко В. / URL: <http://biomedtech.kpi.ua/article/view/292729> (doi.org/10.20535/2617-8974.2023.12.292729).
 - Давидович І. В., Бабенко В. О. Методології аналізу зображень для діагностичних застосувань (огляд) // Біомедична інженерія і технологія. – 2024. – № 14 (2). – С. 54-63. – DOI: [10.20535/2617-8974.2024.14.302135](https://doi.org/10.20535/2617-8974.2024.14.302135).
 - Настенко С., Линник М., Гончарук М., Давидович І., Лутченко В., Бабенко В., Долінчук Л. / Результати застосування методів машинного навчання у дослідженні COVID-19 асоційованої кардіопульмональної патології за даними комп'ютерної томографії // Innovative Biosystems and Bioengineering. – 2025. – Т. 9, № 4. – С. 16-27. – DOI: [10.20535/ibb.2025.9.4.325335](https://doi.org/10.20535/ibb.2025.9.4.325335). (URL: <https://ibb.kpi.ua/article/view/325335>)

Рисунок Б.1 – Акт впровадження в ДУ «Національний науковий центр фтизіатрії, пульмонології та алергології ім. Ф.Г. Яновського» (1 сторінка)

3.4 Давидович І.В., Настенко Є.А. Визначення специфічних структурних патернів гістологічно підтверджених форм раку легень за даними КТ-зображень // Давидович І.В., Настенко Є.А. / Наука і техніка сьогодні. — 2025. — № 11 (52). — С. 2036-2057. — DOI: [10.52058/2786-6025-2025-11\(52\)](https://doi.org/10.52058/2786-6025-2025-11(52)).

4. Де і коли впроваджено: Відділ епідеміологічних та організаційних проблем фтизіопульмонології
5. Ефективність впровадження у відповідності з критеріями, викладеними у джерелі інформації: впровадження технології дозволило підвищити точність діагностики раку легень на основі алгоритмів машинного навчання.
6. Зауваження, пропозиції: запропонована методика відповідає потребам клінічної практики.

Відповідальний за впровадження: _____

/д. мед. н. / Липшик М.І. /

«09» грудня 2025 р.

Рисунок Б.2 – Акт впровадження в ДУ «Національний науковий центр фтизіатрії, пульмонології та алергології ім. Ф.Г. Яновського» (2 сторінка)