Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра Комп'ютерного моделювання та інтелектуальних технологій
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____другий (магістерський)_____
«Дослідження моделей згортальних нейронних мереж глибинного навчання для задач сегментації зображень комп'ютерної томографії»
(тема)

Виконав:
здобувач __2026___ року навчання,
групи СПРм-24-1, Гаєвський Антон Володимирович____
(власне ім'я, прізвище)

Спеціальність ___122 Комп'ютерні науки_
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системне проектування_
(повна назва освітньої програми)

Керівник проф. каф. КМІТ Нечипоренко А. С.
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри _____ _____ _____
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет_____Комп'ютерних наук_____

Кафедра__Комп'ютерного моделювання та інтелектуальних технологій____

Рівень вищої освіти_____другий (магістерський)_____

Спеціальність__122 _Комп'ютерні науки_____
(код і повна назва)

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

Освітня програма_____Системне проектування_____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

«____»_____ 20 ___ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві_____Гаєвському Антону Володимировичу_____
(прізвище, ім'я, по батькові)

1. Тема роботи ____«Дослідження моделей згортальних нейронних мереж глибинного навчання для задач сегментації зображень комп'ютерної томографії»_____
затверджена наказом університету від 06 квітня 2026 р. № _268_
2. Термін подання здобувачем роботи до екзаменаційної комісії _12 _травня__ 2026р.
3. Вихідні дані до роботи _1. Клінічний датасет КТ-досліджень голови з еталонною сегментацією клиноподібної пазухи (534 кейси). 2. Технічна документація фреймворка nnU-Net та бібліотеки MONAI. 3. Передтреновані ваги моделі SuPreM. 4.Результати ручного експертного перегляду масок контрольної вибірки. 5.Відкриті публікації з медичної сегментації, виявлення шуму розмітки та ансамблевих методів глибокого навчання.
4. Перелік питань, що потрібно опрацювати в роботі Аналіз предметної області та постановка задачі сегментації клиноподібної пазухи на КТ-зображеннях. Дослідження особливостей КТ-даних і чинників, що ускладнюють автоматичну сегментацію. Огляд класичних, 2D/2.5D та 3D методів сегментації медичних зображень. Обґрунтування вибору архітектури nnU-Net 3D full resolution. Аналіз впливу шуму експертної розмітки на навчання та оцінювання моделей.Підготовка даних і проведення незалежного експертного перегляду контрольної вибірки. Удосконалення процесу сегментації КТ-зображень на основі виявлення аномалій розмітки. Побудова показника аномальності та перевірка його здатності ранжувати проблемні випадки. Оцінювання впливу очищення даних на якість сегментації. Аналіз можливостей застосування розробленого підходу для контролю якості та активного навчання._____
5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Схема замкненого циклу валідації детектора аномалій розмітки (1 аркуш формату А4); комп'ютерна ілюстрація

КТ-знімка черепа з виділеною клиноподібною пазухою (1 аркуш формату А4); графік порівняння моделей за показником Dice (1 аркуш формату А4); графік порівняння моделей за показником HD95 (1 аркуш формату А4); комп'ютерні ілюстрації типових дефектів розмітки клиноподібної пазухи (1 аркуш формату А4); схема практичних сценаріїв використання детектора аномалій розмітки (1 аркуш формату А4); схема використання детектора в циклі активного навчання (1 аркуш формату А4).

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	20.01.2026	Виконано
2	Аналіз нормативних та наукових джерел	01.02.2026	Виконано
3	Формування датасету та ручний експертний перегляд	15.02.2026	Виконано
4	Проєктування експериментальної методики	20.02.2026	Виконано
5	Реалізація програмних модулів і алгоритмів	01.03.2026	Виконано
6	Проведення експериментів і обробка результатів	12.03.2026	Виконано
7	Написання та оформлення пояснювальної записки	01.04.2026	Виконано
8	Перевірка на академічний плагіат	11.05.2026	
9	Підготовка презентації та доповіді	11.05.2026	
10	Захист перед ЕК	13.05.2026	

Дата видачі завдання 19 лютого 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 111 с., 14 рис., 30 табл., 42 джерел.

АВТОМАТИЧНА СЕГМЕНТАЦІЯ, КЛИНОПОДІБНА ПАЗУХА, КОМП'ЮТЕРНА ТОМОГРАФІЯ, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, ГЛИБИННЕ НАВЧАННЯ, NN U-NET, ШУМ РОЗМІТКИ, КОНТРОЛЬ ЯКОСТІ ДАНИХ.

Об'єкт дослідження – процес автоматичної сегментації анатомічних структур на зображеннях комп'ютерної томографії за допомогою згорткових нейронних мереж глибинного навчання.

Предмет дослідження – моделі тривимірної сегментації КТ-зображень клиноподібної пазухи та методи попередньої обробки даних і оцінювання якості експертної розмітки в умовах її неоднорідності.

Мета роботи – дослідження моделей згортальних нейронних мереж глибинного навчання для сегментації зображень комп'ютерної томографії та обґрунтування процедури попередньої обробки даних, яка враховує неоднорідну якість наявної експертної розмітки.

Методи дослідження – теоретичний аналіз методів сегментації медичних зображень, побудова тривимірних CNN-моделей, використання фреймворку nnU-Net, експертний перегляд контрольної вибірки, ансамблеве оцінювання прогнозів, розрахунок метрик сегментації та статистичний аналіз впливу якості розмітки на результати навчання моделей.

У результаті роботи обґрунтовано доцільність використання тривимірних моделей для сегментації клиноподібної пазухи на КТ-зображеннях і запропоновано підхід до вдосконалення процесу сегментації через виявлення потенційно дефектних випадків експертної розмітки.

ABSTRACT

Master’s thesis contains: 111 pages, 14 figures, 30 tables, 42 references.

AUTOMATIC SEGMENTATION, SPHENOID SINUS, COMPUTED TOMOGRAPHY, CONVOLUTIONAL NEURAL NETWORKS, DEEP LEARNING, NN U-NET, LABEL NOISE, DATA QUALITY CONTROL.

The object of the study is the process of automatic segmentation of anatomical structures in computed tomography images using deep convolutional neural networks.

The subject of the study is three-dimensional CT image segmentation models for the sphenoid sinus, as well as methods of data preprocessing and expert annotation quality assessment under heterogeneous annotation quality.

The aim of the study is to investigate deep convolutional neural network models for computed tomography image segmentation and to substantiate a data preprocessing procedure that takes into account the heterogeneous quality of available expert annotations.

The research methods include theoretical analysis of medical image segmentation methods, development of three-dimensional CNN models, use of the nnU-Net framework, expert review of the control subset, ensemble-based evaluation of predictions, calculation of segmentation metrics, and statistical analysis of the influence of annotation quality on model training results.

As a result of the study, the feasibility of using three-dimensional models for sphenoid sinus segmentation in CT images was substantiated, and an approach to improving the segmentation process through the detection of potentially defective expert annotations was proposed.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

КТ – комп’ютерна томографія;

МРТ – магнітно-резонансна томографія;

НДКТ – низькодозова комп’ютерна томографія;

НМ – нейронна мережа;

ШІ – штучний інтелект;

мм – міліметр;

п.п. – процентний пункт;

2D – two-dimensional – двовимірний;

2.5D – two-and-a-half-dimensional – псевдотривимірний;

3D – three-dimensional – тривимірний;

ACDC – Automated Cardiac Diagnosis Challenge – бенчмарк медичної сегментації серця;

AMOS – Abdominal Multi-Organ Segmentation – бенчмарк багатоорганної сегментації органів черевної порожнини;

AP – Average Precision – середня точність;

ASSD – Average Symmetric Surface Distance – середня симетрична поверхнева відстань;

AUC – Area Under the Curve – площа під кривою;

CBCT – Cone Beam Computed Tomography – конусно-променева комп’ютерна томографія;

CLI – Command-Line Interface – інтерфейс командного рядка;

CNN – Convolutional Neural Network – згорткова нейронна мережа;

CSV – Comma-Separated Values – формат даних із розділенням комами;

DICOM – Digital Imaging and Communications in Medicine – стандарт цифрових медичних зображень і комунікацій;

Dice – Dice Similarity Coefficient – коефіцієнт подібності Dice;

DiceCE – Dice Cross-Entropy loss – функція втрат, що поєднує Dice loss і перехресну ентропію;

F1 – F1-score – гармонійне середнє точності та повноти;

FCN – Fully Convolutional Network – повністю згорткова мережа;

FN – False Negative – хибнонегативний результат;

FP – False Positive – хибнопозитивний результат;

GLCM – Gray-Level Co-occurrence Matrix – матриця співзустрічальності рівнів сірого;

GPU – Graphics Processing Unit – графічний процесор;

GT – Ground Truth – еталонна розмітка;

H1 – перша дослідна гіпотеза;

HD95 – Hausdorff Distance 95th percentile – 95-й перцентиль відстані Гаусдорфа;

HU – Hounsfield Unit – одиниця шкали Гаунсфілда;

IoU – Intersection over Union – перетин над об’єднанням;

JSON – JavaScript Object Notation – текстовий формат обміну даними;

KiTS – Kidney Tumor Segmentation Challenge – бенчмарк сегментації нирок і ниркових пухлин;

LBP – Local Binary Patterns – локальні бінарні шаблони;

M1 – базова модель PlainConvUNet 3D у складі nnU-Net;

M2 – модель SegResNet з попередньо навченими вагами SuPreM;

M3 – модель SegResNet з випадковою ініціалізацією;

MedNeXt – масштабована згорткова архітектура для медичної сегментації;

MedSAM – медична адаптація Segment Anything Model;

MONAI – Medical Open Network for AI – фреймворк глибинного навчання для медичних зображень;

NIfTI – Neuroimaging Informatics Technology Initiative – формат нейровізуалізаційних даних;

nnFormer – volumetric medical image segmentation via a 3D transformer – трансформерна модель для тривимірної медичної сегментації;

nnU-Net – self-configuring U-Net – самоналаштовуваний фреймворк для медичної сегментації;

NSD – Normalized Surface Dice – нормалізований поверхневий коефіцієнт Dice;

OOF – Out-of-Fold – прогнозування поза фолдом;

PlainConvUNet – варіант U-Net із простими згортковими блоками;

PyTorch – фреймворк глибинного навчання;

ReLU – Rectified Linear Unit – випрямлена лінійна активаційна функція;

ROC – Receiver Operating Characteristic – робоча характеристика приймача;

ROC-AUC – Receiver Operating Characteristic Area Under the Curve – площа під ROC-кривою;

SAM – Segment Anything Model – модель сегментації за підказкою;

SAM-Med3D – адаптація Segment Anything Model до тривимірних медичних даних;

SegResNet – сегментаційна залишкова нейронна мережа;

STAPLE – Simultaneous Truth and Performance Level Estimation – метод одночасного оцінювання істинної розмітки та якості експертів;

STU-Net – scalable and transferable U-Net – масштабована та переносима U-Net-подібна модель;

SuPreM – Supervised Pre-training Model – модель попереднього навчання для тривимірної медичної сегментації;

Swin UNETR – Swin Transformer UNETR – трансформерна архітектура для медичної сегментації;

TN – True Negative – істиннонегативний результат;

TP – True Positive – істиннопозитивний результат;

TotalSegmentator – інструмент автоматичної багатоорганної сегментації КТ-зображень;

U-Net – U-подібна згорткова архітектура для сегментації зображень;

UNETR – U-Net Transformer – трансформерна U-Net-подібна архітектура для медичної сегментації;

V-Net – тривимірна згорткова архітектура V-подібного типу для медичної сегментації;

ViT – Vision Transformer – візуальний трансформер.

ВСТУП

У сучасних умовах активного впровадження методів комп'ютерного зору у клінічну практику задача автоматизованої сегментації анатомічних структур на зображеннях комп'ютерної томографії набуває визначального значення для якісної об'ємної оцінки анатомії, передопераційного планування та стандартизації кількісних показників. Згортальні нейронні мережі глибинного навчання розглядаються як основний інструмент сучасної медичної сегментації, оскільки забезпечують вищу точність та узгодженість результатів порівняно з класичними алгоритмічними підходами. Окремим клінічно значущим, але порівняно мало представленим у літературі об'єктом сегментації є клиноподібна пазуха, точне виділення якої є важливим для передопераційного планування транссфеноїдального доступу та аналізу анатомічних варіантів у ділянці основи черепа. Попри значний прогрес у розвитку моделей та збільшення обсягів навчальних корпусів, залишається невирішеною прикладна проблема забезпечення стабільної якості еталонної розмітки, від якої безпосередньо залежать як процес навчання, так і подальша оцінка моделей сегментації.

Актуальність роботи зумовлена поєднанням кількох тенденцій. У клінічних датасетах для медичної сегментації систематично спостерігається неоднорідна якість експертних масок, оскільки розмітка зазвичай виконується в умовах обмеженого часу, різними фахівцями та за різними протоколами. У досліджуваному наборі клиноподібної пазухи ця проблема є особливо вираженою: близько 15 % випадків мають явні дефекти розмітки, а ще приблизно 23 % характеризуються сумнівною якістю. Паралельно зростає увага наукової спільноти до критики стандартної процедури оцінювання сегментаційних моделей за метрикою Dice, оскільки виміряні на тестовому наборі показники у такому випадку відображають не лише властивості моделі, а й похибку самих еталонних масок. За відсутності формалізованого механізму контролю якості розмітки на етапі попередньої

обробки даних виникають такі проблеми, як завищення або заниження виміряних метрик, систематичне відтворення дефектів розмітки моделлю під час навчання та помилкові висновки під час порівняння архітектур.

Метою кваліфікаційної роботи є дослідження моделей згортальних нейронних мереж глибинного навчання для сегментації зображень комп'ютерної томографії та обґрунтування процедури попередньої обробки даних, яка враховує неоднорідну якість наявної експертної розмітки. Для досягнення поставленої мети пропонується поєднати ансамбль тривимірних CNN-моделей сегментації з диверсифікованими архітектурними передумовами, незалежну експертну оцінку контрольної вибірки та процедуру перетворення прогнозів ансамблю у композитну оцінку підозрілості розмітки кожного випадку. У роботі експериментально перевіряються гіпотези щодо узгодженості сигналу запропонованого детектора з незалежною експертною оцінкою, його переваги над ручною фільтрацією як інструмента контролю якості, доцільності повністю автоматичного очищення навчальної вибірки та потенціалу того самого сигналу як критерію вибору кейсів у режимі активного навчання. Дослідження виконано на клінічному датасеті КТ-зображень голови зі 534 досліджень, що включає тренувальний пул і дві незалежні контрольні підвибірки з різних клінічних установ.

Отримані результати можуть бути використані у медичних інформаційних системах, що передбачають побудову моделей сегментації анатомічних структур на КТ-зображеннях, у наукових проєктах з розробки нових сегментаційних архітектур, а також у клінічних робочих процесах, де експертний перегляд автоматичної розмітки потребує оптимізації часу радіолога. Запропонована процедура автоматичного виявлення підозрілих випадків розмітки здатна слугувати інструментом контролю якості наявних анотацій, пріоритизації експертного перегляду та обґрунтованої оцінки потенціалу подальшого вдосконалення моделей на конкретному датасеті.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Сегментація медичних зображень

Сегментація медичних зображень є одним із базових етапів комп’ютерного аналізу візуальних діагностичних даних, оскільки саме вона забезпечує просторове виділення анатомічних структур, патологічних утворень або функціонально важливих областей на зображенні. На відміну від задачі класифікації, де результатом є належність усього зображення до певного класу, сегментація передбачає піксельно або воксельно точне визначення меж об’єкта. Для тривимірних медичних даних, зокрема комп’ютерної томографії, ця задача може бути формалізована як відображення вхідного об’єму з інтенсивностями вокселів у відповідну міткову маску:

$$f: \mathbb{R}^{H\times W\times D} \to \{0,\dots,K\}^{H\times W\times D}. \quad (1.1)$$

де H, W і D визначають просторові розміри томографічного об’єму;

K – кількість класів сегментації.

У найпростішому випадку виконується бінарна сегментація, коли кожен воксель належить або до фону, або до цільової структури. У складніших задачах застосовується багатокласова семантична сегментація, де одночасно виділяється декілька органів або анатомічних областей. Для медичних зображень за допомогою комп'ютерної томографії (КТ) найчастіше використовується саме семантичний підхід із наявністю фонового класу, тоді як інстансна та паноптична сегментація більш характерні для задач, де необхідно розділяти окремі екземпляри однотипних об’єктів.

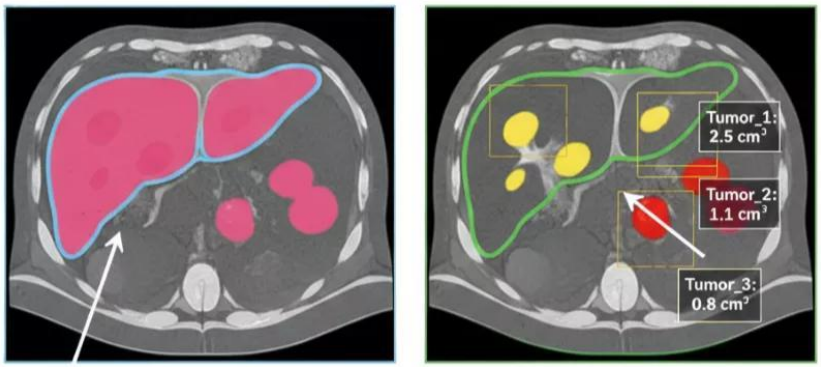

Рисунок 1.1 – Приклад семантичного та інстансного сегметування

Сучасні огляди підкреслюють, що сегментація є критичною складовою медичного аналізу зображень, оскільки вона безпосередньо пов’язана з кількісним вимірюванням, плануванням лікування та підтримкою клінічного рішення [1].

Медична сегментація застосовується до різних модальностей візуалізації, серед яких КТ, магнітно-резонансна томографія, ультразвукове дослідження, рентгенографія та гістопатологічні зображення. Кожна з цих модальностей має власні фізичні особливості, типи шуму, просторову роздільну здатність і характер контрасту між тканинами.

У межах цієї роботи основна увага приділяється КТ-зображенням голови, оскільки вони дають змогу чітко візуалізувати кісткові структури та повітроносні порожнини черепа. Важливою особливістю є використання шкали Гаунсфілда, яка відображає рентгенівську щільність тканин і дозволяє розрізняти повітря, кістку, м’які тканини та патологічні зміни. Водночас КТ-дані мають низку технічних обмежень, серед яких анізотропія розміру вокселя, наявність артефактів від руху або металевих конструкцій, а також підвищений шум у НДКТ. Усі ці фактори ускладнюють автоматичне виділення дрібних або анатомічно варіабельних структур, тому задача сегментації потребує не лише аналізу окремих зрізів, а й урахування повного тривимірного контексту.

Типовий пайплайн сегментації медичних КТ-зображень охоплює декілька взаємопов’язаних етапів.

1. Спочатку виконується збір даних у форматі DICOM, після чого вони можуть бути перетворені у формат NIfTI для подальшої обробки в дослідницьких та нейромережевих середовищах.

2. Далі застосовується попередня обробка, яка може включати обрізання діапазону HU, ресемплінг до уніфікованого просторового кроку, нормалізацію інтенсивностей і приведення об’ємів до узгодженого формату.

3. Наступним етапом є створення еталонної розмітки експертом, яка береться за істинне значення для навчання та перевірки моделі. Після навчання виконується валідація, інференс на нових КТ-дослідженнях і постобробка отриманої маски. Наприклад, видалення хибних малих компонентів зв’язності або згладжування меж.

Таким чином, сегментація не є ізольованою операцією, а виступає частиною повного аналітичного процесу, де якість кожного етапу впливає на точність кінцевого результату.

Клінічне значення сегментації визначається тим, що вона дає змогу перейти від візуальної оцінки зображення до кількісного аналізу. У практичних задачах сегментаційні маски використовуються для планування променевої терапії, передопераційного моделювання, оцінювання об’єму органів і патологічних ділянок, побудови біомаркерів, навігації під час хірургічних втручань і моніторингу динаміки захворювання.

У випадку сегментації клиноподібної пазухи особливе значення має її зв’язок із транссфеноїдальним доступом до гіпофіза та структурами основи черепа. Клиноподібна пазуха є повітроносною порожниною, яка характеризується значною індивідуальною варіабельністю пневматизації, зокрема конхального, преселлярного, селлярного та постселлярного типів. Її тонкі кісткові стінки та близькість до внутрішньої сонної артерії, зорового нерва і кавернозного синуса підвищують вимоги до точності просторового виділення цієї області [2].

На відміну від природних зображень, медичні зображення мають низку специфічних особливостей, які значно ускладнюють побудову автоматичних методів сегментації.

По-перше, у таких задачах часто спостерігається виражений дисбаланс класів, коли область інтересу займає лише незначну частину всього об’єму. Для клиноподібної пазухи ця проблема є особливо помітною, оскільки цільова структура є відносно малою порівняно з повним КТ-об’ємом голови.

По-друге, медичні датасети зазвичай містять обмежену кількість пацієнтів, оскільки їх збір, анотація та перевірка потребують участі фахівців.

По-третє, анатомічна варіабельність між пацієнтами призводить до того, що модель повинна не просто запам’ятовувати типову форму об’єкта, а узагальнювати просторові закономірності. Крім того, ціна помилки у медичній сегментації є значно вищою, ніж у багатьох задачах загального комп’ютерного зору, оскільки хибно пропущені ділянки можуть мати більше клінічне значення, ніж окремі хибнопозитивні включення. Саме тому в медичній сегментації важливими є не лише середні значення точності, а й стабільність результатів на складних клінічних випадках. Глибокі нейронні мережі (НМ) стали провідним напрямом у цій області саме через здатність автоматично виділяти ознаки з великих обсягів візуальних даних, хоча їх ефективність суттєво залежить від якості розмітки, обсягу навчальної вибірки та коректності валідації [3].

Для оцінювання якості сегментації використовують як об’ємні, так і поверхневі метрики. Найпоширенішими є Dice Similarity Coefficient та IoU/Jaccard, які оцінюють ступінь перекриття між прогнозованою та еталонною маскою. Precision і Recall дають змогу окремо проаналізувати кількість хибнопозитивних і хибнонегативних вокселів, що є важливим для клінічної інтерпретації помилок. Для оцінювання точності меж застосовуються Hausdorff distance, HD95, Average Surface Distance і

Normalized Surface Dice. Поверхневі метрики є особливо важливими у задачах сегментації малих структур, оскільки навіть незначне зміщення межі може суттєво впливати на відносні показники якості. У сучасних роботах підкреслюється, що вибір метрик повинен відповідати клінічній задачі, оскільки одна лише метрика Dice не завжди повно відображає якість сегментації, особливо для малих або складних за формою об’єктів [4].

Ручна анотація медичних зображень залишається золотим стандартом формування еталонних масок, однак вона є трудомісткою, залежить від досвіду експерта та може займати значний час для однієї КТ-серії. Напівавтоматичні інструменти, такі як інтерактивне виділення контурів або редактори сегментацій у спеціалізованих медичних програмах, частково зменшують навантаження на фахівця, але все одно потребують контролю й ручного виправлення.

У зв’язку з цим сучасні підходи дедалі частіше базуються на глибинному навчанні, зокрема на згортальних НМ, здатних враховувати просторовий контекст і складні закономірності форми. Для сегментації клиноподібної пазухи така автоматизація є особливо доцільною, оскільки ця структура є малою, анатомічно варіабельною та клінічно важливою, а її точне виділення потребує аналізу всього 3D-об’єму КТ. Саме тому подальший розгляд доцільно спрямувати на моделі глибинного навчання для медичної сегментації та архітектури, які здатні працювати з тривимірними томографічними даними.

1.2 Огляд сучасного стану розвитку методів сегментації зображень

Розвиток методів сегментації медичних зображень відбувався поступово і значною мірою відображає загальну еволюцію комп’ютерного зору. На ранніх етапах, приблизно від 1980-х до початку 2000-х років, переважали класичні алгоритмічні підходи, що базувалися на пороговій обробці, аналізі інтенсивностей, нарощуванні областей, активних контурах,

та морфологічних операціях. Такі методи були відносно простими для інтерпретації та могли давати прийнятні результати для структур із чітким контрастом, однак вони суттєво залежали від якості зображення, параметрів алгоритму та попередніх припущень про форму або інтенсивність цільового об’єкта.

Наступним етапом стало поширення методів класичного машинного навчання, які використовували ручне конструювання ознак. У період приблизно з 2005 до 2015 року активно застосовувалися випадкові ліси, методи на основі атласів, графові моделі; класифікатори на базі локальних текстурних, інтенсивнісних і просторових характеристик. Такі підходи були гнучкішими за суто алгоритмічні методи, однак потребували ретельного добору ознак та часто мали обмежену здатність до узагальнення на нові клінічні центри або інші протоколи сканування [5].

Різкий перехід до глибинного навчання пов’язують із успіхом згорткових НМ у задачах природних зображень після 2012 року, однак для сегментації важливою віхою стала поява повністю згорткових мереж. Fully Convolutional Network (FCN) сформувала наскрізний підхід до семантичної сегментації, де модель безпосередньо перетворює вхідне зображення у карту класів. Для медичних зображень ключовою архітектурою стала U-Net, запропонована у 2015 році. Її структура дала змогу поєднати глибокі семантичні ознаки з локальною просторовою деталізацією, що є критично важливим для точного відновлення меж органів і малих анатомічних структур. Саме тому U-Net стала базовою архітектурною ідеєю для великої кількості наступних моделей медичної сегментації.

Подальший розвиток був спрямований на адаптацію U-Net до тривимірних медичних даних і підвищення точності сегментації об’ємних структур. V-Net запропонувала використання повноцінної 3D-архітектури та функції втрат на основі Dice, що було особливо важливим за умов дисбалансу класів. 3D U-Net розширила принципи на об’ємні зображення та дозволила враховувати просторові зв’язки між сусідніми зрізами. Надалі

з’явилися модифікації, спрямовані на покращення передавання ознак між рівнями мережі, зокрема U-Net++.

Паралельно розвивалися архітектури сімейства DeepLab, які використовували розширені згортки та розражене просторове пірамідне об'єднання для врахування контексту на різних масштабах. Проте для медичної 3D-сегментації найбільш стійкою базовою конструкцією залишалася саме U-подібна архітектура, оскільки вона природно поєднує локалізацію об’єкта та багаторівневу семантичну інформацію.

Окремим етапом розвитку стало формування самоналаштовуваних фреймворків. Найбільш відомим серед них є nnU-Net, який не пропонує принципово новий тип шару, але систематизує повний процес побудови сегментаційної моделі. Його основна ідея полягає в тому, що якість медичної сегментації визначається не лише архітектурою мережі, а й сукупністю рішень щодо попередньої обробки, вибору просторового кроку, розміру патчів, конфігурації 2D або 3D мережі, навчання, інференсу та постобробки. У роботі Isensee та співавторів nnU-Net описано як метод, який автоматично конфігурує передобробку, налаштування архітектури, тренування та постобробку для нової задачі, що зробило його сильним універсальним базової моделі для біомедичної сегментації [6].

Значення публічних бенчмарків у цій області полягає в тому, що вони дають змогу оцінювати методи не на окремих малих вибірках, а на стандартизованих наборах даних із чіткими протоколами порівняння. Medical Segmentation Decathlon був створений саме як багатозадачний бенчмарк для перевірки того, чи здатний метод, ефективний на різних задачах і модальностях, узагальнюватися на нові задачі [7]. KiTS21 зосереджений на автоматичній сегментації нирок, пухлин і кіст у КТ, причому його тестовий набір включав дані з іншої установи, що підвищувало вимоги до узагальнення моделей [8]. AMOS22 запропонував 500 КТ і 100 МРТ досліджень з воксельними анотаціями 15 органів, що зробило його важливим набором для оцінювання багатоорганної

сегментації в різних клінічних умовах [9]. TotalSegmentator, у свою чергу, показав можливість автоматичної сегментації 104 анатомічних структур на КТ-зображеннях, що демонструє перехід від вузьких моделей до більш універсальних систем аналізу тіла [10].

У період з 2021 до 2024 року значну увагу отримали архітектури на основі трансформерів, зокрема UNETR, Swin UNETR і nnFormer, а також гібридні підходи, які поєднують згорткові блоки з механізмами глобального контексту. Пізніше з’явилися масштабовані CNN-архітектури, такі як MedNeXt і STU-Net, а також моделі нового покоління, орієнтовані на підхід базової моделі і сегментацією з можливістю підказки. Проте сучасні порівняльні дослідження показують, що новизна архітектури сама по собі не гарантує переваги над якісно налаштованою U-Net. У роботі Isensee та співавторів 2024 року прямо зазначено проблему упередженності до інновацій, тобто тенденцію перебільшувати переваги нових архітектур через слабкі базові налаштування, недостатні датасети, різні обчислювальні бюджети або неповну валідацію. Автори показують, що CNN-based U-Net моделі, реалізовані у фреймворку nnU-Net, стабільно демонструють сильні результати на шести 3D-наборах даних, а KiTS, AMOS і ACDC визначаються як найбільш придатні для бенчмаркінгу 3D-сегментації через низький статистичний шум і здатність краще розрізняти методи [11].

Отже, сучасний стан розвитку методів сегментації медичних зображень можна охарактеризувати не як просту заміну класичних моделей новими архітектурами, а як поступове ускладнення всієї експериментальної методології. Якщо ранні підходи були зосереджені переважно на алгоритмі виділення меж, то сучасні системи враховують повний цикл роботи з даними, від нормалізації і ресемплінгу до вибору конфігурації мережі, схеми навчання, інференсу та постобробки.

1.3 Потенціал застосування методів штучного інтелекту для обробки зображень комп’ютерної томографії

Комп’ютерна томографія є однією з ключових модальностей сучасної медичної візуалізації, оскільки забезпечує пошарове або об’ємне представлення внутрішніх структур організму на основі реконструкції рентгенівських проєкцій. У результаті реконструкції кожному вокселю КТ-зображення відповідає числове значення, виражене в одиницях Гаунсфілда, які характеризують відносну рентгенівську щільність тканин. Типовими орієнтирами є значення близько -1000 HU для повітря, 0 HU для води, позитивні значення для кісткових структур і значно вищі значення для металевих об’єктів.

Така фізична природа КТ робить її особливо придатною для аналізу кісток, повітроносних порожнин, кальцифікатів та інших структур із вираженою різницею щільності. Водночас вона створює низку викликів для алгоритмів штучного інтелекту, оскільки значення HU можуть змінюватися залежно від параметрів сканування, реконструкційного алгоритму, напруги трубки та наявності артефактів [12].

Окремою особливістю КТ є тривимірна структура даних, яка часто має анізотропну воксельну сітку. Просторова роздільна здатність у площині зрізу може бути значно кращою, ніж відстань між сусідніми зрізами, хоча для КТ голови зазвичай використовуються тонші зрізи, що дає змогу отримувати більш детальне представлення анатомії основи черепа. Для моделей штучного інтелекту це означає необхідність коректного врахування розміру вокселя, оскільки однакова кількість вокселів у різних дослідженнях може відповідати різним фізичним розмірам.

Додатковими проблемами є артефакти від жорсткішання пучка, часткова об'ємність, рух пацієнта, металеві артефакти від стоматологічних конструкцій або хірургічних матеріалів, а також шум у НДКТ. У задачах

сегментації ці фактори можуть призводити до розмиття меж, хибного виділення областей або втрати дрібних анатомічних деталей.

Методи штучного інтелекту в обробці КТ-зображень застосовуються для широкого спектра задач. До них належать класифікація, наприклад скринінг патологій легень або тріаж гострих станів, детекція локальних об’єктів, таких як вузлики, крововиливи або переломи, сегментація органів і патологічних ділянок, реконструкція зображень для зменшення шуму або роботи з неповними проєкціями, а також радіомічний аналіз для прогнозування клінічних результатів. Систематичні огляди сучасних застосувань штучного інтелекту в КТ підкреслюють, що найбільший практичний потенціал мають методи, які поєднують автоматизований аналіз з підтримкою клінічного рішення, підвищенням відтворюваності та оптимізацією робочого процесу радіолога [12].

У клінічній практиці КТ-аналіз на основі штучного інтелекту активно розвивається в онкології, кардіології, травматології, оториноларингології та щелепно-лицевій діагностиці. Для теми цієї роботи особливе значення має автоматична сегментація параназальних пазух і, зокрема, клиноподібної пазухи. У дослідженні Whangbo та співавторів було запропоновано багатокласову 3D-сегментаційну модель для чотирьох областей параназальних пазух, а саме лобової, решітчастої, клиноподібної та верхньощелепної пазух, на основі КТ-зображень пацієнтів із синуситом [13]. Автори окремо підкреслюють клінічну значущість точної сегментації параназальних пазух для зменшення ризику хірургічних ускладнень і використання в системах хірургічної навігації. У більш новій роботі Song та співавторів було проведено порівняння CNN, Vision Transformer і гібридних мереж для сегментації параназальних пазух на КТ-зображеннях, де оцінювалися як метрики точності, так і обчислювальна ефективність моделей [14].

Клиноподібна пазуха є клінічно релевантною, але порівняно рідко окремо досліджуваною структурою в ШІ-літературі. Все це пояснюється її

малим розміром, складною формою, значною варіабельністю пневматизації та близькістю до критичних анатомічних структур основи черепа.

Рисунок 1.2 – Приклад зрізу КТ-знімка черепа з виділеною клиновидною пазухою

Переваги застосування штучного інтелекту в КТ полягають у швидкості, відтворюваності, можливості цілодобового використання та формуванні об’єктивних кількісних показників. Для задачі сегментації це означає, що модель може автоматично створювати маску цільової структури, оцінювати її об’єм, просторове положення та співвідношення з іншими анатомічними зонами. Водночас впровадження таких методів у клінічну практику пов’язане з низкою обмежень. Найбільш суттєвими є зсув домену між різними сканерами та протоколами, обмеженість якісно анотованих даних, потреба в зовнішній валідації, складність інтерпретації помилок моделі та питання відповідальності за клінічне рішення. Саме тому штучний інтелект у КТ доцільно розглядати не як повну заміну фахівця, а як інструмент стандартизації, прискорення та кількісного посилення діагностичного процесу.

Отже, потенціал штучного інтелекту для обробки КТ-зображень полягає не лише у підвищенні точності окремих алгоритмів, а й у зміні всієї логіки роботи з томографічними даними. Для клиноподібної пазухи це особливо важливо, оскільки її точне виділення може бути корисним для передопераційного планування транссфеноїдального доступу, аналізу

анатомічних варіантів і зниження ризику помилок під час втручань у ділянці основи черепа. У зв’язку з цим подальший розгляд у другому розділі доцільно спрямувати на методи сегментації, починаючи з класичних підходів як базових орієнтирів і переходячи до CNN-архітектур, здатних працювати з повним тривимірним контекстом КТ-об’єму.

1.4 Визначення мети та основних завдань дослідження

У першому розділі розглянуто сутність сегментації медичних зображень, особливості КТ-даних і клінічне значення автоматичного виділення анатомічних структур. Показано, що клиноподібна пазуха є складним об’єктом для сегментації через малий розмір, індивідуальну варіабельність форми та близькість до критичних структур основи черепа. Аналіз сучасного стану галузі підтвердив доцільність застосування методів глибинного навчання для цієї задачі та необхідність урахування повного тривимірного контексту КТ-об’єму.

У типовому сценарії розробки моделі автоматичної сегментації цільової анатомічної структури вважається, що еталонна розмітка достатньо точно відповідає реальній анатомії, а основна відповідальність за якість кінцевої моделі покладається на вибір архітектури, обсяг навчальної вибірки та налаштування процедури навчання. Проте попередній аналіз контрольної вибірки клиноподібної пазухи, виконаний у межах роботи, показав, що це припущення не відповідає дійсності для значної частини доступних даних: близько 15 % випадків мають явні дефекти розмітки, а ще приблизно 23 % характеризуються сумнівною якістю. У такому випадку виміряні на тестовому наборі показники моделі відображають не лише її властивості, а й похибку самих еталонних масок, тому стандартна логіка послідовного навчання моделі та подальшого виміру значення метрики Dice [15] втрачає коректність у тому вигляді, в якому вона зазвичай застосовується. Це створює потребу у поєднанні класичного дослідження

моделей сегментації КТ з процедурою попередньої обробки даних, що враховує неоднорідну якість наявної експертної розмітки.

Об'єктом дослідження є моделі згортальних нейронних мереж глибинного навчання для тривимірної сегментації зображень комп'ютерної томографії.

Предметом дослідження є методи попередньої обробки даних та оцінювання якості розмітки при навчанні CNN-моделей сегментації КТ-зображень клиноподібної пазухи, зокрема ансамблеві підходи до автоматичного виявлення потенційно дефектних випадків у наявних анотаціях.

Окремої уваги потребує термінологічне уточнення. У літературі з виявлення помилок розмітки терміном детектор традиційно позначають програмну процедуру, яка зіставляє кожному прикладу з набору даних числову оцінку підозрілості його розмітки на основі узгодженості прогнозів незалежно навчених моделей [16; 17]. У межах даної роботи така процедура реалізована на основі ансамблю тривимірних згорткових мереж із диверсифікованими архітектурними передумовами; її результатом є відносне ранжування кейсів за ризиком наявності дефекту в експертній розмітці, а сама процедура надалі для стислості позначається як детектор аномалій розмітки.

Метою магістерської кваліфікаційної роботи є дослідження моделей згортальних нейронних мереж глибинного навчання для сегментації зображень комп'ютерної томографії та обґрунтування процедури попередньої обробки даних, яка враховує неоднорідну якість наявної експертної розмітки.

Для досягнення цієї мети необхідно виконати такі задачі:

– провести аналіз сучасних моделей згортальних нейронних мереж глибинного навчання для сегментації медичних зображень та обґрунтувати вибір тривимірних архітектур, придатних для використання як інструмент попередньої обробки даних КТ;

– проаналізувати анатомічні особливості клиноподібної пазухи як цільової структури автоматичної сегментації на КТ;

– виконати попередню обробку даних КТ-досліджень та провести аналіз якості наявної експертної розмітки контрольної вибірки;

– запропонувати методологічно обґрунтований підхід до автоматичного виявлення потенційно дефектних випадків розмітки на етапі попередньої обробки даних з використанням ансамблю CNN-моделей сегментації;

– провести серію експериментів з обраним набором моделей CNN на даних КТ та емпірично перевірити запропонований підхід до контролю якості розмітки як складову етапу попередньої обробки;

– здійснити аналіз отриманих результатів сегментації та якості розмітки, оцінити ефективність обраних моделей CNN та визначити напрямки подальшого вдосконалення процесу попередньої обробки даних і навчання моделей.

2 МЕТОДИ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ

2.1 Аналіз класичних методів та моделей сегментації медичних зображень

Класичні методи сегментації медичних зображень об'єднує спільний принцип: рішення про належність вокселя до цільової області приймається на основі заздалегідь сформульованого правила, яке явно задає критерій належності, а саме інтенсивнісний, морфологічний, геометричний або статистичний. На відміну від нейромережевих моделей, де відповідне відображення формується автоматично під час навчання за рахунок мінімізації функції втрат на анотованих даних, у класичних підходах дослідник самостійно визначає аналітичну форму вирішального правила та параметри його застосування. Така постановка задачі формує дві принципові властивості цього класу методів: високу інтерпретованість кожного локального рішення та обмежену здатність до узагальнення на нові клінічні умови, коли реальні характеристики зображень відхиляються від припущень, закладених у модель [15, 16].

Класичні методи доцільно розглядати як послідовність підходів зі зростаючим обсягом інформації, що враховується при формуванні рішення. На найнижчому рівні рішення базується виключно на властивостях окремого вокселя; на наступних рівнях у правило послідовно входять інформація про сусідні вокселі, статистична структура цілого зображення, обмеження на форму цільової межі, попередні анатомічні знання та зрештою результати навчання на ручно сконструйованих ознаках.

2.1.1 Методи на основі локальних інтенсивнісних критеріїв

Найпростіший клас вирішальних правил враховує лише значення інтенсивності у конкретному вокселі. Для зображення $I(x)$, де x позначає просторову координату вокселя, бінарна сегментаційна маска формується через порівняння з пороговим значенням:

$$M(x) = \begin{cases} 1, & I(x) \geq T \\ 0, & I(x) < T \end{cases} \tag{2.1}$$

де T– порогове значення.

Питання вибору параметра T розв'язується або експертним заданням значення з фізичним змістом (наприклад, порога між повітрям і м'якими тканинами на КТ), або автоматичним його обчисленням з гістограми зображення. У методі Оцу T обирається таким, що максимізує міжкласову дисперсію інтенсивностей:

$$\sigma_B^2(T) = \omega_0(T) \cdot \omega_1(T) \cdot [\mu_0(T) - \mu_1(T)]^2 \tag{2.2}$$

де $\omega_0(T)$, $\omega_1(T)$– частки пікселів двох класів, отриманих при заданому порозі;

$\mu_0(T)$, $\mu_1(T)$– середні значення інтенсивності в цих класах.

Альтернативним інтенсивнісним підходом є виявлення меж об'єкта через локальну зміну значень. Норма градієнта зображення кількісно характеризує крутизну зміни інтенсивності та використовується операторами Sobel, Canny та Laplacian-of-Gaussian для побудови карт країв:

$$\| \nabla I(x) \| = \sqrt{\left(\frac{\partial I}{\partial x_1}\right)^2 + \left(\frac{\partial I}{\partial x_2}\right)^2 + \left(\frac{\partial I}{\partial x_3}\right)^2} \tag{2.4}$$

Спільним обмеженням усіх інтенсивнісних правил є те, що рішення про вокселі формуються незалежно одне від одного. У задачі сегментації клиноподібної пазухи це призводить до принципової неоднозначності: усі повітроносні порожнини черепа, такі як клиноподібна пазуха, носова порожнина, решітчасті комірки, зовнішнє повітря, мають практично ідентичні значення HU поблизу −1000, а тонкі кісткові перегородки між ними можуть мати локальні дефекти або природні сполучення (наприклад, sphenoethmoidal recess). Як наслідок, порогова обробка коректно ідентифікує всі повітроносні ділянки разом, але не здатна автоматично відокремити саме клиноподібну пазуху від суміжних структур, а крайові методи можуть пропускати тонкостінні ділянки, де градієнт послаблений артефактами часткового об'ємного ефекту.

2.1.2 Методи з урахуванням просторової зв'язності

Природним розширенням локальних правил є включення інформації про сусідні вокселі. У методі нарощування областей із початкової точки маска формується ітеративним розширенням стартового зерна за критерієм інтенсивнісної подібності до вже виділеної області. Нехай $R^{(k)}$ позначає область на k-му кроці алгоритму, а μ_R її середню інтенсивність. Тоді сусідній воксель x включається до області, якщо виконується умова [20]:

$$| I(x) - \mu_R | < \tau \tag{2.3}$$

де τ– допустимий поріг відхилення інтенсивності.

У тривимірному випадку додатково задається тип сусідства 6-, 18- або 26-зв'язність, що визначає, скільки потенційних сусідів розглядається на кожному кроці розширення. Значно загальніше формулювання задачі дають графові методи, які подають сегментацію як задачу оптимізації на повному графі, де кожен воксель є вершиною, а ребра кодують просторову та

інтенсивнісну подібність сусідніх вершин. Підсумкова мітка кожної вершини обирається з умови мінімізації енергетичного функціонала:

$$E(L) = \sum_i D_i\,(L_i) + \lambda \sum_{(i,j)\in N} V_{ij}\,(L_i, L_j) \qquad (2.7)$$

де L_i – мітка вершини i;

$D_i(L_i)$ – штраф за невідповідність вершини певному класу за локальними даними;

$V_{ij}(L_i, L_j)$ – штраф за різні мітки у подібних сусідніх вершин;

λ – параметр балансу між локальною відповідністю даним і просторовою гладкістю;

N – множина пар сусідніх вершин.

Включення просторового члена у критерій (2.7) дозволяє формувати цілісні області навіть там, де локальна інтенсивність неоднозначна, що частково розв'язує проблему фрагментації виходу. Проте обидва підходи цього класу залишають невирішеною критичну для клиноподібної пазухи проблему: ані критерій (2.3), ані функціонал (2.7) не містять явного механізму для відсічення тонких сполучень між пазухою та суміжними повітроносними структурами. Алгоритм нарощування областей при стартовому зерні всередині пазухи з високою ймовірністю «протече» крізь природне сполучення з носовою порожниною, оскільки інтенсивність по цьому шляху є однорідною, а графовий розріз обере той розріз, що мінімізує загальну енергію без розуміння, де саме повинна закінчуватись цільова анатомічна структура.

2.1.3 Методи статистичного розкладу зображення

Окремий клас підходів формулює сегментацію як задачу розбиття множини вокселів на групи зі схожими характеристиками без попереднього

задання порогів. У методі k-середніх така задача розв'язується мінімізацією сумарної квадратичної відстані вокселів до центрів кластерів:

$$J_{km} = \sum_{i=1}^{k} \sum_{x \in C_i} \| f(x) - c_i \|^2 \quad (2.5)$$

де $f(x)$– вектор ознак вокселя x;

C_i – i-й кластер;

c_i – центр i-го кластера;

k – кількість кластерів.

Узагальненням цього підходу є нечітка кластеризація, у якій кожен воксель характеризується ступенем належності до кожного кластера:

$$J_{fcm} = \sum_{i=1}^{k} \sum_{x} u_i\,(x)^m \cdot \| f(x) - c_i \|^2, m > 1 \quad (2.6)$$

де $u_i(x) \in [0,1]$ – ступінь належності вокселя x до кластера i, з умовою $\sum_i u_i\,(x) = 1$;

m – параметр нечіткості, що регулює різкість переходу між кластерами.

Перевагою кластерних методів є відсутність явних порогів і здатність моделювати плавні переходи між тканинами, що відповідає фізичній природі КТ-зображень. Проте кластеризація формує групи вокселів за статистичною подібністю ознак. Вокселі клиноподібної пазухи, носової порожнини та зовнішнього повітря потрапляють в один кластер за критерієм інтенсивності, оскільки всі вони відповідають повітрю. Розв'язання цієї проблеми потребує додаткової фільтрації компонентів зв'язності, морфологічної обробки або ручного відбору цільової групи, що повертає метод у напівавтоматичний режим [21].

2.1.4 Методи з обмеженнями на форму межі

Спільною слабкістю розглянутих вище класів є відсутність явного контролю над геометричними властивостями отриманої межі, таких як гладкість, неперервність, локальна кривизна. Деформівні моделі усувають цю прогалину, формулюючи сегментацію як задачу еволюції контура (у двовимірному випадку) або поверхні (у тривимірному) під дією комбінації внутрішніх і зовнішніх сил. У моделі Чан-Везе, що не потребує наявності сильного градієнта на межі, мінімізується енергетичний функціонал [22]:

$$E(C, c_1, c_2) = \mu \cdot \mathrm{L}(C) + \lambda_1 \int_{\mathrm{in}(C)} (I - c_1)^2 dx + \lambda_2 \int_{\mathrm{out}(C)} (I - c_2)^2 dx \tag{2.8}$$

де C – контур, що розділяє внутрішню та зовнішню області;

c_1, c_2 – середні інтенсивності всередині та зовні контуру;

$\mu, \lambda_1, \lambda_2$ – параметри регуляризації.

У тривимірних задачах поверхня сегментації зазвичай задається імпліцитно через нульовий рівень допоміжної функції $\phi(x, t)$, еволюція якої описується рівнянням:

$$\frac{\partial \phi}{\partial t} + F \cdot \| \nabla \phi \| = 0 \tag{2.9}$$

де F– швидкість руху поверхні в напрямку нормалі, що залежить від локальних характеристик зображення та параметрів регуляризації.

Введення геометричної регуляризації у функціонали (2.8)-(2.9) дозволяє отримувати гладкі замкнені поверхні навіть на зашумлених даних. Однак у задачі сегментації клиноподібної пазухи саме геометричні припущення можуть стати джерелом помилок. Алгоритм формально продовжує мінімізувати енергетичний функціонал та зберігати гладкість

поверхні навіть тоді, коли контур через тонке кісткове сполучення вже перейшов у носову порожнину та отриманий розв'язок є локально оптимальним з точки зору формули (2.8), але анатомічно некоректним. Крім того, кінцевий результат суттєво залежить від початкової ініціалізації контуру, що у клінічних умовах вимагає інтерактивного втручання оператора.

2.1.5 Методи з використанням анатомічних знань

Усі попередні підходи виводять рішення безпосередньо з зображення, що сегментується, без явного посилання на еталонну анатомію. Атласні методи зміщують цей баланс. Вони використовують попередньо розмічений анатомічний зразок як джерело знань про типове розташування та форму цільових структур. Сегментація нового об'єму $I_{\text{нов}}$ зводиться до пошуку просторового перетворення ϕ, що приводить атласне зображення $I_{\text{атл}}$ у відповідність до нового, після чого мітки атласу переносяться через знайдене перетворення:

$$\phi^* = \arg\min_{\phi} [S(I_{\text{нов}}, I_{\text{атл}} \circ \phi) + \lambda \cdot R(\phi)] \qquad (2.10)$$

де S – міра подібності між зображеннями;

$R(\phi)$ – регуляризатор, що обмежує допустимі деформації;

λ – параметр балансу між якістю реєстрації та гладкістю деформації.

У багатоатласних модифікаціях використовується кілька попередньо розмічених зразків, мітки з яких об'єднуються методами злиття на зразок STAPLE, що дозволяє зменшити вплив помилок окремого атласу. Для задачі клиноподібної пазухи атласний підхід стикається з принциповим обмеженням анатомічної варіабельності: типи пневматизації від конхального до постселлярного формують настільки різні форми

порожнини, що жоден фіксований атлас не покриває весь спектр реальних випадків, а використання багатьох атласів збільшує обчислювальну складність без гарантії повного охоплення варіативності.

2.1.6 Методи статистичного навчання з ручними ознаками

Логічним завершенням еволюції класичних підходів є системи, у яких вирішальне правило формується не аналітично дослідником, а через статистичне навчання на анотованих даних з використанням заздалегідь сконструйованих ознак. У моделі випадкового лісу ймовірність належності вокселя до цільового класу обчислюється як середнє по ансамблю дерев рішень:

$$P(L = c \mid f(x)) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[\, h_t(f(x)) = c] \tag{2.11}$$

де T– кількість дерев у ансамблі;

$h_t(f(x))$ – прогноз t-го дерева для вокселя з вектором ознак $f(x)$;

$\mathbb{I}[\cdot]$ – індикаторна функція класу.

Ключова відмінність цього класу від попередніх полягає у тому, що структура вирішального правила h_t отримується автоматично з тренувальних даних, а не задається явно. Однак вектор ознак $f(x)$, в який входять інтенсивність, координати, локальні статистики, текстурні дескриптори (GLCM або LBP) залишається продуктом ручного інженерного вибору. Усе це створює асиметрію, де класифікатор може вивчити складні нелінійні залежності у заданому ознаковому просторі, але не здатен переглянути сам цей простір, якщо ознаки виявляться неоптимальними. Саме на цьому місці пролягає принципова межа між класичним статистичним навчанням і глибинним. Згорткові нейромережі формують

ієрархію ознак автоматично, спираючись лише на сирі вокселі та анотовані маски.

Розглянуті класи методів утворюють послідовну прогресію зростання інформаційної бази вирішального правила. Від інтенсивності окремого вокселя через врахування локальної зв'язності, статистичних закономірностей зображення, обмежень на форму межі, попередніх анатомічних знань до автоматичного навчання у просторі ручно сконструйованих ознак. На кожному рівні цієї прогресії розв'язується одне з обмежень попереднього рівня, але водночас вводиться нове джерело залежності. Жоден з підходів не пропонує механізму, що одночасно враховував би тривимірну зв'язність, нелінійну структуру даних, форму цільової межі та анатомічний контекст без явного задання відповідних компонентів дослідником.

Для задачі сегментації клиноподібної пазухи на КТ-зображеннях цей кумулятивний дефіцит проявляється особливо різко. Цільова структура одночасно має мінімальний інтенсивнісний контраст з суміжними повітроносними порожнинами, тонкі та нерідко перфоровані кісткові межі, високу варіабельність форми між пацієнтами та тривимірну топологію зі складними рецесами і перегородками. Жоден з класичних методів не в змозі коректно об'єднати всі ці аспекти у межах однієї моделі, що фактично робить їх застосування для основної задачі дослідження непрактичним.

Водночас було б некоректно стверджувати, що класичні методи у сучасних медичних пайплайнах втратили значення. Вони залишаються корисними як інструменти попередньої обробки (нормалізація інтенсивностей, морфологічні операції на кісткових масках), формування початкових масок для напівавтоматичного анотування, фільтрації компонентів зв'язності у постобробці результатів нейромережевих моделей та як інтерпретовані базові орієнтири для оцінки приросту точності, що забезпечується більш складними підходами. Саме у такому допоміжному контексті їх роль доцільно враховувати при побудові повного пайплайну

сегментації, тоді як основне вирішальне правило має формуватися методами, здатними автоматично навчати тривимірне просторове представлення з анотованих даних — згортковими нейронними мережами глибинного навчання, розгляду яких присвячені наступні підрозділи.

2.2 Методи та моделі сегментації на основі нейромереж глибинного навчання із фокусом на 2D та 2.5D

На відміну від класичних підходів, розглянутих у попередньому підрозділі, де правила сегментації задаються явно через інтенсивнісні, морфологічні або графові критерії, у методах глибинного навчання відображення між вхідним зображенням і вихідною маскою формується автоматично через мінімізацію функції втрат на навчальній вибірці.

Формально задача може бути подана як пошук параметрів θ моделі f_θ, які мінімізують емпіричний ризик:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(X_i), Y_i) \tag{2.12}$$

де (X_i, Y_i) – пари «зображення–маска» з навчальної вибірки;

ℓ – обрана функція втрат, наприклад Dice, бінарна перехресна ентропія або їх комбінація.

Залежно від розмірності вхідного представлення X та структури згорткових операцій усередині f_θ виокремлюють три фундаментально різні класи моделей: двовимірні (2D), псевдотривимірні (2.5D) та тривимірні (3D). У цьому підрозділі розглядаються перші два класи, тоді як 3D-підходи, як основні для практичної частини роботи, винесені в окремий підрозділ 2.3.

Двовимірні моделі обробляють тривимірний КТ-об'єм $V \in \mathbb{R}^{H \times W \times D}$ як упорядковану послідовність із D незалежних аксіальних зрізів $\{S_z\}_{z=1}^{D}$, де

$S_z = V[:,:,z] \in \mathbb{R}^{H\times W}$. Процес сегментації об'єму у такому випадку формалізується послідовністю трьох кроків:

$$S_z \to f_\theta^{2D}(S_z) = M_z \to V_M = \text{stack}(M_1, M_2, \dots, M_D) \quad (2.13)$$

де $M_z \in \{0, \dots, K\}^{H\times W}$ – двовимірна маска для z-го зрізу;

$V_M \in \{0, \dots, K\}^{H\times W\times D}$ – реконструйована тривимірна маска.

Принциповою особливістю такого процесу є те, що передбачення для зрізу S_z формується незалежно від сусідніх зрізів S_{z-1}та S_{z+1}, тобто модель не має прямого доступу до інформації про просторову неперервність структури вздовж осі Z. Усі згорткові операції всередині f_θ^{2D}є двовимірними, що значно зменшує кількість параметрів та обчислювальних затрат у порівнянні з 3D-аналогами. Виходячи з цього, виникая можливість використовувати більший розмір батчу, застосовувати передтреновані на ImageNet кодувальники та фактично розгортати тренувальну вибірку, оскільки одна КТ-серія з D зрізами надає D незалежних тренувальних прикладів замість одного об'ємного.

Псевдотривимірні моделі займають проміжне положення між 2D і 3D і реалізують компроміс між обчислювальною економністю та збереженням просторового контексту. У літературі виокремлюють три основні стратегії 2.5D-обробки, які різняться способом подання сусідньої інформації моделі.

Перша стратегія – мультивидове об'єднання. У цій схемі тривимірний об'єм представляється як три незалежні набори двовимірних зрізів у трьох ортогональних площинах:

$$S_z^{ax} = V[:,:,z], S_y^{co} = V[:,y,:], S_x^{sa} = V[x,:,:] \quad (2.14)$$

На кожній із площин навчається окрема 2D-модель f_θ^{ax}, f_θ^{co}, f_θ^{sa}, після чого їх передбачення для одного й того ж вокселя об'єднуються.

Найпростішим варіантом об'єднання є зважене усереднення ймовірностей за класами:

$$p(v) = \alpha \cdot p^{ax}(v) + \beta \cdot p^{co}(v) + \gamma \cdot p^{sa}(v), \alpha + \beta + \gamma = 1 \qquad (2.15)$$

Більш складні варіанти використовують окрему мережу злиття, яка приймає три набори ймовірностей як вхід і формує підсумкову маску. Такий підхід дозволяє врахувати ознаки, що краще проявляються в одній із площин, наприклад поздовжню структуру судини або поверхневий рельєф органа, однак потребує тренування трьох моделей замість однієї та узгодження їх результатів.

Друга стратегія полягає у використанні сусідніх зрізів як каналів зображення. Для передбачення маски центрального зрізу S_z на вхід 2D-моделі подається стек із $(2k+1)$ послідовних зрізів, що трактуються як канали:

$$X_z = \text{concat}(S_{z-k}, \dots, S_z, \dots, S_{z+k}) \in \mathbb{R}^{H \times W \times (2k+1)} \qquad (2.16)$$

Така структура дає змогу моделі бачити обмежений контекст уздовж осі Z без переходу до повноцінних тривимірних згорток. Параметр k зазвичай обирається в межах від одного до трьох, що відповідає стеку з 3, 5 або 7 зрізів. Перевагою цього підходу є збереження ефективності 2D-навчання: модель залишається повністю двовимірною за структурою згорток, а додатковий контекст входить через канальний вимір без зростання кількості параметрів у глибших шарах.

Третя стратегія – гібридні 2D та 3D-архітектури, у яких згортки явно розщеплюються за просторовими осями. Усередині зрізу застосовуються двовимірні згортки розміру $3 \times 3 \times 1$, а між зрізами – окремі одновимірні згортки $1 \times 1 \times k_z$. Подібна до 3D-обробки за значно меншої кількості параметрів, оскільки повна тривимірна згортка $3 \times 3 \times 3$ може бути

замінена послідовною парою операторів $(3 \times 3 \times 1)$ і $(1 \times 1 \times 3)$. Прикладами таких архітектур є MNet, що адаптивно об'єднує мілкі та глибокі представлення для роботи з анізотропними даними [23], та Z-Net, яка явно враховує різну роздільну здатність уздовж різних осей через спеціалізовані блоки згорток [24].

Доцільність застосування 2D і 2.5D-підходів безпосередньо пов'язана з геометричними характеристиками вхідних даних. Кількісно ступінь анізотропії воксельної сітки можна охарактеризувати співвідношенням розмірів вокселя за різними осями:

$$R_{aniso} = \frac{\max(s_x, s_y, s_z)}{\min(s_x, s_y, s_z)} \tag{2.17}$$

де s_x, s_y, s_z – фізичні розміри вокселя за відповідними осями.

У літературі прийнято вважати дані сильно анізотропними при $R_{aniso} > 3$. У такому режимі застосування ізотропних 3D-згорток стикається з низкою принципових обмежень.

По-перше, оператор пулінгу $2 \times 2 \times 2$, що зменшує просторову розмірність на кожному рівні кодувальника, однаково зменшує кількість зрізів вздовж низькороздільної осі, що при невеликій початковій кількості зрізів призводить до швидкої втрати структурної інформації.

По-друге, рецептивне поле 3D-згортки в анізотропному просторі охоплює асиметричний фізичний обсяг: одне й те саме ядро $3 \times 3 \times 3$ покриває суттєво більший фізичний відстань по осі Z, ніж по осях X та Y, що може призводити до неконтрольованого змішування ознак із різних анатомічних рівнів.

По-третє, при невеликій кількості зрізів тривимірна згортка має обмежену кількість позицій для застосування вздовж осі Z, що ускладнює процес навчання просторових залежностей у цьому напрямку.

Визначення оптимального класу архітектури для задачі сегментації клиноподібної пазухи, що розглядається у роботі, потребує врахування фактичних характеристик вхідних даних. Початкове співвідношення розмірів вокселя у досліджуваному наборі становить приблизно 1:1:8, що формально відповідає режиму сильної анізотропії за критерієм (2.17). Однак стандартний крок попередньої обробки передбачає ресемплінг до уніфікованого просторового кроку, після якого ефективна анізотропія знижується до значень близько 1:1:3 і знаходиться на межі вказаного критерію. Крім того, клиноподібна пазуха є замкненою об'ємною структурою з тонкими кістковими перегородками, рецесами та значно змінною формою у різних типах пневматизації. Її коректне відновлення потребує неперервної тривимірної поверхні без розривів між сусідніми зрізами, які можуть формуватися при незалежній 2D-обробці та важко усуваються методами постобробки.

Слід також врахувати, що систематичні порівняння 2D, 2.5D і 3D-архітектур на стандартизованих наборах медичних даних, включно з тими, що мають анізотропну структуру, вже проведені у межах публічних бенчмарків і узагальнених оглядів, на які зроблено посилання у підрозділі 1.2. Через це проведення власного аналогічного порівняння у межах дипломної роботи на меншому за обсягом датасеті не додало б статистичної значущості порівняно з уже наявними результатами. Натомість обґрунтованим є вибір 3D-підходу як основного, оскільки він дозволяє повноцінно використовувати об'ємний контекст у межах автоматично налаштованого фреймворку, здатного адаптивно враховувати залишкову анізотропію вхідних даних на рівні плану навчання.

Таким чином, 2D і 2.5D-методи становлять зрілу дослідницьку галузь із розвиненим математичним апаратом і широким набором архітектур, проте для розглянутої задачі їх використання визнано недоцільним з огляду на анатомічні особливості цільової структури та технічні параметри даних.

Подальший розгляд у підрозділі 2.3 присвячено 3D-моделям сегментації як основі для практичної частини дослідження.

2.3 Методи та моделі сегментації на основі нейромереж глибинного навчання для 3D сегментації

Розглянуті у попередньому підрозділі двовимірні та псевдотривимірні підходи реалізують компроміс між обчислювальною економністю та обробкою об'ємного контексту, проте за наявності КТ-об'ємів зі складною тривимірною морфологією цільової структури їх застосування призводить до втрати частини просторової інформації. У цьому підрозділі розглядаються тривимірні моделі глибинного навчання, які виконують безпосередньо об'ємні операції над вхідним об'ємом і є основним архітектурним класом для практичної частини дослідження. На відміну від 2D і 2.5D-моделей, де згорткові операції розщеплюються за окремими площинами або реалізуються через канальне представлення сусідніх зрізів, тривимірні архітектури обробляють об'єм як єдиний геометричний об'єкт із збереженням ізотропної просторової взаємодії між сусідніми вокселями за всіма трьома осями.

Формально тривимірна сегментаційна модель f_θ^{3D} задає відображення вхідного об'єму у відповідну воксельну маску без проміжного розщеплення на зрізи. Базовою операцією усередині такої моделі є тривимірна згортка, яка для вхідного тензора $X \in \mathbb{R}^{C_{in} \times H \times W \times D}$ та ядра $K \in \mathbb{R}^{C_{out} \times C_{in} \times k_h \times k_w \times k_d}$ формує вихідний тензор за правилом:

$$Y_{c,h,w,d} = \sum_{c'=1}^{C_{in}} \sum_{i=1}^{k_h} \sum_{j=1}^{k_w} \sum_{l=1}^{k_d} K_{c,c',i,j,l} \cdot X_{c',h+i,w+j,d+l} \tag{2.18}$$

де C_{in}, C_{out}– кількість вхідних і вихідних каналів ознак;

k_h, k_w, k_d– розміри ядра за просторовими осями;

h, w, d– координати вокселя у вихідному тензорі.

На відміну від двовимірного випадку, у формулі (2.18) сумування виконується одночасно за трьома просторовими вимірами, що забезпечує безпосереднє врахування вокселів у локальній тривимірній околиці. Як наслідок, кількість параметрів і обчислювальна складність зростають лінійно за третім просторовим виміром у порівнянні з двовимірним аналогом, що формує основне обмеження тривимірних моделей. Для типового ядра розміру $3 \times 3 \times 3$ кількість мультиплікативних операцій у $k_d = 3$ рази більша, ніж для двовимірного ядра 3×3, а вимоги до обсягу графічної пам'яті залежать не лише від параметрів моделі, а й від розміру тензорів проміжних активацій, що зростають кубічно з лінійним розміром вхідного об'єму. Усе це формує два ключові інженерні наслідки:

1. Необхідність роботи з невеликими розмірами батчу (зазвичай від одного до чотирьох прикладів)
2. Потреба у патч-орієнтованому навчанні на просторових фрагментах об'єму замість обробки повного КТ-дослідження за один прохід.

У сучасних дослідженнях тривимірні моделі сегментації медичних зображень можна систематизувати за п'ятьма архітектурними класами, кожен з яких відрізняється способом організації об'ємного представлення та процесом передачі інформації між рівнями моделі.

Перший клас становлять тривимірні згорткові мережі з U-подібною структурою, яка є прямим розширенням U-Net на об'ємні дані. До нього належать 3D U-Net, V-Net, а також їх модифікації із залишковими блоками.

Принциповою особливістю цих моделей є наявність кодувальника, що послідовно зменшує просторову розмірність при одночасному збільшенні кількості каналів ознак; та декодувальника, що відновлює просторову розмірність до вихідної, та skip-з'єднань, які передають високороздільні ознаки безпосередньо з відповідних рівнів кодувальника. Загальний процес

обробки об'єму у такій архітектурі може бути формалізований як послідовність операцій:

$$X \xrightarrow{E_1} F_1 \xrightarrow{E_2} F_2 \ldots \to F_L \xrightarrow{D_L} G_{L-1} \to \cdots \xrightarrow{D_1} G_0 \to M \tag{2.19}$$

де E_i – i-й блок кодувальника, що включає згорткові, нормалізаційні та активаційні операції з подальшим зменшенням просторової розмірності;

F_i – тензори ознак на i-му рівні кодувальника;

D_i – i-й блок декодувальника, що приймає на вхід об'єднання F_i зі skip-з'єднання та результат відновлення розмірності з рівня G_{i+1};

M – підсумкова воксельна маска.

Детальний математичний опис цих архітектур та функцій втрат, що використовуються при їх навчанні, наведено у підрозділі 3.1.

Другий клас представлений каскадними архітектурами, у яких сегментація виконується у два етапи з різною просторовою роздільною здатністю.

На першому етапі модель обробляє низькороздільну версію об'єму, отриману ресемплінгом до більш грубого просторового кроку, що дозволяє охопити повний анатомічний контекст у межах одного патчу.

На другому етапі результат низькороздільної сегментації подається як додатковий канал у модель повної роздільної здатності, яка уточнює межі цільової структури. Такий процес можна описати послідовністю:

$$V \xrightarrow{\text{ресемплінг}} V_{lr} \xrightarrow{f_\theta^{lr}} M_{lr} \xrightarrow{\text{передискретизація}} \widetilde{M}_{lr} \xrightarrow{\text{конкатенація з } V} f_\theta^{hr} \to M \tag{2.20}$$

де V_{lr} – низькороздільна версія об'єму;

M_{lr} – попередня сегментаційна маска, отримана у низькій роздільній здатності;

$\widetilde{M}_{lr}$ – передискретизована до повної роздільної здатності груба маска;

f_θ^{hr} – модель повної роздільної здатності.

Каскадний підхід є особливо корисним при роботі з великими КТ-об'ємами, де патч повної роздільної здатності не охоплює всю цільову структуру і модель не має доступу до глобального анатомічного контексту під час прийняття локальних рішень про сегментацію.

Третій клас становлять гібридні архітектури з трансформерним кодувальником, у яких локальні згорткові ознаки доповнюються механізмом самоуваги для моделювання довгодистанційних залежностей між частинами об'єму. У такій схемі вхідний об'єм або його патч розбивається на регулярну сітку об'ємних токенів розміру $p_h \times p_w \times p_d$, кожен з яких лінійно проєктується у вектор фіксованої розмірності. Послідовність токенів $T = (t_1, t_2, \dots, t_N)$, де $N = (H/p_h) \cdot (W/p_w) \cdot (D/p_d)$, обробляється трансформерним кодувальником через оператор самоуваги:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.21}$$

де Q, K, V– матриці запитів, ключів і значень, отримані лінійним проєктуванням послідовності токенів;

d_k– розмірність вектора ключа.

До цього класу належать UNETR із суто трансформерним кодувальником і згортковим декодувальником [25], Swin UNETR із ієрархічним трансформером Swin, що використовує локальні віконні механізми уваги для зменшення обчислювальної складності [26], а також nnFormer, у якому згорткові та трансформерні блоки чергуються на різних рівнях моделі [27]. Перевагою таких архітектур є здатність моделювати глобальний контекст без необхідності глибокого стека згорткових шарів, проте їх тренування потребує суттєво більших обсягів даних для досягнення стабільної точності.

Четвертий клас об'єднує масштабовані згорткові архітектури нового покоління, які запозичують структурні рішення з трансформерних мереж, але реалізують їх у межах суто згорткової парадигми. Прикладами є MedNeXt, що використовує адаптовані для медичної сегментації блоки ConvNeXt із збільшеним розміром ядра та відокремленими просторовими і канальними згортками [28], а також STU-Net, який пропонує сімейство масштабованих варіантів U-Net із кількістю параметрів від десятків мільйонів до мільярда [29].

Усі ці архітектури покликані відтворити переваги трансформерів: велике рецептивне поле та канальне змішування, без використання операції самоуваги, що зберігає обчислювальну ефективність при високій точності.

П'ятий клас представлений самоналаштовуваними фреймворками, серед яких центральне місце займає nnU-Net [6]. Принципова відмінність цього підходу від попередніх полягає в тому, що він не пропонує нову нейромережеву архітектуру, а формалізує повний процес побудови сегментаційного рішення для довільного датасету. Замість окремої моделі f_θ задається відображення:

$$D \to \mathcal{F}(D) = \pi \to f_{\theta(\pi)} \tag{2.22}$$

де D – навчальний датасет;

$\mathcal{F}$ – функція автоматичного конфігурування, що формує вектор гіперпараметрів π на основі статистичних характеристик датасету (просторовий крок вокселя, розподіл інтенсивностей, розміри об'ємів, ступінь анізотропії);

$f_{\theta(\pi)}$ – модель, побудована та навчена за згенерованою конфігурацією.

Замість одного фіксованого варіанту фреймворк автоматично формує три паралельні конфігурації: двовимірну, тривимірну повної роздільної здатності та каскадну тривимірну, після чого через 5-кратну перехресну валідацію обирається оптимальна конфігурація або їх ансамбль.

Описаний процес знімає з дослідника необхідність ручного підбору гіперпараметрів і одночасно забезпечує систематичний підхід до питань попередньої обробки, навчання, інференсу та постобробки.

Тренувальний процес тривимірних моделей принципово відрізняється від двовимірного випадку через зазначені вище обмеження пам'яті. Замість обробки повного об'єму модель навчається на просторових патчах фіксованого розміру, що випадково вирізаються з тренувальних об'ємів. Інференс на повному об'ємі реалізується методом ковзного вікна, де об'єм покривається перекривними патчами, для кожного з яких модель формує локальні передбачення, а підсумкова маска отримується агрегацією локальних передбачень із вагами, що зменшуються до меж патчу. Формально процес агрегації може бути записаний як:

$$M(v) = \frac{\sum_{i \in \mathcal{P}(v)} w_i(v) \cdot p_i(v)}{\sum_{i \in \mathcal{P}(v)} w_i(v)} \tag{2.23}$$

де $M(v)$ – підсумкова ймовірність належності вокселя vдо цільового класу;

$\mathcal{P}(v)$ – множина патчів, що містять воксель v;

$p_i(v)$ – локальне передбачення i-го патчу для вокселя v;

$w_i(v)$ – ваговий коефіцієнт, що залежить від відстані вокселя до центру патчу і зазвичай задається через гаусову функцію.

Зважування за гаусом зменшує внесок передбачень із крайових ділянок патчу, де модель має менше контекстної інформації, що покращує узгодженість сегментації на межах патчів. Додатково під час інференсу часто застосовується тестова аугментація через дзеркальні відображення об'єму за всіма допустимими осями з подальшим усередненням передбачень, що дозволяє підвищити стабільність результату на коштах багаторазового пропуску об'єму через модель.

Окремої уваги потребує вибір методу нормалізації всередині блоків кодувальника та декодувальника. У стандартних згорткових мережах для природних зображень зазвичай застосовується пакетна нормалізація (Batch Normalization), яка обчислює статистики активацій по всьому батчу. У тривимірних медичних моделях через малий розмір батчу така нормалізація стає нестабільною, оскільки оцінки середнього та дисперсії формуються лише на одному-двох прикладах. Тому стандартним вибором у 3D-сегментації є Instance Normalization, яка нормує активації незалежно для кожного прикладу та кожного каналу, що забезпечує стабільну роботу при будь-якому розмірі батчу. Аналогічні міркування зумовлюють використання активаційної функції Leaky ReLU замість звичайної ReLU для запобігання повному обнуленню градієнтів у глибоких рівнях мережі.

Емпіричні результати тривимірних моделей на публічних бенчмарках підтверджують стабільну перевагу автоматично налаштованих фреймворків над окремими спеціалізованими архітектурами. У великомасштабному порівняльному дослідженні Isensee та співавторів встановлено, що CNN-моделі у фреймворку nnU-Net досягають конкурентних або кращих результатів за актуальні трансформерні архітектури на шести 3D-наборах даних, причому KiTS, AMOS і ACDC визнано найбільш придатними для бенчмаркінгу через найвищу здатність розрізняти методи [11].

Окремо виявлено, що значна частина опублікованих переваг трансформерних або гібридних моделей виникає не за рахунок принципових архітектурних відмінностей, а через нерівні умови порівняння — слабкі базові реалізації U-Net, обмежені обчислювальні бюджети для базових моделей, відсутність належної 5-кратної валідації або тестування на занадто малих наборах даних. Аналогічна тенденція спостерігається у бенчмарках TotalSegmentator [10] та Medical Segmentation Decathlon [7], де моделі на основі nnU-Net стабільно знаходяться у топі лідербордів.

Для задачі сегментації клиноподібної пазухи, що розглядається у роботі, наявні безпосередні прецеденти успішного застосування саме

фреймворку nnU-Net на близьких клінічних задачах. У роботі Gülşen та співавторів модель nnU-Net v2 застосована для одночасної сегментації клиноподібної пазухи та структур середньої частини основи черепа на CBCT-об'ємах, причому для самої пазухи досягнуто Dice-коефіцієнта 0,96, що є найвищим показником серед усіх досліджених анатомічних структур у цій ділянці [30].

У дослідженні Alekseeva та співавторів той самий фреймворк застосовано до сегментації клиноподібної пазухи на КТ-зображеннях у задачі біометричної ідентифікації, де при тренуванні на 426 КТ-сканах із п'ятикратною перехресною валідацією отримано Dice 0,946 та HD95 близько 1,5 мм на зовнішньому тестовому наборі [31].

Обидві роботи демонструють, що автоматично сконфігурований 3D-пайплайн nnU-Net забезпечує клінічно прийнятну точність на цій конкретній анатомічній структурі без необхідності ручного підбору архітектурних і тренувальних параметрів.

З огляду на викладене, тривимірні згорткові моделі сегментації становлять основний архітектурний клас для задачі автоматичного виділення клиноподібної пазухи на КТ-зображеннях. Серед розглянутих категорій архітектур найбільш обґрунтованим вибором для практичної частини роботи є фреймворк nnU-Net у конфігурації 3D повної роздільної здатності, що зумовлено сукупністю чинників: систематичним домінуванням цього підходу на публічних бенчмарках при умовах справедливого порівняння [11], автоматичним налаштуванням повного пайплайну попередньої обробки, навчання та інференсу, що знімає необхідність ручного підбору гіперпараметрів, наявністю прямих прецедентів успішного застосування саме до сегментації клиноподібної пазухи [32], а також відкритою реалізацією фреймворку, що забезпечує відтворюваність експериментів. Виходячи з усього перерахованого, у наступному розділі роботи буде широко розглянуто математичний апарат та

повний пайплайн саме цієї архітектури як основного інструмента дослідження.

2.4 Шум розмітки в медичній сегментації та методи його виявлення

Підходи до сегментації, розглянуті в попередніх підрозділах, незалежно від того, чи йдеться про класичні алгоритми, чи про нейромережеві моделі, ґрунтуються на важливому припущенні: «Еталонна маска Y, яка використовується під час навчання та валідації, достатньо точно відображає реальну анатомічну структуру». Саме припущення розглядає мінімізацію функції втрат як процес наближення прогнозу моделі до бажаної сегментаційної маски. Однак у реальних медичних наборах даних експертна розмітка не завжди є повністю тотожною анатомічно істинній масці.

Джерела шуму у медичній розмітці можна систематизувати за чотирма основними групами. Першу групу складає інтер- та інтраіндивідуальна варіабельність експертів. Одні й ті самі КТ-зображення, розмічені різними фахівцями або тим самим фахівцем у різний час, дають Dice-узгодженість у межах 80–95% залежно від структури та досвіду оператора. До другої групи включають відмінності у протоколах розмітки, де одна частина датасету може бути розмічена за кістковою межею структури, а інша за межею слизової оболонки чи запального процесу, що формує не випадкову, а структуровану невідповідність між підмножинами. Третя група — це обмеження у часі. У клінічній та дослідницькій практиці на розмітку одного випадку зазвичай виділяється від п'яти до п'ятнадцяти хвилин, що недостатньо для ретельної тривимірної сегментації складних структур. До останньої, проте не менш важливої входять технічні дефекти. Помилки реєстрації координат між зображенням і маскою, дублювання зрізів зі зміщенням, частково завантажені або обірвані файли.

Наведена нижче таблиця 2.1 систематизує основні типи дефектів, спостережуваних у медичних датасетах для сегментації, включно з тими, що були виявлені у датасеті клиноподібної пазухи, який використовується у практичній частині дослідження.

Таблиця 2.1 — Типологія дефектів розмітки у медичній сегментації

Тип дефекту	Просторовий масштаб	Характер	Типове джерело виникнення
Воксельний шум	Окремі вокселі	Випадковий	Нестабільність інтерактивних інструментів сегментації, помилки растеризації
Граничний шум	Тонкі смуги поблизу межі	Систематичний або випадковий	Неоднозначність межі при частковому об’ємному ефекті, обмежений час розмітки
Регіональний пропуск	Цілі компоненти структури	Систематичний	Розмітка лише однієї з парних структур, пропуск складних рецесів
Регіональне розширення	Сусідні анатомічні утворення	Систематичний	Включення запаленої слизової оболонки у маску повітроносної порожнини
Об’ємний зсув	Уся маска зміщена	Технічний	Помилки реєстрації, дублювання зрізів зі зміщенням
Заміна структури	Уся маска неправильна	Системний	Розмітка сусідньої анатомічної області замість цільової
Протокольна неоднорідність	Між підмножинами датасету	Систематичний	Зміна правил розмітки між клінічними центрами або періодами

Принципово важливим є те, що різні типи дефектів вимагають різних стратегій виявлення. Воксельний і граничний шум зазвичай не можна повністю усунути, тому що вони відображають фундаментальну невизначеність процесу анотування і присутні навіть у бенчмаркових датасетах високої якості. Регіональні дефекти, об'ємні зсуви та заміни структур, навпаки, є грубими помилками, які можна і потрібно виявляти автоматично перед використанням датасету для навчання та оцінки моделей.

Наявність шуму розмітки у медичних датасетах є типовою проблемою, а не поодиноким винятком. У дослідженнях, присвячених якості експертних масок, показано, що навіть відносно невеликі дефекти на

межі об’єкта можуть помітно знижувати значення Dice-коефіцієнта, а регіональні помилки, зокрема пропуски частини структури або включення сусідніх областей, мають ще сильніший вплив на результат сегментації [33], [34]. Додатковою проблемою є те, що дві незалежні розмітки одного й того самого КТ-дослідження не завжди повністю збігаються між собою. Для різних анатомічних структур рівень узгодженості може суттєво змінюватися, а для малих або складних за формою об’єктів він зазвичай є нижчим. Усе це означає, що якість моделі не може оцінюватися без урахування похибки самої еталонної маски [33], [35].

2.4.1 Концепція досяжної стелі якості

Наявність шуму розмітки безпосередньо впливає на коректність оцінювання моделей сегментації. Виміряна на тестовому наборі метрика розбіжності прогнозу з розміткою концептуально є сумою двох компонент: істинної похибки моделі відносно анатомічно правильної маски та внеску шуму розмітки у виміряне значення. Формально цю декомпозицію можна записати як:

$$\mathcal{E}_{\text{виміряна}} = \mathcal{E}_{\text{модель}} + \Delta_{\text{шум}} \tag{2.24}$$

де $\mathcal{E}_{\text{виміряна}}$ – виміряна похибка моделі відносно доступної розмітки;

$\mathcal{E}_{\text{модель}}$– істинна похибка моделі відносно анатомічно правильної маски;

$\Delta_{\text{шум}}$– внесок шуму розмітки у виміряну метрику.

Прямим наслідком декомпозиції (2.24) є те, що навіть ідеальна модель з нульовою істинною похибкою матиме ненульове виміряне значення, обмежене внеском шуму. У результаті формується обмеження на максимально можливе значення метрики для тестового набору із

зашумленою розміткою. Іншими словами, навіть модель із високою здатністю до узагальнення не зможе стабільно перевищити рівень, який задається не її архітектурою, а точністю, повнотою та узгодженістю еталонних масок. Тому за наявності дефектів розмітки виміряна якість сегментації відображає не лише властивості моделі, а й похибки самого тестового набору.

З цього випливають дві принципові проблеми оцінювання моделей сегментації під шумом розмітки:

1. Значна частка декларованого розриву між сучасними алгоритмами та клінічно прийнятним рівнем точності може бути зумовлена не недосконалістю архітектури, а шумом розмітки.
2. Різні моделі по-різному реагують на шум, тому порівняння на однаковому зашумленому наборі може давати рейтинг, що не відповідає порівнянню на чистих даних.

Обидві проблеми мотивують необхідність систематичного виявлення зашумлених прикладів у тестових наборах перед публікацією результатів.

2.4.2 Підходи до виявлення зашумленої розмітки

У літературі сформувалося чотири принципово різні підходи до автоматичного виявлення зашумлених міток у задачах сегментації, що відрізняються джерелом сигналу, який використовується для оцінки якості розмітки. Узагальнене порівняння цих підходів за ключовими властивостями наведено у таблиці 2.2, а їхній зміст детальніше розглянуто нижче.

Таблиця 2.2 — Порівняльна характеристика підходів до виявлення шуму розмітки

Підхід	Джерело сигналу	Розмітка	Моделі	Перевага	Обмеження
Аналіз метричних відхилень	Відхилення метрики для прикладу	Так	1	Низька вартість	Не відрізняє шум від складного випадку
Confident Learning	Незгода прогнозу з розміткою	Так	1	Формалізований підхід	Залежить від калібрування
Неузгодженість ансамблю	Розбіжність між прогнозами	Ні	Декілька	Не потребує розмітки	Висока обчислювальна вартість
Попередньо навчений оракул	Незгода зовнішньої моделі з розміткою	Так	1	Незалежний від шуму набору	Залежить від якості оракула

Першим розглядається аналіз метричних відхилень, який є найпростішим за реалізацією. Для всіх прикладів датасету обчислюють метрику розбіжності між прогнозом моделі та розміткою, зазвичай Dice, HD95 або їх комбінацію. Після цього приклади з аномально високими значеннями цієї метрики розглядають як кандидати на наявність шуму. Основне обмеження підходу полягає в тому, що він не дозволяє розрізнити дві різні причини високої похибки. До першої належать випадки із зашумленою розміткою, до другої належать приклади з коректною розміткою, але об’єктивно складною анатомією, з якою модель не справляється. Унаслідок цього серед виявлених аномалій може з’являтися значна кількість анатомічно складних, але правильно розмічених випадків.

Другий підхід полягає в тому що, зашумлені приклади характеризуються не лише високою похибкою моделі, а й поєднанням двох ознак, а саме високою впевненістю моделі у власному прогнозі та істотною розбіжністю цього прогнозу з наявною розміткою. Для класифікаційних задач підхід реалізується через побудову матриці спільних ймовірностей між істинними та зашумленими мітками, а для сегментації його

узагальнюють на воксельний рівень. Принципово важливою умовою є використання out-of-fold прогнозів, отриманих за допомогою перехресної валідації. Порівняно з простим ранжуванням за значенням метрики Confident Learning є більш обґрунтованим, оскільки враховує не лише сам факт розбіжності, а й упевненість моделі в альтернативному прогнозі. Водночас практична ефективність методу суттєво залежить від того, наскільки коректно відкалібровані ймовірності конкретної моделі.

Третій підхід ґрунтується на аналізі неузгодженості моделей ансамблю. У цьому випадку використовують декілька незалежно навчених моделей і оцінюють розбіжність між їхніми прогнозами без звернення до наявної розмітки. Якщо моделі ансамблю відрізняються архітектурою, випадковою ініціалізацією або складом навчальних підвибірок, стійка розбіжність між ними для певного прикладу може вказувати на структурну невизначеність вхідного зображення, а не на окрему помилку однієї моделі. Відмінність цього підходу від попередніх полягає в тому, що оцінка підозрілості розмітки формується виключно на основі прогнозів моделей.

Четвертий підхід передбачає використання попередньо навчених моделей як незалежних оракулів і потребує окремого розгляду через його концептуальну важливість для сучасної медичної сегментації.

2.4.3 Попередньо навчені моделі

Попередньо навчена модель сегментації, чиї параметри сформовані на великому зовнішньому корпусі медичних зображень і потім лише незначно адаптовані до цільового датасету, концептуально несе анатомічні знання, що статистично незалежні від специфічних артефактів розмітки конкретного цільового датасету. Якщо у цільовому датасеті присутній систематичний шум, такий як послідовне розширення межі пазухи на одному й тому ж типі випадків, то навчена з нуля модель з високою ймовірністю відтворить цей шум, оскільки інтерпретуватиме його як

корисний сигнал. Попередньо навчена модель, чий «попередній досвід» формувався на даних з принципово інакшим розподілом помилок, не матиме підстав для відтворення цього специфічного зміщення і дасть прогноз, що систематично відхилятиметься від зашумленої розмітки саме в зашумлених кейсах.

Сучасні попередньо навчені моделі для медичної сегментації КТ становлять окрему підгалузь. Наведена нижче таблиця 2.3 узагальнює найбільш релевантні моделі цієї категорії.

Таблиця 2.3 — Сучасні попередньо навчені моделі для медичної сегментації КТ

Модель	Рік	Корпус попереднього навчання	Архітектурна основа	Ключова особливість
TotalSegmentator	2023	1228 КТ із 104 розміченими структурами	nnU-Net	Універсальна повнотіла сегментація
SuPreM	2024	22 КТ-органи із самопідсиленим попереднім навчанням	SegResNet	Попередньо навчена основа для перенесення знань
MedSAM	2024	1,5 млн пар «зображення–маска» з різних модальностей	ViT-кодувальник + SAM-декодер	Сегментація за підказкою
SAM-Med3D	2024	21 тис. тривимірних медичних об’ємів	3D ViT + SAM-декодер	Адаптація SAM до тривимірних медичних даних

Перевагою використання таких моделей як квалітативних оракулів є те, що відповідний детектор шуму можна реалізувати з мінімальними обчислювальними витратами: достатньо однієї моделі замість ансамблю, причому її лише донавчають на одному фолді, а не тренують з нуля з повною перехресною валідацією. Емпіричні дослідження показують, що такий «дешевий» детектор досягає якості виявлення шуму, близької до повного ансамблю з кількох незалежно навчених моделей, при істотно менших ресурсних витратах.

Розглянуті у цьому підрозділі положення, зокрема типологія дефектів розмітки, поділ виміряної похибки на складову моделі та складову шуму,

підходи до виявлення зашумлених прикладів і поняття кривої досяжної межі якості, формують теоретичне підґрунтя для методології практичної частини дослідження. На цій основі обґрунтовується використання ансамблю моделей різного походження для незалежного оцінювання якості розмітки, побудова композитного показника аномальності, перевірка детектора шуму на вручну переглянутій контрольній вибірці та оцінювання моделей не лише на повному, а й на очищеному тестовому наборі. Практична реалізація цих принципів і результати їх експериментальної перевірки на задачі сегментації клиноподібної пазухи подаються у наступних розділах роботи.

2.5 Вдосконалення сценарію сегментації зображень КТ.

Виконаний у попередніх підрозділах аналіз сучасного стану методів сегментації медичних зображень та клінічних особливостей клиноподібної пазухи дозволяє чітко окреслити напрям подальшого вдосконалення процесу обробки КТ-даних. Як було показано вище, базовою архітектурною основою для практичної частини доцільно обрати тривимірні згорткові моделі повної просторової роздільної здатності, а додатковим фактором, який потребує окремого опрацювання у складі попередньої обробки даних, є неоднорідна якість наявної експертної розмітки. У наступному розділі викладено підхід до автоматичного виявлення потенційно дефектних випадків розмітки, який реалізується через ансамбль тривимірних CNN-моделей сегментації та зіставлення їхніх прогнозів із незалежним експертним переглядом контрольної вибірки.

У другому розділі проаналізовано класичні, двовимірні, псевдотривимірні та тривимірні підходи до сегментації медичних зображень. Встановлено, що класичні методи залишаються корисними як допоміжні інструменти попередньої та заключної обробки. Обґрунтовано вибір тривимірних моделей і фреймворку nnU-Net як найбільш придатної основи для практичної частини роботи.

3 УДОСКОНАЛЕННЯ ПРОЦЕСУ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ КТ

У межах дослідження виявлено, що для покращення результатів сегментації зображень КТ запропоновано розширити стандартний процес попередньої обробки даних окремою процедурою автоматичного виявлення потенційно дефектних випадків експертної розмітки. Її реалізація передбачає виконання таких етапів: визначення складу використаного датасету, ручну перевірку контрольної вибірки та систематизацію виявлених дефектів розмітки, побудову протоколу валідації, обґрунтування вибору моделей ансамблю, визначення набору метрик для оцінювання окремих випадків, формування методів обчислення оцінки підозрілості розмітки, калібрування порогу та перенесення детектора з контрольної вибірки на навчальний пул за допомогою cross-fold інференсу.

3.1 Програмна реалізація та використані бібліотеки

Програмну частину роботи реалізовано як відтворюваний дослідницький хід роботи на Python. Набір скриптів і експериментальних модулів охоплює підготовку КТ-даних і масок, навчання та inference сегментаційних моделей, розрахунок метрик якості, аналіз помилок розмітки, а також формування таблиць, звітів і візуалізацій (табл. 3.1). Реалізація орієнтована на керовані експерименти дипломної роботи, а не на окремий користувацький застосунок.

КТ-об'єми та бінарні маски оброблялися у форматах NIfTI/DICOM із збереженням просторової геометрії. Для запусків фіксувалися сталі розбиття вибірки, manifest-файли та проміжні результати у CSV/JSON. Такий підхід забезпечує повторюваність обчислень і дає змогу відтворити окремі етапи без ручного перенесення параметрів між скриптами.

Навчання й inference виконувалися двома основними гілками. Базова модель M1 запускалася через nnU-Net v2 як сильний baseline для 3D-

сегментації. Моделі M2/M3 реалізовано у PyTorch/MONAI на базі SegResNet, з використанням попередньо навчених ваг SuPreM або випадкової ініціалізації. Для тривимірних об'ємів застосовано patch-based навчання та sliding-window inference, що дає змогу працювати з великими КТ без завантаження повного об'єму в пам'ять GPU.

Оцінювання винесено в окремий CLI-етап. Для кожного випадку обчислювалися Dice, IoU, HD95, ASSD, NSD@1mm/2mm, boundary IoU та Volume Similarity, після чого per-case результати зберігалися у CSV, а агреговані підсумки у JSON. Додаткові скрипти будували статистику, кореляційний аналіз метрик і діагностичні графіки для інтерпретації якості сегментації.

Окремий блок реалізації пов'язаний із пошуком потенційних аномалій розмітки. Anomaly score формувався за кількома каналами, серед яких metric residual, cross-model disagreement, pretrained confidence та ensemble consensus. Після ранжування результати зіставлялися з ручним review і використовувалися як допоміжний сигнал контролю якості, без заміни експертного рішення.

Таблиця 3.1 – Характеристика інструментів

Бібліотека / інструмент	Роль у реалізації
Python 3.10+	Основна мова реалізації скриптів підготовки даних, навчання, оцінювання та звітності
PyTorch, TorchVision	Навчання та inference нейронних мереж, робота з CUDA
MONAI	Медичний deep learning pipeline, SegResNet, sliding-window inference, DiceCE loss
nnU-Net v2	Сильний baseline для 3D-сегментації
NumPy, SciPy	Масиви, морфологічні операції, distance transform і статистичні тести
NiBabel, SimpleITK, pydicom	Робота з медичними форматами NIfTI/DICOM та просторовою геометрією
Matplotlib, PIL	Візуалізації, діагностичні звіти

Лістинг 3.1. Фрагмент реалізації розрахунку метрик сегментації

```
def dice(pred: np.ndarray, target: np.ndarray, eps: float = 1e-8) -> float:
    intersection = float(np.logical_and(pred, target).sum())
    return (2.0 * intersection + eps) / (float(pred.sum()) + float(target.sum()) + eps)
def surface(mask: np.ndarray) -> np.ndarray:
    if not np.any(mask):
        return np.zeros_like(mask, dtype=bool)
    eroded = binary_erosion(mask, structure=np.ones((3, 3, 3), dtype=bool), border_value=0)
    return mask & ~eroded
def evaluate_case(pred_path: Path, target_path: Path, threshold: float) -> dict[str, Any]:
    pred, pred_spacing = load_binary(pred_path, threshold)
    target, target_spacing = load_binary(target_path, 0.5)
    if pred.shape != target.shape:
        raise ValueError(f"Shape mismatch: pred={pred.shape}, target={target.shape}")
    pred_surf = surface(pred)
    target_surf = surface(target)
    spacing = target_spacing
    sd_p2t = distance_transform_edt(~target_surf, sampling=spacing)[pred_surf]
    sd_t2p = distance_transform_edt(~pred_surf, sampling=spacing)[target_surf]
    hd95_val = float(max(np.percentile(sd_p2t, 95), np.percentile(sd_t2p, 95)))
    assd_val = float((sd_p2t.mean() + sd_t2p.mean()) / 2.0)
    return {
        "dice": dice(pred, target),
        "hd95_mm": hd95_val,
        "assd_mm": assd_val,
        "nsd_2mm": _nsd_from_surfaces(pred_surf, target_surf, spacing, 2.0),
        "volume_similarity": volume_similarity(pred, target),
    }
```

Лістинг 3.2. Фрагмент inference для 3D-об'єму в MONAI/SegResNet

```
def predict_case(model: nn.Module, case: CaseRecord, device: torch.device) -> tuple[np.ndarray,
VolumeData]:
    volume: VolumeData = load_volume(case)
    image_tensor: torch.Tensor = torch.from_numpy(volume.image[None, None, ...]).to(device)
    with torch.no_grad():
        with autocast(device_type=device.type, enabled=bool(AMP_ENABLED and device.type == "cuda")):
            logits: torch.Tensor = sliding_window_inference(
                image_tensor,
                roi_size=PATCH_SIZE,
                sw_batch_size=SW_BATCH_SIZE,
                predictor=model,
                overlap=INFER_OVERLAP,
                mode="gaussian",
            )
            probability: torch.Tensor = torch.softmax(logits, dim=1)[0, 1]
    return probability.float().cpu().numpy(), volume
```

Лістинг 3.3. Фрагмент рангового об'єднання каналів anomaly score

```
methods = {
    "score_A_residual": scores_a,
    "score_B_disagreement": scores_b,
    "score_C_confidence": scores_c,
    "score_D_consensus": scores_d,
}
raw_scores: dict[str, list[float]] = {}
for method_name, method_scores in methods.items():
    raw_scores[method_name] = [
        method_scores.get(cid, {}).get(method_name, 0.0) for cid in case_ids
    ]
ranks: dict[str, np.ndarray] = {}
for method_name, values in raw_scores.items():
    arr = np.array(values, dtype=float)
    ranks[method_name] = stats.rankdata(arr, method="average") / len(arr)
for i, cid in enumerate(case_ids):
    rank_values = [ranks[m][i] for m in methods]
    composite_scores[cid] = {
        "rank_A": float(ranks["score_A_residual"][i]),
        "rank_B": float(ranks["score_B_disagreement"][i]),
        "rank_C": float(ranks["score_C_confidence"][i]),
        "rank_D": float(ranks["score_D_consensus"][i]),
        "score_E_composite": float(np.mean(rank_values)),
    }
```

Обрана структура реалізації розділяє підготовку даних, навчання, inference, оцінювання та аналіз розмітки на відтворювані етапи. Завдяки цьому результати можна перевіряти повторно, а окремі компоненти pipeline оновлювати без зміни всієї експериментальної схеми.

3.2 Дані КТ та їх попередня обробка

Експериментальна частина роботи виконана на клінічному датасеті КТ-зображень голови, що містить 534 дослідження з еталонною розміткою клиноподібної пазухи. Розподіл даних на функціональні підмножини виконано до початку експериментів і протягом усіх етапів роботи залишався незмінним з метою унеможливлення витоку інформації між тренувальною та оцінювальною частинами. Систематизований склад датасету з зазначенням ролі кожної підмножини у дослідженні наведено у таблиці 3.2.

Таблиця 3.2 – Склад датасету та схема використання підмножин

Підмножина	Кількість досліджень	Призначення
Тренувальний пул	430	5-фолдна перехресна валідація, навчання моделей ансамблю, OOF-інференс
Внутрішня контрольна вибірка	52	Експертний перегляд масок, вялідація детектора аномалій
Зовнішня контрольна вибірка	52	Експертний перегляд масок, перевірка детектора на даних з іншої клініки

Тренувальна частина датасету використовувався для побудови та валідації моделей сегментації за схемою 5-фолдної перехресної валідації, тоді як обидві контрольні підвибірки виступали у ролі холдауту для незалежної оцінки якості як самих моделей, так і запропонованого детектора аномалій. Зовнішня контрольна підвибірка отримана з клінічної установи, що відрізняється від джерела тренувального пулу як протоколом сканування, так і складом пацієнтів, що дає змогу частково оцінити стійкість детектора до зсуву предметної області. Об'єднання обох контрольних підвибірок у єдиний holdout зі 104 досліджень виконане лише на етапі

експертного перегляду та аналізу аномалій, не передбачаючи їх використання у процедурі навчання моделей.

Поділ тренувального пулу на 5 фолдів виконано із збереженням приблизно однакового розподілу за обсягом цільової структури і типом пневматизації клиноподібної пазухи між фолдами, з метою зменшення варіативності результатів між окремими фолдами під час перехресної валідації.

Необхідною умовою валідації детектора аномалій розмітки є наявність зовнішньої експертної оцінки, причинно незалежної від процедури формування скорингу. У роботі цю роль виконує ручний експертний перегляд кожної маски з контрольної вибірки в обсязі 104 кейсів. Особливістю запропонованого протоколу виступає те, що судження про якість розмітки формується людиною, яка має доступ до повного 3D-об'єму КТ-дослідження та оригінальної маски, але не має доступу до прогнозів жодної з моделей ансамблю на цьому ж дослідженні. Така схема забезпечує статистичну незалежність експертної оцінки від сигналу детектора, що становить необхідну передумову для коректного обчислення метрик якості ранжування.

Кожній масці контрольної вибірки під час перегляду присвоюється один з трьох статусів за ступенем серйозності виявлених дефектів. Розподіл досліджень за цими статусами наведено у таблиці 3.3.

Таблиця 3.3 – Розподіл кейсів контрольної вибірки за статусом експертного перегляду

Статус	Опис	Внутрішня вибірка, (n=52)	Зовнішня вибірка, (n=52)	Загалом
ok	Розмітка узгоджується з анатомією, явних дефектів немає	36	28	64
doubt	Існують підстави для сумніву, але дефект не є очевидним	12	12	24
exclude	Виявлено явний дефект, зокрема зсув маски, пропуск частини структури або розмітку іншої анатомічної області	4	12	16

Розподіл показує, що зовнішня контрольна підвибірка містить істотно більше випадків з явними дефектами розмітки, ніж внутрішня (12 проти 4), що узгоджується з очікуваним вищим рівнем шуму у даних, отриманих з іншої клінічної установи. Загалом частка випадків зі сумнівною або явно дефектною розміткою у контрольній вибірці становить 38 %, що відповідає літературним оцінкам поширеності шуму розмітки у медичних датасетах і слугує емпіричною мотивацією для подальшої роботи з детектором аномалій.

Окремою методологічною деталлю є обґрунтування того, які саме статуси використовуються як цільова змінна під час калібрування детектора. У роботі прийнято рішення використовувати тільки статус exclude як hard ground truth, а статус doubt виключати з калібрувальної цілі. Підставою для такого рішення виступає відмінність характеру невизначеності між двома проміжними класами. Статус exclude відповідає випадкам, у яких дефект розмітки може бути зафіксований об'єктивно: наявність зсуву маски, пропуск однієї з симетричних структур, явне розмічення неправильної анатомічної області. Натомість статус doubt відображає суб'єктивну невпевненість експерта в окремих граничних випадках, які можуть бути коректно розміченими, але мають незвичну конфігурацію. Включення doubt у калібрувальну ціль розмиває оптимальний поріг детектора і зменшує його якість як ранжуючого механізму. Емпіричне підтвердження цього рішення наводиться у підрозділі 4.1, де показано, що значення ROC-AUC всіх п'яти методів скорингу систематично зростає при переході від цілі exclude+doubt до цілі exclude only. У всіх подальших обчисленнях під виразом «дефектні випадки» розуміються виключно кейси зі статусом exclude, що складають 16 досліджень з 104.

Вагомим обмеженням запропонованої схеми виступає те, що експертний перегляд виконано автором роботи, а не радіологом-фахівцем. Цей фактор частково компенсується трьома методологічними рішеннями.

По-перше, для калібрування використовуються лише явно дефектні випадки, для яких міжоператорна узгодженість зазвичай висока і не залежить істотно від кваліфікації окремого спостерігача.

По-друге, остаточна оцінка валідності детектора у Розділі 4 здійснюється комбінацією критеріїв, серед яких узгодженість з експертною класифікацією, статистична значущість проти випадкового базису та якісний аналіз окремих дефектних кейсів, що зменшує залежність висновків від суб'єктивних особливостей одного оглядача.

По-третє, інженерний характер виявлених дефектів (зсунуті маски, пропуски структури) дозволяє верифікацію через зовнішнє програмне порівняння обсягів та просторових координат масок з типовим розподілом по датасету.

3.2.1 Типологія виявлених дефектів розмітки

Аналіз 16 кейсів зі статусом exclude та порівняння з прогнозами моделей ансамблю на цих самих кейсах дозволили виокремити п'ять основних категорій дефектів розмітки, що зустрічаються у досліджуваному датасеті. Знання цієї типології суттєве для подальшого аналізу: різні категорії дефектів виявляються різними методами скорингу і мають різну вагу для подальших застосувань детектора. Систематизована типологія наведена у таблиці 3.4, а характерні приклади виявлених дефектів представлено на рисунку 3.1.

Таблиця 3.4 – Типологія дефектів розмітки клиноподібної пазухи у контрольній вибірці

Категорія дефекту	Частка серед exclude	Сигнатура у прогнозах моделей
Зсунута маска	≈ 50 %	Dice близький до 0 у всіх моделей ансамблю одночасно
Часткова розмітка, одна сторона	≈ 20 %	Dice 0,3–0,5, високий рівень FN, пропуск симетричної частини
Розширена розмітка, включення запалення	≈ 15 %	Dice 0,7–0,85, високий рівень FP у м’якотканинній зоні

Дублікати зрізів	≈ 10 %	Просторова невідповідність маски та КТ-об’єму
Розмітка іншої структури	≈ 5 %	Dice близький до 0 у всіх моделей ансамблю одночасно

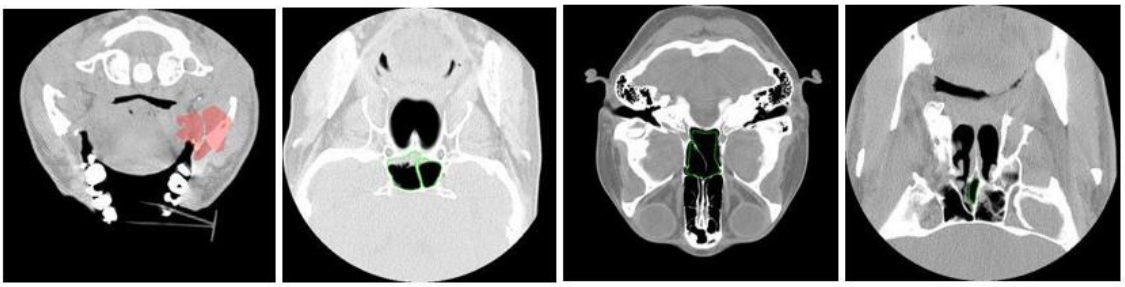

Рисунок 3.1 – Дефекти розмітки клиноподібної пазухи у контрольній вибірці

Найпоширенішим типом дефекту виступає зсунута маска, для якої характерним є одночасний обвал якості сегментації по всіх моделях ансамблю незалежно від їх архітектури. Така сигнатура є оптимальною для виявлення детектором: розбіжність між узгодженим прогнозом ансамблю і наявною розміткою формує сигнал, що погано пояснюється варіативністю окремих моделей.

Часткова розмітка, у якій одна з симетричних половин клиноподібної пазухи систематично пропущена, дає менш виражений, але стабільний сигнал у формі високого рівня хибнонегативних вокселів у регіоні пропуску. У роботі такі випадки залишаються коректно ранжованими композитним детектором, але потребують більшого пасу композитної оцінки для впевненої ідентифікації порівняно з повністю зсунутими масками.

Розширена розмітка, у якій маска включає не лише повітряну порожнину, а й прилеглу запалену слизову оболонку, виступає найбільш складним для детектора типом дефекту. Прогноз моделі з вагою SuPreM, орієнтованої на сегментацію за повітряною щільністю, систематично відхиляється від такої розмітки в напрямку звуження, що призводить до

високого FP-сигналу. Однак у цьому ж типі випадків розмітка може бути коректною з точки зору інших клінічних задач, тому видалення таких кейсів з тренувального пулу не завжди дає очікуваний приріст якості.

Дублікати зрізів та помилкова розмітка іншої анатомічної структури зустрічаються рідше і виявляються переважно через просторові інваріанти, не пов'язані безпосередньо з якістю прогнозів моделей. У межах роботи такі випадки виявлялися додатковою перевіркою цілісності файлів і просторових атрибутів NIfTI-об'єктів і маркувалися окремо.

Загалом запропонована типологія охоплює близько 95 % виявлених дефектів і використовується у Розділі 4 як основа для якісного аналізу окремих кейсів і для змістовної інтерпретації механізму, через який детектор формує сигнал.

3.2.2 Попередня обробка даних та поділ на частини

Усі КТ-дослідження пройшли єдину процедуру попередньої обробки, узгоджену з логікою роботи nnU-Net. Вона передбачала приведення зображень до уніфікованого просторового кроку, автоматично визначеного за характеристиками тренувального пулу, нормалізацію інтенсивностей у шкалі Гаунсфілда з обрізанням значень до 5-го та 95-го перцентилів тренувального розподілу, а також поділ кожного об’єму на тривимірні фрагменти для подальшої обробки моделями ансамблю. Усі параметри попередньої обробки були однаковими для всіх моделей, що забезпечувало зіставність отриманих результатів.

Тренувальний пул було детерміновано поділено на п’ять непересічних підмножин для проведення перехресної валідації. Той самий поділ застосовано і в експериментах із моделями родини SegResNet, які навчалися у середовищі MONAI, щоб зберегти однакові умови порівняння між архітектурами. Рішення щодо штучного розширення даних і схеми навчання моделей розглянуто у підрозділі 3.2.

Контрольна вибірка, що містила 104 дослідження, не використовувалася під час формування навчальних підмножин і не залучалася до процедур штучного розширення даних, що забезпечило її незалежність від навчального процесу та дозволило використовувати її для об’єктивної перевірки детектора аномалій розмітки.

3.3 Архітектура детектора аномалій розмітки

Побудова детектора аномалій розмітки спирається на три компоненти: незалежну експертну оцінку контрольної вибірки, ансамбль моделей сегментації та процедуру перетворення їхніх прогнозів у композитну оцінку підозрілості розмітки кожного випадку. У межах розділу визначено протокол перевірки детектора, склад ансамблю, порядок формування прогнозів, набір метрик для кожного окремого випадку, способи оцінювання підозрілості розмітки, процедуру калібрування порогу та механізм перенесення детектора на тренувальну частину датасету.

3.3.1 Замкнений цикл валідації з незалежною експертною оцінкою

Перевірка детектора аномалій розмітки не може ґрунтуватися на тій самій розмітці, якість якої ставиться під сумнів. Тому як еталон для його оцінювання використано незалежний експертний перегляд контрольної вибірки. Експертна оцінка не залучалася ні до навчання моделей ансамблю, ні до побудови показника аномальності, що забезпечує її незалежність від самого детектора.

Загальну логіку протоколу подано на рисунку 3.2. Вона охоплює три рівні: незалежну експертну оцінку, детектор аномалій розмітки та три напрями його подальшого використання.

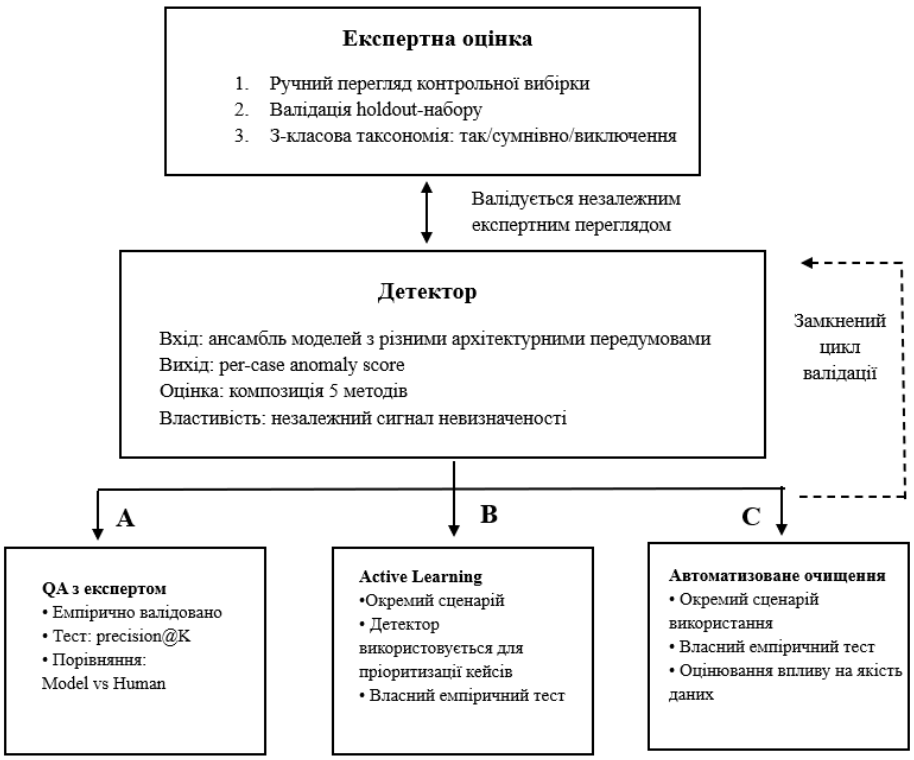

Рисунок 3.2 – Замкнений цикл валідації детектора аномалій розмітки з незалежною експертною оцінкою

На вхід детектора надходять КТ-зображення Xта прогнози трьох моделей сегментації $\{f_1(X), f_2(X), f_3(X)\}$, побудованих на різних архітектурних передумовах. На виході формується скалярна оцінка підозрілості розмітки, яка відображає відносний ризик того, що наявна маска містить дефект розмітки. Якість ранжування оцінюють за площею під ROC-кривою та точністю серед перших Kвипадків, використовуючи незалежний експертний висновок як еталон.

Після перевірки детектор застосовується у трьох різних сценаріях: для контролю якості наявної розмітки за участі експерта, для відбору неанотованих зображень у задачі активного навчання та для автоматичного очищення тренувальної вибірки. Придатність до ранжування випадків за аномальністю і придатність до конкретного практичного застосування розглядаються окремо. Наявність якісного сигналу є необхідною умовою

для подальшого використання детектора, але сама по собі не гарантує однакової ефективності в усіх сценаріях.

Модель сегментації не може виступати незалежним оцінювачем розмітки, на якій вона навчалася. У тренувальному пулі розмітка є ціллю навчання, тому модель частково відтворює її властивості разом із можливими дефектами. На тестовій вибірці прогноз лише порівнюється з наявною маскою, але не визначає її коректність. Через це еталонна оцінка детектора має формуватися окремо, за участі людини, на даних, що не брали участі у внутрішніх циклах навчання моделей.

3.3.2 Ансамбль моделей сегментації

Сигнал детектора формується не за прогнозом однієї моделі, а за узгодженістю та розбіжностями між кількома моделями. Інформативність такого підходу залежить від складу ансамблю. Надто подібні моделі дають майже однакові прогнози й додають мало нової інформації, тоді як надмірно різні архітектури можуть створювати нестабільність, не пов’язану з якістю розмітки. Тому ансамбль сформовано з трьох 3D-моделей повної просторової роздільної здатності, які мають близьку якість сегментації, але відрізняються архітектурою або джерелом початкових ваг.

Перш ніж конкретизувати склад ансамблю, доцільно коротко охарактеризувати ті архітектурні класи, з яких обрано моделі, та зафіксувати їх ключові структурні відмінності.

PlainConvUNet 3D у складі nnU-Net. Базовою архітектурою моделі M1 є тривимірний варіант U-Net із згортковими блоками без залишкових з'єднань як 2D-прототип [36] і узагальнений на тривимірний випадок [37]. У межах фреймворку nnU-Net [6] глибина мережі, кількість каналів, розмір патчу та параметри навчання автоматично адаптуються до характеристик конкретного датасету, що знімає з дослідника необхідність ручного підбору гіперпараметрів. Архітектурною особливістю PlainConvUNet є послідовне

застосування простих згорткових блоків Conv-IN-LReLU у кодувальнику та декодувальнику, що забезпечує стабільну збіжність на медичних даних обмеженого обсягу.

SegResNet. Моделі M2 та M3 побудовані на архітектурі SegResNet [38] і реалізованій у фреймворку MONAI [39]. SegResNet є залишковим варіантом тривимірної U-подібної мережі: кодувальник складається з послідовності залишкових блоків зростаючої глибини, між якими виконується просторовий downsampling, а декодувальник реалізує симетричне відновлення просторової роздільної здатності з skip-з'єднаннями. Наявність residual-блоків полегшує градієнтну поширюваність на глибших рівнях і робить архітектуру придатною для попереднього навчання на великих корпусах.

SuPreM як джерело попередньо навчених ваг. Початкова ініціалізація моделі M2 виконана за допомогою ваг SuPreM[40]. SuPreM є набором попередньо навчених на великому корпусі КТ-зображень тривимірних моделей сегментації, включно з варіантами SegResNet, призначеним для подальшого донавчання на цільових клінічних задачах. Використання SuPreM як ініціалізації для M2 дає змогу частково перенести у локальну модель представлення, сформовані поза наявним зашумленим датасетом, і тим самим забезпечити архітектурну диверсифікацію ансамблю не лише за рахунок різниці мереж, але й за рахунок різних джерел вагової інформації.

Узагальнено три моделі покривають дві осі диверсифікації: різну архітектуру і різне джерело початкових ваг. Така комбінація обрана не для отримання найкращої окремої моделі, а для забезпечення часткової незалежності помилок між елементами ансамблю, що є необхідною передумовою для подальшого виявлення дефектів розмітки за узгодженістю/розбіжностями прогнозів.

Таблиця 3.5 – Склад ансамблю моделей сегментації

Ідентифікатор	Архітектура	Початкова ініціалізація	Кількість епох	Dice на валідаційній підмножині 0
M1	nnU-Net v2, PlainConvUNet 3D, режим 3d_fullres	випадкова	1000	0,929
M2	MONAI SegResNet, залишкова 3D U-Net	SuPreM	145	0,897
M3	MONAI SegResNet, залишкова 3D U-Net	випадкова	145	0,893

Модель M1 представляє автоматично налаштований фреймворк nnU-Net і виконує роль найсильнішої базової моделі ансамблю. Моделі M2 та M3 побудовані на однаковій архітектурі SegResNet, але відрізняються способом ініціалізації. M2 використовує ваги SuPreM, отримані внаслідок попереднього навчання на зовнішньому корпусі медичних КТ-даних, тоді як M3 навчається з випадкової ініціалізації. Така побудова дозволяє окремо оцінити внесок архітектури та внесок попереднього навчання.

Поєднання M1, M2 і M3 дає три різні джерела судження про форму цільової структури. M1 та M3 повністю формують свої представлення на досліджуваному датасеті, тоді як M2 починає навчання з ваг, отриманих на зовнішніх даних. За рахунок цього M2 частково зберігає незалежність від можливих систематичних дефектів локальної розмітки і може бути корисною для виявлення випадків, де моделі, навчені з нуля, відтворюють помилку тренувального набору.

Усі три моделі навчалися на одному поділі даних, що містив 343 дослідження у тренувальній частині та 87 у валідаційній. Для M2 і M3 використовувалися однакові гіперпараметри, зокрема швидкість навчання, схема її зменшення, розмір пакета та набір перетворень даних. Відмінність між ними полягала лише у початкових вагах. Модель M1 навчалася за автоматично сформованою конфігурацією nnU-Net без ручного коригування параметрів. Усі три моделі досягли стабільного рівня якості, достатнього для подальшого використання в ансамблі.

3.3.3 Формування прогнозів і набору метрик для окремих випадків

Після навчання кожна модель сформувала прогнози для всіх 104 досліджень контрольної вибірки. Для кожного випадку зберігалися дві форми виходу: бінарна сегментаційна маска та ймовірнісна карта класів. Бінарні маски використовувалися для обчислення стандартних метрик сегментації, а ймовірнісні карти, для оцінювання впевненості моделі у зонах розбіжності з розміткою.

Для кожного поєднання «випадок, модель» обчислювалися об’ємні та поверхневі метрики. До першої групи увійшли Dice, IoU, precision, recall, подібність об’ємів і відносна похибка об’єму. До другої, HD95, ASSD, NSD з порогами 1 і 2 мм, boundary IoU та нормалізовані варіанти HD95 і ASSD. Сукупно було отримано $104 \times 3 \times 16$значень, які надалі використовувалися для побудови всіх показників аномальності без повторного виконання інференсу.

Перед обчисленням поверхневих метрик із прогнозів видалялися малі компоненти зв’язності обсягом менше 64 вокселів. Така обробка зменшувала вплив випадкових шумових включень поза цільовою структурою і не змінювала сегментацію клиноподібної пазухи, об’єм якої істотно перевищував зазначений поріг.

3.3.4 Методи оцінювання підозрілості розмітки розмітки

Для виявлення потенційно некоректних масок використано чотири базові підходи та один композитний показник. Методи адаптовані у межах роботи на основі загальних принципів виявлення помилок розмітки та оцінювання модельної невизначеності, систематизованих у літературі з confident learning [17], аналізу зашумленої розмітки [16] та ансамблевих методів оцінювання невизначеності у глибинному навчанні [41; 42]. Кожен базовий метод фіксує окремий тип сигналу, пов'язаний із можливим шумом

розмітки. Об'єднання результатів виконується через рангове злиття, що дозволяє поєднати показники різного масштабу без прямого підбору ваг.Метод А. Метричне відхилення на рівні окремого випадку

Перший метод оцінює, наскільки якість сегментації конкретного випадку відхиляється від типової якості моделі на контрольній вибірці. Для моделі mі випадку iобчислюється стандартизоване відхилення Dice:

$$z_m(i) = \frac{\mu_m - Dice_m(i)}{\sigma_m}, \qquad (3.1)$$

де μ_mі σ_mє середнім значенням та стандартним відхиленням Dice для моделі mна контрольній вибірці.

Аналогічні значення розраховуються для HD95 та ASSD, але з урахуванням того, що більші значення цих метрик відповідають гіршій якості. Підсумковим показником методу є найбільше стандартизоване відхилення серед усіх метрик і моделей. Перевага підходу полягає у простоті, однак він не розрізняє дефектну розмітку та анатомічно складний, але коректно розмічений випадок.

Метод В. Розбіжність усередині ансамблю

Другий метод порівнює узгодженість прогнозів моделей між собою з їхньою узгодженістю з наявною розміткою:

$$score_B(i) = \frac{1}{3}\sum_{j<k} Dic\, e_{j,k}(i) - \frac{1}{3}\sum_{m} Dic\, e_{m,Y}(i), \qquad (3.2)$$

де $Dice_{j,k}(i)$є значенням Dice між прогнозами моделей jі k, а $Dice_{m,Y}(i)$є Dice між прогнозом моделі mта наявною розміткою Y.

Високе значення показника виникає тоді, коли моделі ансамблю узгоджені між собою, але одночасно відхиляються від розмітки. Така ситуація є характерною ознакою потенційного дефекту маски. Метод має

найменшу залежність від наявної розмітки серед усіх базових підходів, однак його застосування потребує кількох незалежно навчених моделей.

Метод C. Упевненість попередньо навченої моделі у зонах розбіжності

Третій метод використовує ймовірнісну карту моделі M2, яка має зовнішню попередню ініціалізацію. Показник збільшується тоді, коли модель із високою впевненістю прогнозує клас, відмінний від наявної розмітки:

$$score_C(i) = \frac{1}{2}(conf_{FP}(i) + conf_{FN}(i)) \cdot (1 - H(p_{dis}(i))) \cdot \frac{|D_i|}{|Y_i \cup \hat{Y}_i|} \qquad (3.3)$$

де $conf_{FP}(i)$і $conf_{FN}(i)$є середньою впевненістю моделі у зонах хибнопозитивних і хибнонегативних розбіжностей, $H(p_{dis}(i))$є нормалізованою ентропією у зонах розбіжності, а $|\ D_i\ |/|\ Y_i \cup \hat{Y}_i\ |$визначає відносний обсяг цих зон.

Перший множник відображає силу незгоди, другий, упевненість прогнозу, третій, просторовий масштаб розбіжності. Метод потребує лише однієї моделі та тому придатний для спрощеного варіанта детектора з нижчими обчислювальними витратами.

Метод D. Узгодженість мажоритарного прогнозу з розміткою

Четвертий метод порівнює наявну розмітку з агрегованим прогнозом ансамблю, сформованим мажоритарним голосуванням на рівні вокселів:

$$\hat{Y}_a(i) = sign\left(\sum_m \hat{Y}_m\,(i) - 1{,}5\right), \qquad (3.4)$$

де $\hat{Y}_m(i)$є бінарним прогнозом моделі m.

Показник методу визначається як поєднання оберненого Dice мажоритарного прогнозу та частки вокселів, які ансамбль одностайно відносить до цільової структури, але які відсутні в розмітці:

$$score_D(i) = (1 - Dice_a(i)) + \gamma \cdot \frac{| \hat{Y}_a \setminus Y_i |}{| \hat{Y}_a \cup Y_i |}, \quad (3.5)$$

де γє ваговим коефіцієнтом, прийнятим рівним 0,5.

Підхід є інтерпретованим, оскільки безпосередньо оцінює ступінь розходження між розміткою та колективним прогнозом ансамблю. Разом із тим він частково дублює сигнал методу С, оскільки обидва підходи спираються на незгоду між прогнозом і наявною маскою.

Метод Е. Композитний показник на основі рангового злиття

П’ятий метод об’єднує чотири попередні показники в єдине ранжування. Для кожного методу випадки впорядковуються за спаданням аномальності, після чого обчислюється середня нормована позиція:

$$score_E(i) = \frac{1}{4N}(rank_A(i) + rank_B(i) + rank_C(i) + rank_D(i)), \quad (3.6)$$

де $rank_X(i)$є позицією випадку iу ранжуванні за методом X, а Nє кількістю досліджень у контрольній вибірці.

Рангове злиття не залежить від абсолютного масштабу базових показників і є стійкішим до окремих викидів, ніж пряме усереднення значень. Саме композитний показник використовується як основний для подальшого калібрування порогу.

Таблиця 3.6 – Зведена характеристика методів оцінювання підозрілості розмітки розмітки

Метод	Джерело сигналу	Залежність від розмітки	Кількість моделей	Перевага	Обмеження
A. Метричне відхилення	стандартизовані метрики для випадку	так	1+	низька обчислювальна вартість	не відрізняє шум від складного випадку
B. Розбіжність ансамблю	узгодженість прогнозів моделей	слабка	3	майже не залежить від розмітки	потребує ансамблю
C. Упевненість попередньо навченої моделі	ймовірнісна карта M2	так	1	низька вартість, окремий зовнішній сигнал	залежить від якості M2
D. Узгодженість мажоритарного прогнозу	агрегований прогноз ансамблю	так	3	висока інтерпретованість	частково дублює C
E. Композитний показник	рангове злиття A+B+C+D	слабка	3	найстійкіше інтегральне ранжування	менша інтерпретованість окремих внесків

3.3.5 Калібрування порогу прийняття рішень

Базові методи та композитний показник повертають неперервні значення, придатні для ранжування випадків за ймовірністю аномалії. Для переходу до бінарного рішення використано ROC-аналіз із критерієм Юдена:

$$J^* = \max_t (TPR(t) - FPR(t)), \tag{3.7}$$

де $TPR(t)$і $FPR(t)$є частками правильно та хибно виявлених позитивних випадків за порогу t.

Цільовим позитивним класом обрано статус exclude, встановлений за результатами експертного перегляду. Випадки doubt не включалися до цілі

калібрування, оскільки вони не мають однозначного висновку про дефект розмітки.

Для перенесення детектора на тренувальний пул додатково використано відносний поріг у вигляді частки найаномальніших випадків $top\text{-}K\%$. Такий підхід є стабільнішим за перенесення абсолютного порога, оскільки розподіл оцінки підозрілості розмітки в контрольній і тренувальній вибірках може відрізнятися. Калібрування виконувалося один раз на повній контрольній вибірці зі 104 досліджень. Через її обмежений обсяг додатковий поділ для перехресної перевірки порогу не застосовувався, а невизначеність оцінок надалі враховувалася через довірчі інтервали, обчислені методом бутстреп-перевибірки.

3.3.6 Перевірка значущості відносно випадкового відбору

Для оцінювання того, чи перевищує детектор рівень випадкового відбору, застосовано перестановочний тест. Для кожного значення K виконувалося 1000 випадкових перестановок індексів, після чого вилучалися $top\text{-}K$ випадків і повторно обчислювалися цільові метрики. Отриманий розподіл використовувався як нульова гіпотеза, а положення фактичного результату детектора в цьому розподілі визначало p-значення.

Силу ефекту додатково оцінювали за коефіцієнтом Cohen's d:

$$d = \frac{x_{\text{detector}} - \mu_{\text{null}}}{\sigma_{\text{null}}}, \tag{3.8}$$

де x_{detector} є результатом детектора, а μ_{null} і σ_{null} є середнім значенням та стандартним відхиленням нульового розподілу. Перестановочний тест не вимагає припущення про нормальність розподілу метрик і тому є придатним для медичних вибірок невеликого обсягу.

3.3.7 Прогнозування поза фолдом на тренувальному пулі

Для застосування детектора до тренувального пулу необхідно отримати прогнози, сформовані моделями, які не навчалися на відповідних дослідженнях. Для цього використано прогнозування поза фолдом, або OOF-підхід. У межах п’ятифолдної перехресної валідації кожен випадок тренувального пулу прогнозується моделлю, навченою на решті чотирьох підмножин.

Для моделі M2 було сформовано п’ять версій, по одній для кожної підмножини перехресної валідації. Після об’єднання прогнозів усі 430 тренувальних випадків отримали оцінки від моделей, які не використовували їх під час навчання. До цих прогнозів застосовувався той самий набір метрик, що й для контрольної вибірки.

Оскільки метод C потребує лише ймовірнісних карт M2, його можна перенести на тренувальний пул без повторення повного OOF-циклу для всіх трьох моделей ансамблю. Такий варіант значно зменшує обчислювальні витрати й водночас зберігає основний зовнішній сигнал, необхідний для виявлення потенційно дефектних масок. На основі отриманих оцінок формуються списки $top\text{-}K\%$ найаномальніших випадків, які надалі використовуються для побудови очищених навчальних вибірок.

У третьому розділі описано підхід до автоматичного виявлення підозрілих випадків розмітки на основі ансамблю моделей сегментації як складову етапу попередньої обробки даних. Визначено склад датасету, виконано ручний перегляд контрольної вибірки, побудовано ансамбль сегментаційних моделей і запропоновано набір показників для оцінювання підозрілих випадків. Розроблений протокол передбачає незалежну експертну оцінку, калібрування порогу та перенесення детектора на тренувальний пул через прогнозування поза фолдом, що створює основу для подальшої експериментальної перевірки його практичної придатності.

4 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНОГО ДОСЛІДЖЕННЯ

Експериментальна частина роботи побудована навколо чотирьох гіпотез, кожна з яких перевіряється на окремому етапі дослідження.

Гіпотеза H1: композитна оцінка підозрілості розмітки, обчислена детектором аномалій розмітки, статистично значимо узгоджується з незалежною експертною оцінкою якості розмітки і за метрикою якості ранжування достовірно перевершує випадковий рівень.

Гіпотеза H2: при оцінюванні моделі на скороченому тестовому наборі модельна фільтрація випадків за оцінкою підозрілості розмітки дає вищі значення метрик, ніж випадкова фільтрація та ручне виключення тієї самої кількості випадків.

Гіпотеза H3: автоматичне виключення випадків з найвищою оцінкою підозрілості розмітки з навчальної вибірки призводить до покращення якості моделі при подальшому навчанні.

Гіпотеза H4: оцінка підозрілості розмітки придатна як критерій вибору кейсів у режимі активного навчання, тобто пріоритетна розмітка випадків з високою оцінкою ефективніше прискорює сходження моделі, ніж випадкова стратегія розмітки.

Гіпотези H1 і H2 перевіряються на контрольній вибірці, гіпотеза H3 за результатами повторного навчання модельного ансамблю на очищеному тренувальному пулі. Підтвердження або спростування кожної з цих гіпотез разом визначають кінцеву характеристику меж застосовності запропонованої процедури.

4.1 Валідація детектора аномалій розмітки на контрольній вибірці

На першому етапі експериментального дослідження перевірено, чи здатний побудований детектор ранжувати випадки за ймовірністю наявності дефекту розмітки, щ обуло представлено раніше у підрозділі 2.4.

4.1.1 Вибір цільового класу для калібрування

Спочатку перевірено, чи доцільно поєднувати категорії doubt та exclude в один позитивний клас. Такий варіант виявився менш придатним для валідації. Після переходу до цільового класу exclude якість ранжування зросла для всіх п’яти методів. Найбільший приріст AUC отримано для ансамблевого консенсусу, композитного показника та методу попередньо навченої моделі (табл. 4.1, рис. 4.1).

Таблиця 4.1 – Порівняння варіантів цільового класу для калібрування детектора

Метод	AUC (exclude+doubt)	AUC (exclude only)	Приріст AUC
A: Metric Residual	0.645	0.681	+0.036
B: Cross-Model Disagreement	0.600	0.670	+0.070
C: Pretrained Confidence	0.624	0.705	+0.081
D: Ensemble Consensus	0.629	0.717	+0.088
E: Composite	0.640	0.722	+0.082

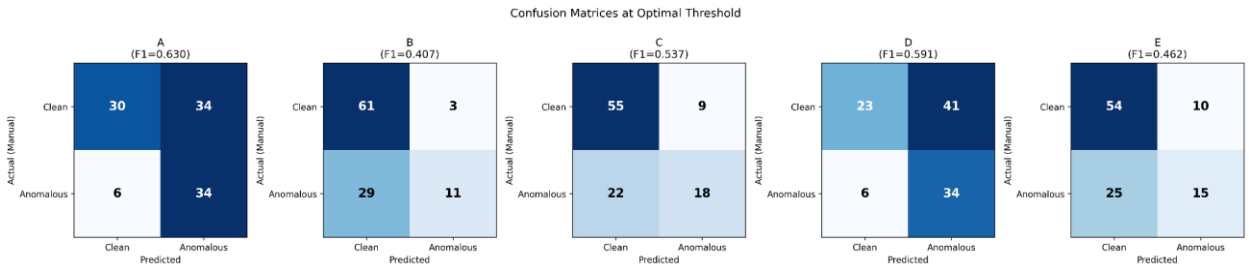

Рисунок 4.1 – Матриця плутанини методів

Отриманий результат показує, що категорія doubt відображає не лише можливі дефекти масок, а й анатомічно складні або протокольно неоднозначні випадки. Її включення до позитивного класу змішує два різні явища, тому для первинної перевірки детектора доцільніше використовувати лише явно дефектні маски. Такий вибір є також більш обережним з огляду на те, що ручний перегляд виконувався одним оглядачем, а не незалежною групою радіологів.

4.1.2 Якість ранжування дефектних випадків

За цільового класу exclude усі п’ять методів показали якість вище випадкового рівня. Найкращий результат отримано для композитного показника, для якого AUC становив 0,722, AP дорівнював 0,439, а кореляція Спірмена з ручною оцінкою була статистично значущою, $\rho = 0{,}272$, $p = 0{,}0053$. Серед окремих методів найближчими до нього виявилися ансамблевий консенсус з AUC 0,717 і впевненість попередньо навченої моделі з AUC 0,705 (табл. 4.2).

Таблиця 4.2 – Показники якості ранжування для методів оцінювання підозрілості розмітки

Метод	AUC	AP	Spearman ρ	p-value Spearman
A: Metric Residual	0.681	0.360	0.264	0.0068
B: Cross-Model Disagreement	0.670	0.430	0.198	0.0442
C: Pretrained Confidence	0.705	0.410	0.243	0.0131
D: Ensemble Consensus	0.717	0.397	0.254	0.0094
E: Composite	0.722	0.439	0.272	0.0053

Позитивний клас складав лише 16 із 104 випадків, тому додатково проаналізовано precision-recall характеристику. Для композитного показника AP дорівнював 0,439 за випадкового рівня 0,154, тобто верхня частина ранжування була майже втричі більш насичена дефектними випадками, ніж випадкова вибірка.

Таблиця 4.3 – Матриці помилок методів у робочій точці за критерієм Юдена

Операційна точка Youden's J	Threshold	TP	FP	FN	TN	Precision	Recall	F1
A: Metric Residual	-0.184	14	48	2	40	0.226	0.875	0.359
B: Cross-Model Disagreement	0.038	8	9	8	79	0.471	0.500	0.485
C: Pretrained Confidence	0.110	10	17	6	71	0.370	0.625	0.465
D: Ensemble Consensus	0.008	12	27	4	61	0.308	0.750	0.436

E: Composite	0.589	12	27	4	61	0.308	0.750	0.436

Робочі точки, визначені за критерієм Юдена (табл. 4.3), показали різні профілі поведінки методів. Метричне відхилення давало найвищу чутливість, але супроводжувалося великою кількістю хибнопозитивних спрацьовувань. Міжмодельна розбіжність була обережнішою, маючи вищу точність за нижчої повноти. Композитний показник і ансамблевий консенсус в обраній точці виявили 12 із 16 явних дефектів, але разом із ними позначили 27 додаткових випадків.

Отримані значення не свідчать про повну заміну експертного перегляду. Вони показують інше: детектор здатний відокремлювати явно дефектні маски від решти вибірки та формувати корисне ранжування для подальшого контролю якості.

4.1.3 Структура сигналу детектора аномалій розмітки

Кореляційний аналіз показав, що чотири базові методи фактично формують два типи сигналу. Першу групу формують методи A, C і D, які відображають розбіжність між прогнозом моделі або ансамблю та наявною розміткою. Між ними спостерігається виражений кореляційний зв’язок: $A - C = 0{,}69$, $A - D = 0{,}74$, $C - D = 0{,}90$. Такий результат є закономірним, оскільки всі три методи оцінюють близьке за змістом явище, але через різні ознаки. Метод A використовує відхилення метрик, метод C враховує впевненість попередньо навченої моделі у зонах незгоди з розміткою, а метод D спирається на розбіжність між мажоритарним прогнозом ансамблю та наявною маскою (табл. 4.4).

Таблиця 4.4 – Кореляції між базовими методами оцінювання підозрілості розмітки

	A: Residual	B: Disagree	C: Pretrained	D: Consensus
A: Residual	1.00	-0.09	0.69	0.74
B: Disagree	-0.09	1.00	0.37	0.35
C: Pretrained	0.69	0.37	1.00	0.90
D: Consensus	0.74	0.35	0.90	1.00

Окремий тип сигналу формує метод B, який характеризує узгодженість прогнозів моделей між собою. Його зв'язок із методом A є майже відсутнім, $\rho = -0{,}09$, а кореляції з методами C і D залишаються помірними, відповідно 0,37 і 0,35. На відміну від інших підходів, метод B відповідає не на питання про відповідність прогнозу наявній розмітці, а на питання про те, чи збігаються висновки кількох моделей незалежно від якості самої маски. Саме тому цей сигнал має особливу цінність у практичному застосуванні, де експертний перегляд для нового випадку ще не виконано, а коректність розмітки наперед невідома.

Практичне значення цього результату полягає також у поясненні високої самостійної ефективності методу C. Показник упевненості попередньо навченої моделі належить до домінантної групи сигналів і досягає $AUC = 0{,}705$проти $0{,}722$у композитного показника. Отже, попередньо навчена та донавчена модель SuPreM забезпечує значну частину діагностичної інформації за менших обчислювальних витрат. Водночас повний композитний показник залишається доцільним для валідації, оскільки поєднує сигнал попередньо навченої моделі з окремою міжмодельною інформацією.

4.1.4 Порівняння модельної та ручної фільтрації

Найбільш прикладним тестом першого етапу стало порівняння ручної та модельної фільтрації за однакової кількості вилучених випадків. Для nnU-Net вилучення 16 випадків за модельним ранжуванням підвищило Dice до 0,947, тоді як ручне вилучення 16 випадків зі статусом exclude дало 0,933. За вилучення 40 випадків різниця також зберігалася: 0,959 для модельної фільтрації проти 0,942 для ручного залишення лише випадків ok (табл. 4.5).

Таблиця 4.5 – Результати ручної та модельної фільтрації для трьох моделей сегментації

Стратегія	N після фільтрації	M1 nnU-Net Dice	M2 SuPreM Dice	M3 Random Dice
No filter	104	0.909	0.890	0.883
Manual v1: без exclude	88	0.933	0.914	0.906
Manual v2: тільки ok	64	0.942	0.916	0.907
Model top-16	88	0.947	0.929	0.919
Model top-40	64	0.959	0.954	0.944
Model optimal	79	0.953	0.941	0.933
Pretrained-only top-16	88	0.947	0.931	0.927

Таблиця 4.6 – Порівняння ручної та модельної фільтрації за однакової кількості вилучених випадків

Порівняння при однаковому N	Manual filter Dice M1	Model filter Dice M1	Перевага model filter
16 вилучено, 88 залишено	0.933	0.947	+1.4 п.п.
40 вилучено, 64 залишено	0.942	0.959	+1.7 п.п.

Перевага модельної фільтрації спостерігалася не лише для Dice. Для nnU-Net за вилучення 16 випадків HD95 зменшився з 5,47 до 3,75 мм, а NSD@2mm зріс з 0,962 до 0,979. За вилучення 40 випадків HD95 знизився з 4,89 до 2,57 мм, а NSD@2mm підвищився з 0,969 до 0,987 (табл.4.6).

Таблиця 4.7 – Зміна Dice, HD95 та NSD@2mm після різних стратегій фільтрації

Стратегія	M1 Dice	M1 HD95, мм	M1 NSD@2mm
No filter	0.909	6.95	0.938
Manual v1: без exclude	0.933	5.47	0.962
Model top-16	0.947	3.75	0.979
Manual v2: тільки ok	0.942	4.89	0.969
Model top-40	0.959	2.57	0.987

На перший погляд такий результат може здаватися суперечливим (рис. 4.2-4.3). Якщо композитний показник має $AUC = 0{,}722$ відносно ручно встановленої категорії exclude, виникає питання, чому модельне ранжування перевершує ручну фільтрацію. Пояснення полягає в тому, що в цих двох випадках оцінюються різні аспекти якості. Ручний перегляд визначає, чи відповідає маска візуальним і протокольним вимогам до коректної розмітки. Модельний показник, своєю чергою, відображає, наскільки окремий випадок порушує узгодженість між прогнозами моделей, наявною розміткою та типовою анатомічною формою, а отже, наскільки сильно він може спотворювати підсумкові метрики. Повна анатомічна істина залишається недоступною без узгодженої оцінки кількох експертів, однак саме розбіжність між ручною оцінкою розмітки та її впливом на метрики робить детектор корисним для контролю якості.

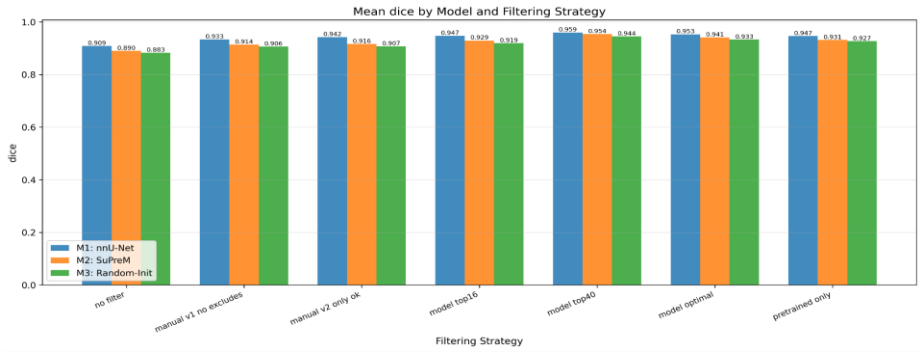

Рисунок 4.2 – Графік порівняння моделей за показником Mean Dice

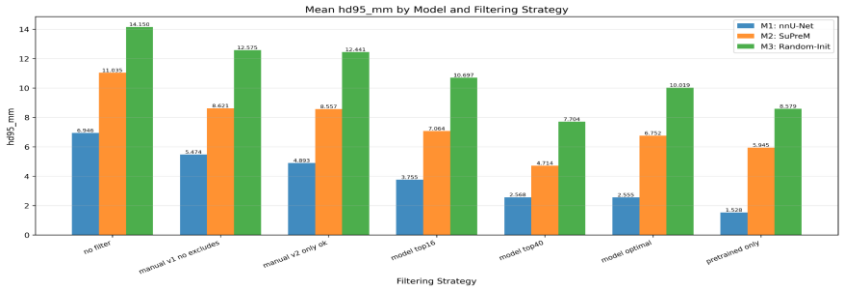

Рисунок 4.3 – Графік порівняння моделей за показником Mean HD95_MM

Отримані результати підтверджують доцільність використання детектора для пріоритизації ручного перегляду. Він не має самостійно ухвалювати клінічні рішення, проте дає змогу сформувати список випадків, перевірка яких найімовірніше буде корисною. У частині з них може бути виявлено явний дефект розмітки, в інших, складну анатомічну ситуацію, що непропорційно впливає на підсумкові метрики.

4.1.5 Перевірка проти випадкового вилучення

Щоб виключити пояснення, за яким метрики зростають лише через саме зменшення вибірки, виконано перестановочне тестування. Для $K = 16$, 24 і 40 випадків по 1000 разів формувалися випадкові підмножини вилучення, після чого їх результати порівнювалися з модельним фільтром.

Для $K = 16$ nnU-Net після модельної фільтрації мав Dice 0,947 проти середнього випадкового значення 0,909, HD95 3,75 мм проти 6,97 мм і NSD@2mm 0,979 проти 0,938. Величина ефекту за Cohen's d становила відповідно 5,73, 4,47 і 5,63. Подібна картина зберігалася і для інших моделей (табл. 4.8, рис. 4.4)

Таблиця 4.8 – Результати перестановочного тестування для $K = 16$

K	Модель	Метрика	Model-based	Random mean	p-value	Cohen's d
16	M1 nnU-Net	Dice	0.947	0.909	<0.001	5.73
16	M1 nnU-Net	HD95, мм	3.75	6.97	<0.001	4.47
16	M1 nnU-Net	NSD@2mm	0.979	0.938	<0.001	5.63
16	M2 SuPreM	Dice	0.929	0.890	<0.001	5.28
16	M3 Random	Dice	0.919	0.883	<0.001	4.54

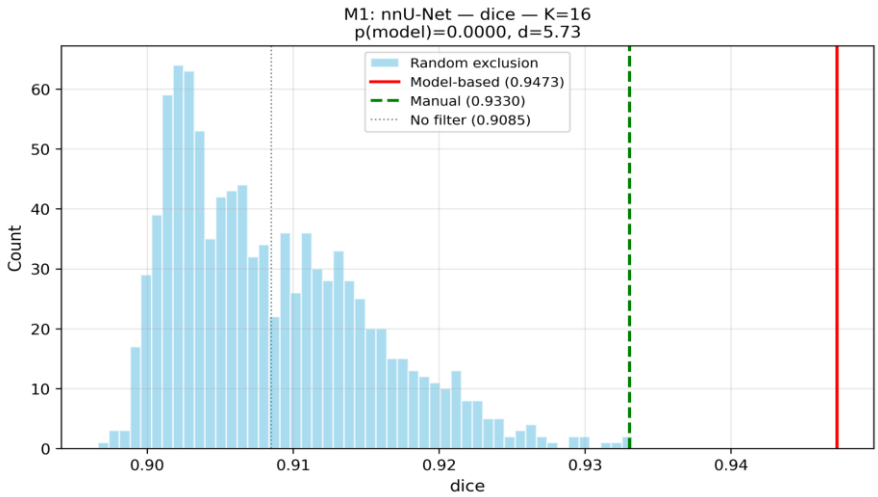

Рисунок 4.4 – Результати перестановочного тестування

Для $K = 40$ ефект також залишався вираженим. Усі 27 комірок аналізу для трьох моделей, трьох метрик і трьох значень K мали $p < 0{,}05$, а Cohen's d лежав у межах від 2,19 до 5,73 (табл.4.9).

Таблиця 4.9 – Результати перестановочного тестування для $K = 40$

K	Модель	Метрика	Model-based	Random mean	p-value	Cohen's d
40	M1 nnU-Net	Dice	0.959	0.908	<0.001	4.13
40	M1 nnU-Net	HD95, мм	2.57	6.91	<0.001	3.05
40	M1 nnU-Net	NSD@2mm	0.987	0.938	<0.001	3.65
40	M2 SuPreM	Dice	0.954	0.889	<0.001	4.78
40	M3 Random	Dice	0.944	0.882	<0.001	4.35

Отже, підвищення метрик не є наслідком випадкового скорочення контрольної вибірки. Детектор справді концентрує у верхній частині ранжування випадки, які мають найбільший вплив на якість оцінювання.

4.1.6 Оцінка впливу шуму розмітки на контрольні метрики

Послідовне вилучення випадків із найвищим композитним показником дозволило оцінити, наскільки сильно шум розмітки занижує виміряну якість моделей. Для nnU-Net Dice на повній контрольній вибірці становив 0,909. Після вилучення 15,4 % найаномальніших випадків він зріс до 0,947, після 25% до 0,953, а після 40 % до 0,960 (табл. 4.10).

Таблиця 4.10 – Зміна Dice після поетапного вилучення найаномальніших випадків

Модель	Raw Dice	15.4% вилучено	25% вилучено	40% вилучено
M1 nnU-Net	0.909	0.947 (+3.9 п.п.)	0.953 (+0.5 п.п. від 15.4%)	0.960 (+0.6 п.п. від 25%)
M2 SuPreM	0.890	0.929 (+3.9 п.п.)	0.942 (+1.3 п.п. від 15.4%)	0.956 (+1.3 п.п. від 25%)
M3 Random-init	0.883	0.919 (+3.6 п.п.)	0.935 (+1.7 п.п. від 15.4%)	0.945 (+1.0 п.п. від 25%)

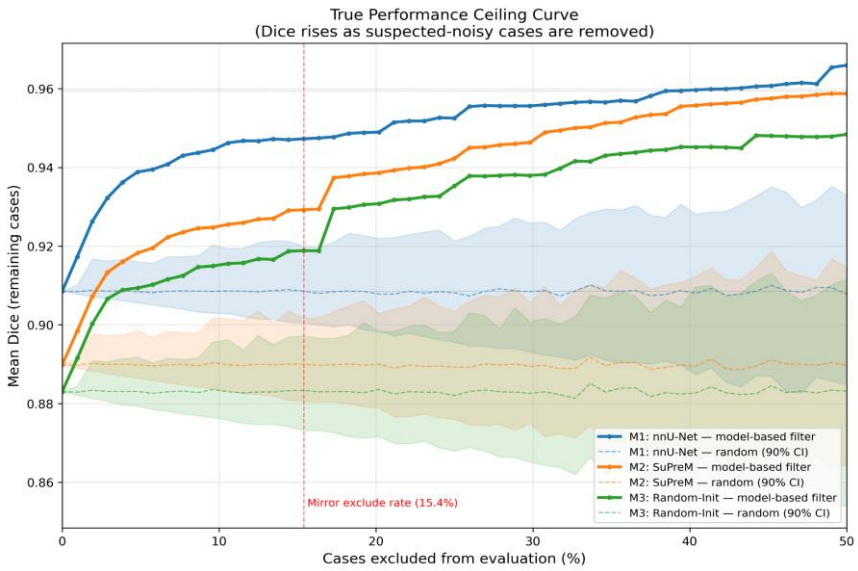

Рисунок 4.5 – Залежність Dice від частки вилучених випадків за композитним показником аномальності

Основний приріст зосереджений у перших 15–25 % ранжування, після чого крива переходить у ділянку насичення. Для nnU-Net різниця між оцінкою без фільтрації та рівнем після агресивнішого очищення становить близько п’яти процентних пунктів. Таку величину доцільно розглядати як емпіричну оцінку заниження метрики, пов’язаного з якістю контрольної розмітки. Вона не є часткою помилкових масок, але показує масштаб впливу шуму розмітки на підсумковий результат.

4.1.7 Зіставлення ручних і модельних позначень

Порівняння ручних і модельних позначень показало, що вони частково перетинаються, але не збігаються повністю. Для top-40 випадків 20 були позначені обома підходами, 20 лише моделлю і 20 лише під час ручного перегляду (табл. 4.11, рис. 4.6).

Таблиця 4.11 – Перетин ручних і модельних позначень для top-40 випадків

Категорія перетину при top-40	N випадків	Інтерпретація
Позначені обома підходами	20	Зона згоди між ручним переглядом і detector score
Позначені лише моделлю	20	Метрично токсичні або складні випадки, які ручний перегляд не відкинув
Позначені лише ручним переглядом	20	Сумнівні для рев'юера випадки, які моделі сегментують стабільно

За top-16 перетин становив лише п’ять випадків. До спільної групи увійшли переважно грубі дефекти, зокрема зміщені маски та розмітка іншої структури. Група, позначена лише моделлю, містила переважно випадки зі зниженим середнім Dice, які могли бути або метрично складними, або мати тонкі дефекти, не зафіксовані під час першого перегляду. Навпаки, частина випадків, позначених лише людиною, мала високий середній Dice, що

свідчило про стабільне відтворення межі моделями попри сумнів під час ручної оцінки.

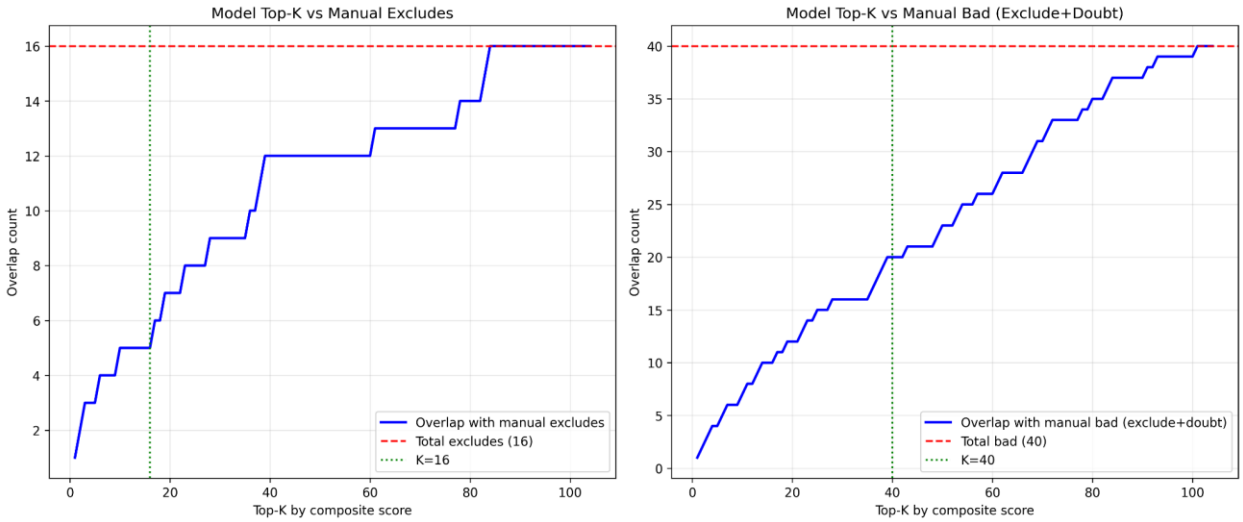

Рисунок 4.6 – Порівняння вибраних випадків

Такий розподіл підтверджує, що ручний перегляд і модельне ранжування відображають різні аспекти якості. Перший оцінює видиму коректність маски, друге виявляє випадки з найбільшим впливом на метрики та узгодженість прогнозів.

4.1.8 Якісний аналіз окремих випадків

Якісний перегляд окремих прикладів підтвердив числові результати. До групи випадків, позначених і людиною, і моделлю, увійшли грубі дефекти розмітки(табл. 4.12). У випадку 1494 середній Dice дорівнював 0,000 через повністю зміщену маску, а у випадку 705 значення 0,294 відповідало неповній розмітці однієї зі сторін клиноподібної пазухи (рис. .4.7-4.8).

Таблиця 4.12 – Випадки, позначені одночасно ручним переглядом і детектором

True positives	Manual	Composite score	Mean Dice	Інтерпретація
1494	exclude	0.998	0.000	Маска повністю зміщена
patient306	exclude	0.993	0.000	Розмічено не ту структуру
705	exclude	0.976	0.294	Неповна розмітка
1469	doubt	0.942	0.621	Розширення або протокольна неоднозначність
1378	exclude	0.894	0.833	Просторова неузгодженість або дефект слайсів

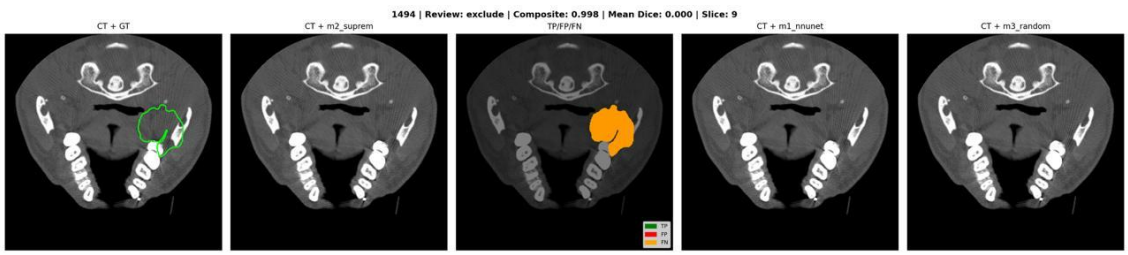

Рисунок 4.7 – Приклади явних дефектів розмітки клиноподібної пазухи: а) повністю зміщена маска у випадку 1494;

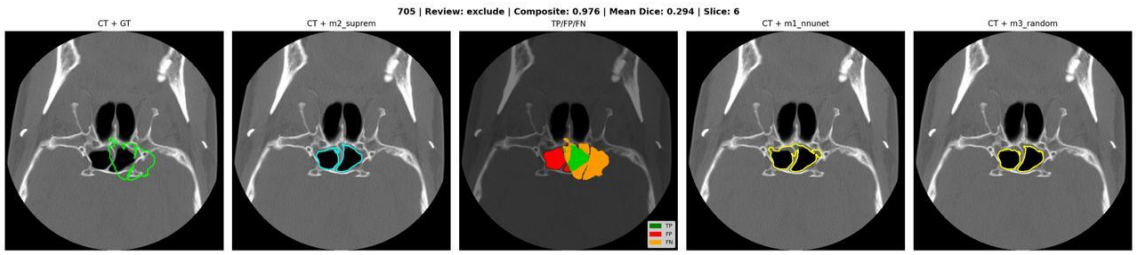

Рисунок 4.8 – Приклади явних дефектів розмітки клиноподібної пазухи: б) неповна розмітка у випадку 705

Випадки, позначені лише моделлю, мали середній Dice у діапазоні 0,743–0,895. Їх не можна автоматично вважати дефектними. Частина з них може відповідати складній анатомії, запальним змінам або нетиповому

контрасту, інша частина може містити дрібні помилки розмітки, що потребують повторного перегляду (4.13).

Таблиця 4.13 – Випадки, позначені лише детектором

Model-only marked	Manual	Composite score	Mean Dice	Можлива інтерпретація
patient1564	ok	0.916	0.743	Складна анатомія або тонкий дефект
patient1570	ok	0.873	0.787	Складний випадок для моделей
1389	ok	0.865	0.803	Метрично токсичний приклад
1603	ok	0.803	0.895	Помірна невідповідність
patient1582	ok	0.750	0.814	Складна, але прийнята розмітка

Випадки, позначені лише під час ручного перегляду, навпаки, мали високий середній Dice, переважно в межах 0,941–0,955. Це свідчить, що категорія doubt часто фіксує не явний дефект маски, а невпевненість щодо складного або нетипового випадку.

Таблиця 4.14 – Випадки, позначені лише ручним переглядом

Human-only marked	Manual	Composite score	Mean Dice	Інтерпретація
1518	doubt	0.563	0.944	Сумнів без явного модельного конфлікту
patient340	doubt	0.526	0.951	Висока відтворюваність межі
patient508	doubt	0.526	0.953	Висока узгодженість моделей із GT
patient225	doubt	0.471	0.941	Протокольна невпевненість
1483	doubt	0.471	0.955	Складний, але когерентний випадок

До контрольної групи ввійшли випадки, де ручна оцінка та оцінка підозрілості розмітки узгоджено вказували на відсутність проблем. Для них середній Dice становив 0,972–0,986.

Таблиця 4.15 – Узгоджено чисті випадки контрольної вибірки

Consensus clean	Manual	Composite score	Mean Dice	Інтерпретація
patient530	ok	0.055	0.972	Узгоджена розмітка
patient461	ok	0.103	0.986	Дуже висока згода
patient517	ok	0.120	0.981	Стабільний випадок
1124	ok	0.125	0.983	Стабільний випадок
patient313	ok	0.147	0.981	Стабільний випадок

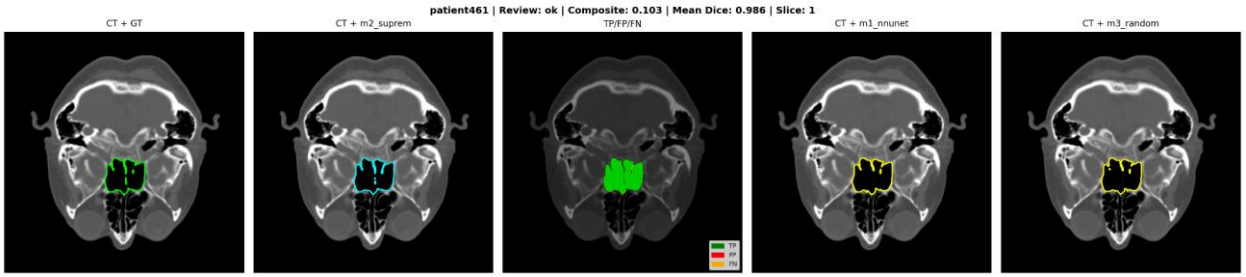

Рисунок 4.9 – Приклад випадку з повною узгодженістю

Узагальнення результатів першого етапу показує, що детектор має практичну цінність як інструмент контролю якості. Композитний показник досяг AUC 0,722 та AP 0,439, а модельна фільтрація перевищила ручну за впливом на Dice, HD95 і NSD@2mm. Найбільш обґрунтованим сценарієм його застосування є пріоритизація випадків для експертного перегляду, де верхня частина ранжування містить як явні дефекти розмітки, так і випадки з найбільшим впливом на якість оцінювання.

4.2 Перевірка автоматичного очищення навчальної вибірки

Після підтвердження придатності детектора для контролю якості було перевірено значно жорсткіший сценарій: автоматичне вилучення підозрілих випадків із навчальної вибірки з подальшим повторним навчанням моделі. Такий експеримент дозволяє встановити, чи можна перетворити корисний інструмент ранжування на повністю автоматичний механізм покращення навчальних даних.

4.2.1 Вибір частки вилучення

Використано два рівні очищення, 5 % і 15,4 %. Значення 15,4 % відповідає частці випадків exclude у контрольній вибірці, тобто 16 із 104 досліджень. Значення 5 % обрано як обережніший режим, у якому верхівка ранжування мала вищу точність і містила меншу кількість потенційно корисних складних випадків.

Такий дизайн дозволяє перевірити не одну довільну точку, а залежність ефекту від частки вилучення. Якщо автоматичне очищення є корисним, то принаймні консервативний режим 5 % мав би бути нейтральним або позитивним для якості моделі.

4.2.2 Очищення навчального пулу та схема чотирьох комірок

Для уникнення витоку інформації кожен випадок тренувального пулу оцінювався моделлю, яка не використовувала його під час навчання. У результаті отримано 430 прогнозів поза фолдом, на основі яких формувалося ранжування за методом C. Для режиму 5 % було вилучено 22 випадки, для 15,4 % – 66 випадків. Після цього модель SuPreM SegResNet навчалася повторно на очищених підмножинах за незмінної архітектури та тих самих гіперпараметрів.

Оцінювання виконувалося за схемою чотирьох комірок: модель, навчена на необробленій або очищеній вибірці, перевірялася на повній або очищеній контрольній вибірці. Така побудова дозволила відокремити вплив очищення навчання від механічного підвищення метрик після очищення тесту.

4.2.3 Результати повторного навчання

Автоматичне очищення не покращило модель навіть у консервативному режимі. За вилучення 5 % навчальних випадків Dice на повній контрольній вибірці зменшився з 0,890 до 0,870, а HD95 зріс з 11,03 до 22,13 мм. Погіршення зберігалося і після оцінювання на очищеній контрольній вибірці (табл. 4.16).

Таблиця 4.16 – Результати повторного навчання моделі після очищення 5 % навчальної вибірки

Train / test, K=5%	Raw test, n=104	Cleaned test, n=99
Raw train, baseline	Dice 0,890; HD95 11,03 мм; NSD@2mm 0,917	Dice 0,918; HD95 8,92 мм; NSD@2mm 0,948
Cleaned train	Dice 0,870; HD95 22,13 мм; NSD@2mm 0,886	Dice 0,898; HD95 20,50 мм; NSD@2mm 0,915

За вилучення 15,4 % ефект посилився. На повній контрольній вибірці Dice знизився до 0,831, а HD95 зріс до 25,05 мм. Навіть після паралельного очищення тесту модель, навчена на скороченій вибірці, залишалася гіршою за базову (табл. 4.17).

Таблиця 4.17 – Результати повторного навчання моделі після очищення 15,4 % навчальної вибірки

Train / test, K=15,4%	Raw test, n=104	Cleaned test, n=88
Raw train, baseline	Dice 0,890; HD95 11,03 мм; NSD@2mm 0,917	Dice 0,929; HD95 7,06 мм; NSD@2mm 0,960
Cleaned train	Dice 0,831; HD95 25,05 мм; NSD@2mm 0,837	Dice 0,873; HD95 21,50 мм; NSD@2mm 0,881

Очищення тестової вибірки саме по собі підвищувало метрики, що очікувано після вилучення складних або дефектних випадків. Проте очищення навчальної вибірки мало протилежний ефект і знижувало якість повторно навченої моделі. Отже, позитивний результат першого етапу не переноситься автоматично на сценарій сліпого вилучення навчальних прикладів.

4.2.4 Залежність ефекту від частки вилучення

Порівняння двох рівнів очищення сформувало чіткий профіль деградації. При вилученні 5 % навчальних прикладів втрата Dice на повному тесті становила 0,020, а при вилученні 15,4 % вже 0,059. Для HD95 погіршення становило відповідно 11,1 та 14,0 мм (табл. 4.18).

Таблиця 4.18 – Залежність ефекту від частки автоматично вилучених навчальних випадків

K	Precision@K на holdout	Вилучено з train pool	C−A Dice	D−B Dice	C−A HD95	D−B HD95	Інтерпретація
0%	-	0	baseline	baseline	baseline	baseline	Навчання без автоматичного очищення
5%	0,60	22	−0,020	−0,021	+11,1 мм	+11,6 мм	Консервативне очищення вже погіршує нижній хвіст
15,4%	0,31	66	−0,059	−0,057	+14,0 мм	+14,4 мм	Агресивніше очищення посилює деградацію

Наявність негативного ефекту вже за 5 % показує, що проблема не зводиться до надто агресивного порогу. Сам сигнал методу C змішує два різні типи випадків: помилкові маски і складні, але коректно розмічені приклади.

4.2.5 Зміни в нижньому хвості розподілу

Середнє зниження Dice не передає повної структури погіршення. Після очищення 15,4 % навчальної вибірки верхня частина розподілу майже не змінилася, тоді як нижній хвіст суттєво просів. На очищеному тесті

значення P5 знизилося з 0,65 до 0,47, P10 з 0,79 до 0,60, а P25 з 0,90 до 0,76. Кількість випадків із Dice нижче 0,7 зросла з 7 до 19, а з Dice нижче 0,8, з 11 до 32 (табл. 4.19).

Таблиця 4.19 – Зміна квантилів Dice після автоматичного очищення навчальної вибірки

Показник нижнього хвоста, K=15,4%	Baseline B	Cleaned D	Зсув
P5 Dice	0,65	0,47	−0,19
P10 Dice	0,79	0,60	−0,19
P25 Dice	0,90	0,76	−0,14
P50 Dice	0,94	0,92	−0,03
P75 Dice	0,96	0,96	0,00
P95 Dice	0,98	0,98	0,00

Результат означає, що модель не стала рівномірно слабшою. Найбільше постраждали саме складні випадки, для яких правильне відтворення межі було критичним.

4.2.6 Зрив межі на складних випадках

Найбільш показовим проявом деградації став зрив межі. Для окремих випадків HD95 зростав у 8–24 рази, хоча зменшення Dice могло виглядати менш драматичним. Наприклад, для випадку 1453 Dice знизився з 0,89 до 0,59, а HD95 зріс з 3,0 до 68,2 мм; для випадку 1190 відповідні зміни становили 0,92 до 0,68 і 3,0 до 71,2 мм (табл. 4.20).

Таблиця 4.20 – Приклади зриву межі після автоматичного очищення навчальної вибірки

Case	Dice baseline -> cleaned	HD95 baseline -> cleaned	Коментар
1453	0,89 -> 0,59	3,0 -> 68,2 мм	Зрив межі, HD95 ×22,7
1190	0,92 -> 0,68	3,0 -> 71,2 мм	Зрив межі, HD95 ×23,7
1158	0,91 -> 0,68	6,0 -> 57,0 мм	Зрив межі, HD95 ×9,5
1410	0,95 -> 0,71	3,0 -> 52,4 мм	Зрив межі, HD95 ×17,5
1179	0,90 -> 0,68	9,0 -> 71,3 мм	Зрив межі, HD95 ×7,9

Такий тип помилки є особливо важливим для медичної сегментації, оскільки просторова похибка межі має інше клінічне значення, ніж локальна втрата невеликої ділянки маски. Саме тому одночасний аналіз Dice та HD95 є необхідним для коректної інтерпретації результатів.

4.2.7 Ручний перегляд вилучених навчальних випадків

Щоб з’ясувати причину деградації, вручну переглянуто 22 випадки, вилучені в режимі 5 %. Серед них 10 випадків відповідали реальним помилкам розмітки, 9 були складними, але коректно розміченими, і ще 3 виявилися хибними спрацюваннями детектора.

Серед складних, але коректних випадків переважали дослідження із запаленням або м’якотканинним заповненням порожнини (рис. 4.10). Для таких даних правильна маска має визначатися кістковою межею, а не наявністю повітря в порожнині. Модель SuPreM, що значною мірою спирається на сигнал повітряної щільності, виявляла такі випадки як аномальні, хоча вони є необхідними для навчання. Унаслідок їх автоматичного вилучення модель втрачала саме ті приклади, які мали б компенсувати її систематичну слабкість.

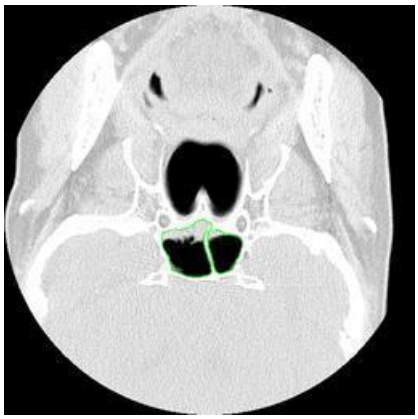

Рисунок 4.10 – Приклади складних, але коректно розмічених випадків, помилково вилучених із навчальної вибірки

Отримані результати узгоджуються між собою. Схема чотирьох комірок показала падіння якості для обох рівнів очищення. Залежність від частки вилучення засвідчила посилення деградації зі збільшенням K. Аналіз нижнього хвоста локалізував погіршення на складних випадках, а приклади зриву межі показали просторовий характер помилки. Ручний перегляд вилучених випадків пояснив механізм негативного ефекту: детектор виявляє не лише помилки розмітки, а й клінічно важливі приклади, на яких конкретна архітектура має невпевненість.

Тому автоматичне очищення навчальної вибірки за сигналом однієї моделі у цій постановці не підтверджено. Негативний результат не спростовує цінність детектора як інструмента контролю якості, але показує обмеження його безекспертного використання.

4.3 Потенціал використання детектора в активному навчанні

Результати попередніх експериментів показали, що детектор аномалій розмітки має практичну цінність, але не є універсальним механізмом автоматичного вилучення прикладів із навчальної вибірки. Під час ручного перегляду випадків, вилучених за найвищими значеннями показника аномальності, було встановлено, що поряд із реальними дефектами розмітки до верхньої частини ранжування потрапляють складні, але коректно розмічені дослідження. Саме вони виявилися важливими для навчання моделі, а їх автоматичне видалення призвело до погіршення результатів повторного навчання.

У сучасних підходах до активного навчання медичних зображень зазвичай виділяють два ключові критерії відбору прикладів: інформативність і представницькість. Інформативними вважаються випадки, для яких модель демонструє високу невизначеність, низьку впевненість або суттєву розбіжність прогнозів. Представницькість потрібна для того, щоб вибірка не складалася лише з дуже схожих складних випадків

і зберігала різноманітність клінічного матеріалу. Саме поєднання цих двох ознак вважається більш надійним, ніж простий відбір лише за невизначеністю.

У межах цієї роботи детектор уже містить придатні для такого сценарію сигнали. Міжмодельна розбіжність не потребує наявної розмітки та може бути обчислена для нерозмічених КТ-досліджень. Показник упевненості попередньо навченої моделі також дозволяє виявляти випадки, які виходять за межі типових для моделі анатомічних представлень. Після додавання механізму контролю різноманітності ці ознаки можуть бути використані для формування черги досліджень, які доцільно передавати експерту на розмітку в першу чергу.

Таблиця 4.21 – Порівняння практичних сценаріїв використання детектора

Сценарій	Вхідні дані	Роль детектора	Роль експерта
Контроль якості наявної розмітки	Розмічені КТ-дослідження	Ранжує підозрілі випадки	Перевіряє верхню частину ранжування та відокремлює дефекти від складних випадків
Автоматичне очищення навчальної вибірки	Розмічені КТ-дослідження	Автоматично вилучає випадки з високим показником аномальності	Не залучається
Активне навчання	Нерозмічений пул КТ-досліджень	Відбирає найбільш інформативні випадки для анотування	Розмічає відібрані дослідження

Практична схема активного навчання може бути побудована у вигляді ітеративного циклу. На початковому етапі модель навчається на невеликій розміченій вибірці. Далі вона формує прогнози для нерозміченого пулу, після чого детектор обчислює показники невизначеності та розбіжності. З отриманого ранжування обирається обмежений набір найбільш інформативних і водночас різноманітних випадків, які передаються експерту для ручної сегментації. Після додавання нових масок до

навчальної вибірки модель перенавчається, і цикл повторюється до досягнення потрібної якості або вичерпання бюджету анотування. Така логіка відповідає сучасному pool-based active learning, де метою є досягнення тієї самої якості за меншої кількості розмічених прикладів або вищої якості за незмінного бюджету анотування.

У поточній роботі активне навчання не перевірялося як окремий експеримент. Не будувалися криві навчання за різних стратегій відбору, не порівнювалися випадковий відбір, відбір за невизначеністю та відбір з урахуванням різноманітності, а також не оцінювалася економія бюджету розмітки. Тому робити висновок про фактичне покращення якості моделі за рахунок запропонованого детектора було б передчасно. Водночас результати попередніх підрозділів створюють обґрунтовану підставу для такого напряму подальшої роботи. Детектор уже підтвердив здатність виділяти випадки, що є складними для модельної системи, а негативний результат автоматичного очищення показав, що ці випадки доцільніше не вилучати, а передавати на експертне опрацювання.

Отже, найбільш коректною інтерпретацією активного навчання в межах цієї роботи є не готовий прикладний результат, а наступний логічний спосіб використання перевіреного сигналу детектора. Якщо для контролю якості він допомагає експерту швидше знаходити проблемні маски, то для активного навчання може допомагати раціональніше витрачати час експерта на розмітку тих КТ-досліджень, які потенційно найбільше розширюють знання моделі про складні варіанти анатомії.

У четвертому розділі експериментально перевірено можливості побудованого детектора. Отримані результати підтвердили його придатність для ранжування випадків і пріоритизації експертного перегляду, оскільки композитний показник досяг AUC 0,722, а модельна фільтрація перевищила ручну за впливом на підсумкові метрики. Водночас автоматичне вилучення підозрілих випадків із навчальної вибірки не покращило модель і призвело до деградації на складних прикладах, тому

найбільш обґрунтованим напрямом подальшого використання детектора є контроль якості за участі експерта та активне навчання, а не повністю автоматичне очищення даних.

ВИСНОВКИ

У дипломній роботі проведено дослідження моделей згортальних нейронних мереж глибинного навчання для тривимірної сегментації клиноподібної пазухи на зображеннях комп'ютерної томографії та запропоновано процедуру попередньої обробки даних, яка враховує неоднорідну якість наявної експертної розмітки. За результатами роботи отримано такі основні результати.

Виконано аналіз сучасних моделей глибинного навчання для сегментації медичних зображень та обґрунтовано вибір тривимірних згорткових архітектур повної просторової роздільної здатності як базової основи для практичної частини. До складу робочого ансамблю включено три моделі з диверсифікованими архітектурними передумовами та джерелами вагової ініціалізації: модель M1 на основі PlainConvUNet 3D у складі фреймворку nnU-Net з випадковою ініціалізацією; модель M2 на основі архітектури SegResNet з попереднім навчанням на вагах SuPreM; модель M3 на основі тієї ж архітектури SegResNet з випадковою ініціалізацією. Усі три моделі досягли стабільної якості сегментації на валідаційній підмножині (Dice 0,929; 0,897; 0,893 відповідно), достатньої для використання у складі ансамблю.

Проаналізовано анатомічні особливості клиноподібної пазухи як цільової структури автоматичної сегментації на КТ. Підтверджено, що мала величина об'єму, значна варіабельність пневматизації та близькість до критичних структур основи черепа роблять цю анатомічну структуру складним об'єктом як для класичних, так і для нейромережевих методів сегментації, що зумовлює доцільність використання тривимірного контексту КТ-об'єму.

Виконано попередню обробку даних КТ-досліджень та проведено аналіз якості наявної експертної розмітки контрольної вибірки зі 104 досліджень. Встановлено, що близько 15 % випадків мають явні дефекти розмітки, а ще

приблизно 23 % характеризуються сумнівною якістю. Сформовано типологію виявлених дефектів, яка охоплює близько 95 % спостережуваних випадків і включає п'ять основних класів: зміщені маски, недосегментацію, надсегментацію, дублікати зрізів та помилкову розмітку іншої анатомічної структури. Цей результат підтверджує об'єктивну необхідність розширення стандартного процесу попередньої обробки даних окремою процедурою контролю якості розмітки.

Запропоновано та реалізовано підхід до автоматичного виявлення потенційно дефектних випадків розмітки на основі ансамблю CNN-моделей сегментації. Підхід поєднує чотири незалежні методи оцінювання підозрілості розмітки (метричне відхилення, міжмодельна розбіжність, упевненість попередньо навченої моделі, ансамблевий консенсус) та композитний показник на основі рангового злиття. Поріг прийняття рішення калібровано за критерієм Юдена на цільовому класі явно дефектних випадків з використанням незалежного експертного перегляду як еталона.

Проведено експериментальну перевірку запропонованого підходу та його застосувань. Композитний показник підозрілості розмітки досяг AUC 0,722 на контрольній вибірці при випадковому рівні 0,5 та статистично значущій кореляції з експертною оцінкою ($\rho = 0{,}272$, $p = 0{,}0053$), що підтверджує гіпотезу H1 про узгодженість сигналу детектора з незалежним експертним судженням. Порівняння модельної та ручної фільтрації показало, що вилучення випадків за модельним ранжуванням підвищило Dice для nnU-Net до 0,947 при $K = 16$ та до 0,959 при $K = 40$, що перевищує результати ручного виключення на тій самій кількості випадків (0,933 та 0,942 відповідно) і підтверджує гіпотезу H2 про валідність детектора як інструменту контролю якості. Гіпотезу H3 про доцільність повністю автоматичного вилучення підозрілих випадків із навчальної вибірки спростовано: при $K = 5$ % виявлено зниження якості на $\Delta\text{Dice} = -0{,}020$, а при $K = 15{,}4$ % деградація сягала $\Delta\text{Dice} = -0{,}057$, що зумовлено помилковим

вилученням анатомічно складних, але коректно розмічених випадків. Гіпотезу H4 щодо придатності того самого сигналу для активного навчання обґрунтовано теоретично з огляду на структурні особливості міжмодельної розбіжності, яка не потребує наявної розмітки і може обчислюватися для нерозмічених КТ-досліджень; її експериментальна перевірка віднесена до напрямків подальших досліджень.

На основі отриманих результатів сформульовано практичні рекомендації щодо вдосконалення процесу попередньої обробки даних і навчання моделей сегментації КТ. Кількісна оцінка обмеження якості сегментації, обумовленого шумом розмітки, для цільового датасету становить близько 5 п.п. за метрикою Dice і може використовуватися як орієнтир для подальших робіт. Встановлено, що скорочений варіант детектора на основі лише попередньо навченої моделі SuPreM забезпечує AUC = 0,705 проти 0,722 у повного композитного показника, що дозволяє реалізувати процедуру контролю якості розмітки з істотно меншими обчислювальними витратами.

Отримані результати обмежені одним датасетом клиноподібної пазухи та одним оглядачем експертного перегляду, тому їх перенос на інші анатомічні структури та клінічні умови потребує окремого підтвердження. Подальшими напрямками розвитку дослідження є експериментальна перевірка застосування побудованого детектора у режимі активного навчання на нерозмічених КТ-досліджень, розширення складу ансамблю за рахунок архітектур інших класів (трансформерних та гібридних), а також проведення повноцінного експертного перегляду незалежною групою радіологів для підвищення достовірності еталонної оцінки.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Medical image segmentation: a comprehensive review of deep learning-based methods / Y. Gao et al. Tomography. 2025. Vol. 11, Iss. 5. URL: https://doi.org/10.3390/tomography11050052 (date of access: 10.05.2026).

2. Sphenoid sinus types, dimensions and relationship with surrounding structures / N. Štoković et al. Annals of Anatomy — Anatomischer Anzeiger. 2016. Vol. 203. P. 69–76. URL: https://doi.org/10.1016/j.aanat.2015.02.013 (date of access: 10.05.2026).

3. A survey on deep learning in medical image analysis / G. Litjens et al. Medical Image Analysis. 2017. Vol. 42. P. 60–88. URL: https://doi.org/10.1016/j.media.2017.07.005 (date of access: 10.05.2026).

4. Towards a guideline for evaluation metrics in medical image segmentation / D. Müller, I. Soto-Rey, F. Kramer. BMC Research Notes. 2022. Vol. 15. URL: https://doi.org/10.1186/s13104-022-06096-y (date of access: 10.05.2026).

5. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches / W. Khan et al. Bioengineering. 2024. Vol. 11, Iss. 10. URL: https://doi.org/10.3390/bioengineering11101034 (date of access: 10.05.2026).

6. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation / F. Isensee et al. Nature Methods. 2021. Vol. 18. P. 203–211. URL: https://doi.org/10.1038/s41592-020-01008-z (date of access: 10.05.2026).

7. The Medical Segmentation Decathlon / M. Antonelli et al. Nature Communications. 2022. Vol. 13. URL: https://doi.org/10.1038/s41467-022-30695-9 (date of access: 10.05.2026).

8. The KiTS21 challenge: automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT / N. Heller et al. arXiv.org. 2023. URL: https://arxiv.org/abs/2307.01984 (date of access: 10.05.2026).

9. AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation / Y. Ji et al. Advances in Neural Information Processing Systems (NeurIPS 2022). 2022. URL: https://arxiv.org/abs/2206.08023 (date of access: 10.05.2026).

10. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images / J. Wasserthal et al. Radiology: Artificial Intelligence. 2023. Vol. 5, Iss. 5. URL: https://doi.org/10.1148/ryai.230024 (date of access: 10.05.2026).

11. nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation / F. Isensee et al. Medical Image Computing and Computer-Assisted Intervention — MICCAI 2024. Cham, 2024. URL: https://arxiv.org/abs/2404.09556 (date of access: 10.05.2026).

12. Artificial intelligence applications in CT imaging: a systematic review / K. Eskandar. iRADIOLOGY. 2025. URL: https://doi.org/10.1002/ird3.70046 (date of access: 10.05.2026).

13. Deep learning-based multi-class segmentation of the paranasal sinuses of sinusitis patients based on computed tomographic images / J. Whangbo et al. Sensors. 2024. Vol. 24, Iss. 6. URL: https://doi.org/10.3390/s24061933 (date of access: 10.05.2026).

14. Comparison of segmentation performance of CNNs, vision transformers, and hybrid networks for paranasal sinuses with sinusitis on CT images / D. Song et al. Scientific Reports. 2025. Vol. 15. URL: https://doi.org/10.1038/s41598-025-17266-w (date of access: 10.05.2026).

15. Dice L. R. Measures of the amount of ecologic association between species. Ecology. 1945. Vol. 26, Iss. 3. P. 297–302. URL: https://doi.org/10.2307/1932409 (date of access: 12.04.2026).

16. Frénay B., Verleysen M. Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems. 2014. Vol. 25, Iss. 5. P. 845–869. URL: https://doi.org/10.1109/TNNLS.2013.2292894 (date of access: 27.03.2026).

17. Confident learning: estimating uncertainty in dataset labels / C. G. Northcutt, L. Jiang, I. L. Chuang. Journal of Artificial Intelligence Research. 2021. Vol. 70. P. 1373–1411. URL: https://doi.org/10.1613/jair.1.12125 (date of access: 03.05.2026).

18. Pham D. L., Xu C., Prince J. L. Current methods in medical image segmentation. Annual Review of Biomedical Engineering. 2000. Vol. 2. P. 315–337. URL: https://doi.org/10.1146/annurev.bioeng.2.1.315 (date of access: 10.05.2026).

19. A review of medical image segmentation algorithms / K. K. D. Ramesh et al. EAI Endorsed Transactions on Pervasive Health and Technology. 2021. Vol. 7, Iss. 27. URL: https://doi.org/10.4108/eai.12-4-2021.169184 (date of access: 10.05.2026).

20. Adams R., Bischof L. Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1994. Vol. 16, Iss. 6. P. 641–647. URL: https://doi.org/10.1109/34.295913 (date of access: 10.05.2026).

21. Pham D. L., Prince J. L. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging. 1999. Vol. 18, Iss. 9. P. 737–752. URL: https://doi.org/10.1109/42.802752 (date of access: 10.05.2026).

22. Chan T. F., Vese L. A. Active contours without edges. IEEE Transactions on Image Processing. 2001. Vol. 10, Iss. 2. P. 266–277. URL: https://doi.org/10.1109/83.902291 (date of access: 10.05.2026).

23. MNet: rethinking 2D/3D networks for anisotropic medical image segmentation / Z. Dong et al. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22). 2022. P. 870–876. URL: https://doi.org/10.24963/ijcai.2022/122 (date of access: 10.05.2026).

24. Z-Net: an anisotropic 3D DCNN for medical CT volume segmentation / P. Li et al. arXiv.org. 2019. URL: https://arxiv.org/abs/1909.07480 (date of access: 10.05.2026).

25. UNETR: transformers for 3D medical image segmentation / A. Hatamizadeh et al. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022. P. 574–584. URL: https://arxiv.org/abs/2103.10504 (date of access: 10.05.2026).

26. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images / A. Hatamizadeh et al. International MICCAI Brainlesion Workshop. Cham, 2022. P. 272–284. URL: https://arxiv.org/abs/2201.01266 (date of access: 10.05.2026).

27. nnFormer: volumetric medical image segmentation via a 3D transformer / H.-Y. Zhou et al. IEEE Transactions on Image Processing. 2023. Vol. 32. P. 4036–4045. URL: https://arxiv.org/abs/2109.03201 (date of access: 10.05.2026).

28. MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation / S. Roy et al. Medical Image Computing and Computer-Assisted Intervention — MICCAI 2023. Cham, 2023. P. 405–415. URL: https://arxiv.org/abs/2303.09975 (date of access: 10.05.2026).

29. STU-Net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training / Z. Huang et al. arXiv.org. 2023. URL: https://arxiv.org/abs/2304.06716 (date of access: 10.05.2026).

30. Deep learning model for automated segmentation of sphenoid sinus and middle skull base structures in CBCT volumes using nnU-Net v2 / İ. T. Gülşen et al. Oral Radiology. 2025. URL: https://doi.org/10.1007/s11282-025-00848-9 (date of access: 10.05.2026).

31. Exploring the sphenoid sinus as a biometric marker for human identification / V. Alekseeva et al. ProfIT AI'25: 5th International Workshop of IT-Professionals on Artificial Intelligence. CEUR Workshop Proceedings. 2026. Vol. 4164. P. 487–493. URL: https://ceur-ws.org/Vol-4164/short11.pdf (date of access: 10.05.2026).

32. nnU-Net: self-adapting framework for U-Net-based medical image segmentation / F. Isensee et al. arXiv.org. 2018. URL: https://arxiv.org/abs/1809.10486 (date of access: 10.05.2026).

33. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis / D. Karimi et al. Medical Image Analysis. 2020. Vol. 65. URL: https://doi.org/10.1016/j.media.2020.101759 (date of access: 10.05.2026).

34. An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation / Ș. Vădineanu et al. Proceedings of Machine Learning Research (MIDL 2022). 2022. Vol. 172. P. 1251–1267. URL: https://proceedings.mlr.press/v172/vadineanu22a.html (date of access: 10.05.2026).

35. Why rankings of biomedical image analysis competitions should be interpreted with care / L. Maier-Hein et al. Nature Communications. 2018. Vol. 9. URL: https://doi.org/10.1038/s41467-018-07619-7 (date of access: 10.05.2026).

36. Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. Cham, 2015. P. 234–241. URL: https://doi.org/10.1007/978-3-319-24574-4_28 (date of access: 18.02.2026).

37. 3D U-Net: learning dense volumetric segmentation from sparse annotation / Ö. Çiçek et al. Medical Image Computing and Computer-Assisted Intervention — MICCAI 2016. Cham, 2016. P. 424–432. URL: https://doi.org/10.1007/978-3-319-46723-8_49 (date of access: 09.04.2026).

38. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Cham, 2019. P. 311–320. URL: https://doi.org/10.1007/978-3-030-11726-9_28 (date of access: 25.03.2026).

39. MONAI: an open-source framework for deep learning in healthcare / M. J. Cardoso et al. arXiv.org. 2022. URL: https://arxiv.org/abs/2211.02701 (date of access: 14.04.2026).

40. SuPreM: a foundation model for 3D medical image segmentation through supervised pre-training / W. Li et al. arXiv.org. 2024. URL: https://arxiv.org/abs/2306.00988 (date of access: 06.05.2026).

41. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30. P. 6402–6413. URL: https://arxiv.org/abs/1612.01474 (date of access: 21.02.2026).

42. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning (ICML). 2016. P. 1050–1059. URL: https://arxiv.org/abs/1506.02142 (date of access: 02.04.2026).