МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Західноукраїнський національний університет
Факультет комп’ютерних інформаційних технологій
Кафедра комп'ютерної інженерії

Копія Андрій Вікторович

Алгоритми формування наборів даних для навчання глибоких нейронних мереж / Algorithms for forming data sets for training deep neural networks

спеціальність: 123 - Комп’ютерна інженерія

освітньо-професійна програма - Комп’ютерна інженерія

Кваліфікаційна робота

Виконав студент групи КІм-21

А. В. Копія

_____

Науковий керівник:

д.т.н., професор, О. М. Березький

_____

ТЕРНОПІЛЬ - 2025

АНОТАЦІЯ

Копія А.В. Алгоритми формування наборів даних для навчання глибоких нейронних мереж. – Рукопис.

Кваліфікаційна робота на здобуття освітнього ступеня «магістр» за спеціальністю 123 «Комп'ютерна інженерія», освітньо-професійна програма. Західноукраїнський національний університет, Тернопіль, 2025.

Робота написана обсягом 76 сторінка і містить 27 ілюстрацій, 10 таблиць, 1 додаток та 37 джерел за переліком посилань. Метою роботи є розроблення алгоритмів формування навчальних наборів даних для глибоких нейронних мереж із урахуванням вимог до їх якості, узагальнювальної здатності в сучасних інформаційних системах.

Методи дослідження. У роботі використано методи машинного навчання та комп'ютерного зору, методи аналізу зображень, а також інструменти програмної інженерії для реалізації алгоритмів у програмному середовищі Paint.NET.

Результати дослідження: проведено аналіз джерел формування навчальних наборів даних, розроблено методику їх структуризації. У практичній частині сформовано датасет імуногістохімічних зображень та проведено експериментальне дослідження ефективності запропонованих алгоритмів.

Результати роботи можуть бути використані під час розроблення систем комп'ютерного зору, у задачах медичної діагностики, у створенні високоякісних навчальних наборів даних, а також у навчальному процесі.

Орієнтовні напрямки розвитку досліджень: удосконалення автоматизованих методів формування датасетів, застосування алгоритмів у клінічних інформаційних системах, створення комплексних платформ для обробки медичних зображень.

КЛЮЧОВІ СЛОВА: НАВЧАЛЬНІ ДАНІ, ГЛИБОКЕ НАВЧАННЯ, СЕГМЕНТАЦІЯ, АНОТАЦІЯ, АУГМЕНТАЦІЯ, ДАТАСЕТ, PAINT.NET.

ANNOTATION

Kopia A.V. Algorithms for the Formation of Datasets for Training Deep Neural Networks. – Manuscript.

Master's qualification work for obtaining the educational degree "Master" in specialty 123 "Computer Engineering", educational and professional program. West Ukrainian National University, Ternopil, 2025.

The work consists of 76 pages and includes 27 illustrations, 10 tables, 1 appendix, and 37 references. The aim of the research is to develop algorithms for forming training datasets for deep neural networks, taking into account requirements for their quality and generalization ability in modern information systems.

Research methods. The study employs methods of machine learning and computer vision, image analysis techniques, as well as software engineering tools for implementing the algorithms in the Paint.NET environment.

Research results: an analysis of data sources for forming training datasets was conducted, and a methodology for their structuring was developed. In the practical part, an immunohistochemical image dataset was created, and an experimental evaluation of the proposed algorithms was carried out.

The results of the work can be used for the development of computer vision systems, medical diagnostic tools, the creation of high-quality training datasets, and in the educational process.

Prospective directions of further research: improvement of automated methods for dataset formation, application of the algorithms in clinical information systems, and development of integrated platforms for medical image processing.

KEYWORDS: TRAINING DATA, DEEP LEARNING, SEGMENTATION, ANNOTATION, AUGMENTATION, DATASET, PAINT.NET.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ГНМ - Глибока нейрона мережа (DNN)

ЗНМ - Згорткова нейрона мережа (CNN)

ПЗ - Програмне забезпечення

ResNet - Залишкова мережа

ReLU - Випрямлений Лінійний Блок

CVAT (Computer Vision Annotation Tool) - Інструмент для Анотації Даних Комп'ютерного Зору

TensorFlow Datasets (TFDS) - бібліотека, яка надає готові до використання набори даних

SMOTE- Техніка Синтетичного Надмірного Семплірування Меншості

IoU (Intersection over Union) - Перетин відносно Об'єднання

FAIR - наукові дані які відкриті і доступні для всіх охочих.

ВСТУП

Актуальність. Стрімкий розвиток технологій глибокого навчання зумовив суттєве зростання потреби у великих, якісних, структурованих та репрезентативних наборах даних. Саме дані формують основу навчання глибоких нейронних мереж, визначають точність, стійкість і узагальнювальну здатність моделей. У сучасних умовах, коли глибоке навчання застосовується у таких критичних сферах, як комп'ютерний зір, медична діагностика, робототехніка, автономні системи та аналітика великих даних, проблема формування правильно підібраних і збалансованих наборів даних стає ключовою.

Актуальність теми дослідження зумовлена тим, що саме процес створення навчальних наборів даних є найбільш трудомістким, ресурсоємним та водночас визначальним етапом розроблення моделей глибокого навчання. Помилки, допущені під час збору, очищення, сегментації, анотації чи структурування даних, безпосередньо впливають на якість навченої моделі та її здатність працювати в реальних умовах. Незбалансованість вибірки, наявність шумів, дублікати, некоректні мітки або недостатня різноманітність прикладів можуть призвести до перенавчання, зміщення моделі та некоректних прогнозів.

Сучасні дослідження з формування наборів даних демонструють важливість алгоритмізації цього процесу. Автоматизовані методи очищення даних, алгоритми сегментації та анотації зображень, техніки аугментації, синтетична генерація даних, а також стандартизовані підходи до структурування датасетів дозволяють значно підвищити якість моделей глибокого навчання. Водночас розвиток інструментів для створення синтетичних даних (Unity Perception, NVIDIA Omniverse) та сучасних засобів анотації (CVAT, LabelMe, Label Studio) потребує дослідження їх ефективності у реальних задачах.

Утім, незважаючи на широкий спектр доступних рішень, питання побудови універсальних алгоритмів формування навчальних наборів даних залишається недостатньо дослідженим. У більшості випадків цей процес залежить від конкретної предметної області, особливостей даних, вимог до точності моделі та обмежень обчислювальних ресурсів. Саме тому необхідність комплексного аналізу, систематизації та розроблення алгоритмів формування наборів даних є надзвичайно актуальною.

Додатковою складністю є варіативність типів даних - від реальних зображень до синтетично згенерованих сцен. Для задач комп'ютерного зору важливими є процеси сегментації, анотації та очищення зображень, що потребують використання спеціалізованих алгоритмів. Складність моделей глибокого навчання висуває нові вимоги до якості, структурованості та повноти наборів даних.

Метою дослідження є розроблення алгоритмів формування навчальних наборів даних для глибоких нейронних мереж із урахуванням вимог до їх якості, узагальнювальної здатності в сучасних інформаційних системах.

Об'єкт дослідження: процеси формування та підготовки навчальних наборів даних для глибоких нейронних мереж.

Предмет дослідження: методи, алгоритми та програмні засоби формування, структуризації, анотації та аугментації навчальних наборів даних.

Завдання дослідження:

1. здійснити аналіз джерел формування навчальних наборів даних а також оцінити їх репрезентативність і збалансованість;

2. розробити структуровану методику формування датасетів, що охоплює попередню обробку, сегментацію, анотацію, аугментацію та генерацію синтетичних даних;

3. розробити алгоритми інтеграції методів сегментації, анотації та аугментації та їх реалізувати у програмному середовищі Paint.NET;

4. провести експериментальне дослідження ефективності алгоритмів формування навчального набору даних.

Наукова новизна отриманих результатів полягає у вдосконаленні алгоритмів формування навчальних наборів даних з метою покращення якості глибоких моделей.

Практичне значення роботи полягає у формуванні датасету імуногістохімічних зображень для задач сегментації зображень у програмному середовищі Paint.NET.

Апробація роботи. Основні результати дослідження були представлені на науково-практичних конференції, опубліковано тези доповіді [1,2]. Кваліфікаційна робота написана згідно методичних рекомендацій [3,4]. Магістерська робота написана в рамках науковому напряму дослідницької групи «комп'ютерні системи штучного інтелекту» під керівництвом професора Березького О.М. [5-14].

Кваліфікаційна робота складається з трьох розділів, висновків, списку використаних джерел та додатків.

Перший розділ, проведено аналіз джерел формування наборів даних та програмних засобів їх підготовки.

Другий розділ, присвячений алгоритмам формування навчальних наборів даних, включаючи попередню обробку, сегментацію, анотацію та генерацію синтетичних зображень.

У третьому розділі виконано практичну частину - формування датасету та дослідження алгоритмів у програмному середовищі Paint.NET.

1 АНАЛІЗ ДЖЕРЕЛ ФОРМУВАННЯ НАВЧАЛЬНИХ НАБОРІВ ДАНИХ

1.1 Поняття глибоких нейронних мереж

Глибокі нейронні мережі - це різновид систем машинного навчання, головна особливість якого полягає у значній структурній глибині: наявності багатьох прихованих шарів між входом та виходом. Ця глибина є критичною, оскільки вона забезпечує якісну трансформацію, дозволяючи мережі вивчати складні ієрархії абстракцій і реалізовувати представницьке навчання. Фактично, ГНМ є технічною основою для всієї дисципліни глибокого навчання [15].

На відміну від неглибоких мереж ГНМ автоматично виявляють оптимальні ознаки для вирішення конкретної задачі.

Глибина мережі дозволяє перетворити високо-розмірні дані (наприклад, мільйони пікселів) на щільний, низько-розмірний простір ознак, де подібні об'єкти згуртовуються, а класи стають легко роздільними.

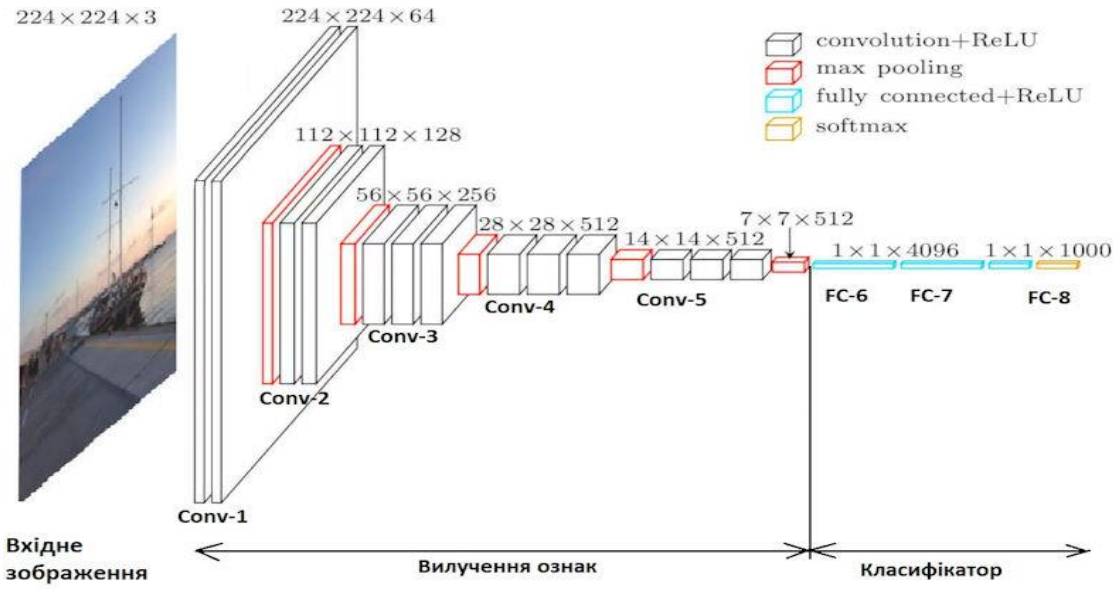

Рисунок 1.1 - Архутектура ЗНМ

Згорткова нейронна мережа (ЗНМ) є одним із найефективніших типів глибоких нейронних мереж, що широко використовується для обробки

візуальної інформації - зокрема зображень, відео, медичних знімків та інших даних із просторовою структурою (рисунок 1.1). Її головна ідея полягає в автоматичному виділенні та узагальненні ознак об’єктів на зображенні, що дає змогу моделі самостійно «бачити» і навчатися розпізнавати складні патерни без ручного опису характеристик [16].

Архітектура ЗНМ складається з кількох послідовно з’єднаних шарів, кожен з яких виконує певну функцію у процесі перетворення даних. На вході мережа отримує багатовимірний тензор, який відображає структуру зображення - його висоту, ширину та кількість каналів (32×32×3 для кольорового RGB-зображення). Вхідний шар не змінює дані, а лише передає їх далі для обробки.

Основу мережі становлять згорткові шари, що виконують операцію згортки - вони проходять фільтрами (ядрами) невеликого розміру (3×3 або 5×5) по всьому зображенню, обчислюючи локальні комбінації пікселів. Кожен фільтр виявляє певну ознаку, таку як краї, кути або текстурні деталі. Результатом є карти ознак, які зберігають найважливішу інформацію про структуру зображення.

Після операції згортки зазвичай застосовується функція активації - найчастіше ReLU (Rectified Linear Unit), яка надає моделі нелінійності. Завдяки цьому мережа здатна розпізнавати складні залежності між пікселями, а не лише прості лінійні співвідношення.

Наступним важливим етапом є пулінг (Pooling) - процедура зменшення розмірності карт ознак, що дозволяє скоротити обчислювальні витрати та зменшити ризик перенавчання. Пулінг зберігає найважливіші характеристики сигналу, зменшуючи при цьому кількість параметрів. Найпоширенішими типами є max pooling (вибір максимального значення з області) та average pooling (усереднення значень).

Після кількох циклів згорткових та пулінгових шарів дані сплющуються у вектор та передаються до повнозв’язних шарів. На цьому етапі нейрони з’єднуються між собою, що дозволяє моделі узагальнювати отримані ознаки та здійснювати прийняття рішення.

Для покращення навчання часто використовуються додаткові елементи, такі як нормалізація пакетів (Batch Normalization) - для стабілізації навчального процесу, та Dropout, який випадково вимикає частину нейронів, запобігаючи перенавчанню.

Завершальним етапом є вихідний шар, який формує кінцевий прогноз. Для задач класифікації зазвичай використовується функція Softmax, що перетворює вихідні значення в ймовірності належності до певного класу, а для задач регресії - лінійна активація.

ЗНМ навчається у процесі послідовного узагальнення ознак: нижчі шари виявляють прості елементи (лінії, краї), середні - складніші структури (форми, частини об’єктів), а верхні - цілісні образи або об’єкти. Завдяки цій ієрархії ЗНМ демонструє надзвичайно високу ефективність у задачах комп’ютерного зору, таких як розпізнавання образів, детекція об’єктів, сегментація сцен та аналіз медичних зображень [17].

Відомо багато архітектур згорткових мереж одніє з них є, LeNet-5 стала першою реалізацією ЗНМ, AlexNet запровадила використання ReLU та GPU для пришвидшення навчання, VGG16 характеризується глибиною та одноманітними фільтрами 3×3, а ResNet завдяки skip-з’єднанням дала змогу створювати надзвичайно глибокі мережі без втрати якості. Більш сучасні моделі, такі як EfficientNet, оптимізують співвідношення глибини, ширини та роздільної здатності, забезпечуючи високу точність при меншій кількості параметрів.

Робота з ГНМ є циклічним процесом, який складається з трьох основних взаємопов’язаних етапів. Цей процес не завершується після одного проходу, а повторюється багаторазово, оскільки кожен наступний цикл дозволяє поступово вдосконалювати якість моделі, підвищуючи точність її прогнозів і здатність до узагальнення.

Першим і одним із найважливіших етапів є підготовка та формування даних. На цьому етапі здійснюється збір необхідної інформації з різних джерел, її перевірка, очищення від шумів, пропущених значень або некоректних

записів. Далі дані нормалізуються, тобто приводяться до однакових масштабів, щоб нейронна мережа могла ефективніше сприймати числові значення. Якщо у вибірці присутні текстові або категоріальні змінні, вони перетворюються у числовий формат, який зрозумілий алгоритмам машинного навчання. Важливо також правильно розділити дані на три підмножини: навчальну, валідаційну та тестову. Навчальна вибірка використовується для безпосереднього навчання моделі, валідаційна - для перевірки проміжних результатів і налаштування параметрів, а тестова - для остаточної оцінки якості після завершення навчання [18]. У деяких випадках, коли обсяг даних невеликий, застосовують аугментацію - штучне збільшення набору даних за рахунок модифікацій існуючих прикладів (обертання зображень, зміна яскравості чи шуму). Усе це сприяє підвищенню стійкості моделі та зменшує ризик перенавчання.

Другий етап, навчання моделі під час якого відбувається основна частина роботи з ГНМ. На цьому етапі визначається архітектура нейронної мережі, тобто скільки шарів вона матиме, який тип нейронів буде використано, які функції активації застосовуватимуться тощо. Далі відбувається процес навчання, у ході якого модель поступово підлаштовує свої вагові коефіцієнти. Це відбувається завдяки алгоритму зворотного поширення помилки, який дозволяє моделі вчитися на своїх помилках. Кожна ітерація полягає у тому, що модель робить прогноз, порівнює його з реальним значенням, обчислює помилку та коригує ваги, щоб у наступний раз помилка стала меншою. Для ефективнішої оптимізації використовуються спеціальні методи. Паралельно із навчанням проводиться валідація, що перевіряє якість поточного стану моделі на окремій частині даних, щоб запобігти перенавчанню та зберегти її здатність узагальнювати нову інформацію.

Третім етапом є оцінювання та вдосконалення моделі. Після того як навчання завершено, модель тестується на незалежних тестових даних, які не використовувались у процесі тренування. Це дозволяє об’єктивно оцінити її здатність працювати з новими прикладами. На цьому етапі розраховуються основні показники ефективності. Якщо результати не відповідають очікуванням

або модель демонструє нестабільну поведінку, проводиться її вдосконалення. Це може включати зміну гіперпараметрів (швидкості навчання, кількості епох, розміру пакета даних), модифікацію архітектури, використання нових методів регуляризації чи навіть збільшення обсягу або якості навчальних даних. Після внесення змін цикл роботи повторюється спочатку, а оновлені дані знову проходять етапи підготовки, навчання та перевірки.

1.2 Аналіз джерел формування наборів даних

Формування якісного набору даних є одним із ключових етапів у побудові та навчанні глибоких нейронних мереж, оскільки саме від даних залежить здатність моделі правильно відображати закономірності реального світу. Якість, повнота, різноманітність і достовірність даних безпосередньо впливають на кінцевий результат - точність прогнозування та стійкість нейронної мережі до нових, невідомих прикладів. Тому під час аналізу джерел даних важливо враховувати не лише кількість інформації, а й її походження, спосіб збору, структуру та актуальність. У сучасних дослідженнях розгляд цих характеристик стає ще більш критичним, адже якість навчальних наборів визначає не лише коректність моделі, а й можливість подальшого масштабування системи.

Джерела даних, які використовуються для навчання глибоких моделей, можна умовно поділити на кілька основних груп: реальні, синтетичні та комбіновані. Кожен із цих типів має свої переваги, недоліки та особливості використання. Проте їх ефективність залежить також від рівня репрезентативності, ступеня контролю над процесом формування та можливостей подальшої адаптації до потреб конкретного дослідження. Саме тому сучасний підхід до аналізу джерел даних включає не лише класифікацію, а й оцінку їх відповідності задачам глибокого навчання.

Одним із найпоширеніших варіантів є використання відкритих наборів даних, які доступні у вільному доступі в мережі Інтернет і часто публікуються дослідницькими установами або великими компаніями. Такі ресурси, як Kaggle, UCI Machine Learning Repository, Google Dataset Search, OpenML, ImageNet, COCO та інші, надають великі обсяги структурованих і добре підготовлених даних для різних типів задач - від класифікації зображень до аналізу тексту чи прогнозування часових рядів. Перевагою цих джерел є їхня перевіреність і наявність описів, які допомагають зрозуміти структуру та зміст вибірки. Проте головним недоліком відкритих даних є те, що вони часто є надто загальними й не завжди підходять під специфіку конкретної задачі, що може потребувати додаткової обробки або адаптації.

Іншим важливим джерелом є синтетичні дані, що формуються спеціально для певного проєкту чи дослідження. У цьому випадку дані можуть бути отримані шляхом експериментів, спостережень, вимірювань або за допомогою спеціалізованих сенсорів. Такий підхід дозволяє створити максимально релевантний набір даних, що точно відповідає вимогам конкретної задачі. Наприклад, для навчання моделі розпізнавання облич можна самостійно створити вибірку фотографій у потрібних умовах освітлення чи ракурсу. Однак головним недоліком власних даних є складність та тривалість процесу їх збору, а також необхідність ретельної перевірки на повноту, збалансованість і відсутність систематичних похибок [25]. Цей розрив призводить до зниження якості узагальнення моделі на реальних даних. Для вирішення цієї проблеми застосовують техніки доменної адаптації, фізично коректний рендеринг та доменну рандомізацію, що дозволяє моделі краще враховувати реальні коливання ознак.

Третю групу становлять комбіновані джерела даних, коли дослідник поєднує кілька типів інформації з різних джерел. Такий підхід дозволяє значно розширити обсяг вибірки, зробити її більш різноманітною та універсальною. Наприклад, частину даних можна взяти з відкритих наборів, а іншу - доповнити власними вимірюваннями чи синтетично згенерованими прикладами. У сфері

комп’ютерного зору часто використовують генерацію штучних зображень за допомогою спеціальних інструментів або нейронних мереж, що дозволяє моделі тренуватись на великій кількості варіацій без необхідності фізично збирати такі дані

.

Таблиця 1.1 -Переваги та недоліки джерел формування наборів даних

Джерела формування наборів даних	Переваги	Недоліки
Реальні дані:	Швидке накопичення великих обсягів інформації. Збір даних у чітко визначених умовах.	Велика собівартість та часозатратність. Етичні та правові ризики. Зниження надійності у критичних або незвичайних ситуаціях.
Синтетичні дані:	Повний контроль над вмістом та метаданими. Відсутність конфіденційності. Можливість генерування необмежену кількість зразків	Неможливість повністю відтворити реалій світу. Втрата точності синтетичних даних при застосуванні до реальних даних

При виборі джерел даних важливо враховувати низку критеріїв, серед яких - достовірність, повнота, репрезентативність, відсутність упередженості та актуальність (таблиця 1.1). Достовірність означає, що дані повинні відображати реальні явища або об’єкти без помилок чи фальсифікацій. Повнота передбачає достатню кількість прикладів для кожного класу чи категорії, щоб уникнути дисбалансу, який може призвести до зміщення результатів. Репрезентативність гарантує, що вибірка охоплює всі можливі варіанти і є типовою для досліджуваної області. Важливим аспектом також є перевірка на наявність

дублікатів, шумів і пропусків, оскільки такі дефекти можуть значно вплинути на процес навчання. Також особливу увагу слід приділяти питанням етики та захисту даних, адже у багатьох випадках набори містять персональну або конфіденційну інформацію. Використання таких даних вимагає дотримання законодавчих норм, зокрема щодо знеособлення інформації, отримання дозволів і збереження конфіденційності користувачів.

Таким чином, комплексний аналіз джерел даних включає не лише класифікацію, а й оцінку їхньої структури, методів збору, ризиків, обмежень та відповідності вимогам конкретної задачі глибокого навчання. Поєднання різних типів джерел, їх ретельна валідація та контроль якості є ключовими чинниками успішної побудови ефективних моделей.

1.3 Програмні засоби формування наборів даних

Формування якісних наборів даних є фундаментальним етапом у процесі побудови та навчання глибоких нейронних мереж. Незалежно від обраної архітектурної моделі, саме тому почався рівень її загальної здатності та продуктивності при вирішенні практичних завдань. Тому в сучасних дослідженнях суттєво важливо використовувати розробці інструментів, які не можуть лише збирати та структурувати дані, але й автоматизувати їхню підготовку, забезпечувати збалансованість вибору, а також створювати синтетичні приклади у випадках, коли реальні дані є недоступними або їхня кількість обмежена [19].

У сучасних системах штучного інтелекту програмні засоби формування даних відіграють роль повноцінного технологічного шару, який забезпечує якість, структурованість та відтворюваність експериментів. Робота з даними більше не обмежується завантаженням та збереженням - вона включає контроль версій, управління метаданими, моніторинг якості вибірки, автоматизацію

збору та анотації. Саме тому вся екосистема інструментів поділяється на кілька класів: бібліотеки фреймворків, сервіси анотації, платформи для аугментації та рушії для синтетичної генерації.

Програмні засоби формування наборів даних можна умовно поділити на декілька категорій: універсальні бібліотеки для роботи з даними у складі фреймворків глибокого навчання, спеціалізовані інструменти для анотування та маркування інформації, бібліотеки для аугментації даних, а також системи для генерації синтетичних зображень.

Кожна група інструментів також відповідає певному етапу життєвого циклу даних. Наприклад, фреймворк-бібліотеки відповідають за читання, потокову обробку та інтеграцію даних у нейронні мережі. Інструменти анотації - за створення навчальних міток для задач комп'ютерного зору. Платформи аугментації забезпечують приріст варіативності вибірки. А синтетичні рушії дозволяють створювати дані там, де реальні зібрати неможливо або надто дорого. У сукупності всі ці компоненти формують цілісний pipeline, що дозволяє створювати великі, чисті та збалансовані набори даних.

Одним із найрозширеніших інструментів є TensorFlow Datasets (TFDS) - це бібліотека з відкритим кодом, яка входить до екосистеми TensorFlow. Вона містить десятки готових наборів даних з різних напрямків - із зображень класифікації завдань до аналізу тексту та обробки звукових сигналів. Головна перевага TFDS виникає в тому, що дослідник може миттєво завантажити вже структуровані дані та розпочати навчання моделі без необхідності самостійної підготовки. Крім того, бібліотека дозволяє створювати власні набори, використовуючи єдиний уніфікований формат. Це значно зменшує витрати на підготовку експериментів і робить наукові дослідження більш відтворюваними. У роботі з великими наборами даних TFDS також відіграє важливу роль у стандартизації формату, що дозволяє зменшити кількість помилок у підготовці даних. Завдяки підтримці кешування та автоматичної валідації структура наборів залишається стабільною, що важливо для відтворюваності експериментів у наукових дослідженнях.

Не менш популярним інструментом є PyTorch Dataset та DataLoader, які входять до складу фреймворка PyTorch. Вони забезпечують зручний механізм завантаження даних, їх попередньої обробки та передачі в модель під час навчання. DataLoader дозволяє ефективно працювати з великими наборами, розподіляючи їх на біти та здійснюючи паралельну обробку. Дослідники можуть легко реалізувати власні класи для читання певних форматів, наприклад медичних знімків чи часових рядів, що робить PyTorch дуже гнучким у наукових експериментах.

Одним із перших популярних рішень став Keras ImageDataGenerator, що дає можливість автоматично зберігати різноманітні перетворення зображення: повороти, віддзеркалення, масштабування, зсув та інші модифікації. Завдяки цьому можна значно збільшити обсяг вибору, зберігаючи при цій різноманітності прикладів. Подібні підходи допомагають уникати проблеми перенавчання нейронних мереж, коли модель сильно піддається реалізується під обмежену кількість даних і втрачає здатність узагальнювати нову інформацію.

Серед сучасних бібліотек варто мати Albumentations - високопродуктивний інструмент для збільшення зображення, який завдяки оптимізованим алгоритмам працює швидше, ніж великих аналогів. Albummentations підтримує велику кількість складних перетворень, виключно зі зміною кольорових характеристик, накладанням шуму, перспективними трансформаціями та ін. Це робить бібліотеку універсальною у програмуванні - від простих класифікаційних завдань до сегментації об’єктів у високороздільних зображеннях. У порівнянні з іншими інструментами Albumentations забезпечує найкраще співвідношення швидкості та складності підтримуваних трансформацій. Це дає змогу ефективно готувати навіть багатомільйонні набори зображень у промислових ML-проєктах. Бібліотека також підтримує механізми інтеграції з PyTorch та TensorFlow, що робить її універсальною незалежно від обраного фреймворку.

Важливе місце в процесі формування наборів даних займають інструменти для розмітки та анотації зображення. У кожному виконанні комп'ютерного зору потрібна наявність міток (наприклад, координат об'єктів або масок сегментації), виникає потреба у використанні спеціалізованих програм. Одним із найвідоміших рішень є LabelImg - проста у використанні утиліта для маркування об'єктів на зображеннях, яка підтримує популярні формати, зокрема Pascal VOC та YOLO. Для більш масштабних проектів застосовуються системи на кшталт CVAT (Computer Vision Annotation Tool), розроблені компанією Intel, та Supervisely - хмарна платформа, яка забезпечує зручний колективний доступ до процесу розмітки (таблиця 1.2). Використання таких інструментів дозволяє створювати великі набори даних для завдань детекції та сегментації об'єктів, що є основою для розробки сучасних систем автономного транспорту чи системи відеоспостереження [18].

Таблиця 1.2 - Розширене порівняння програмних засобів формавання наборів даних.

Інструмент	Тип задач	Підтримувані формати	Складність	Автоматизація	Масштабованість	Інтеграція з ML	Командна робота
Labellmg	Детекція	VOC, YOLO	Низька	Немає	Низька	Частково	Немає
CVAT	Детекція, сегментація	JSON, XML, COCO	Середня	Присутня	Висока	Висока	Присутня
Supervisely	Всі типи CV задач	JSON, COCO, власні	Середня	Висока	Висока	Висока	Присутня
Albumentations	Аугментація	Будь-які зображення	Низька	Часткова	Висока	Повна	Немає

Продовження таблиці 1.2

ImageData Geytrator	Аугмента-ція	Зображення	Низька	Часткова	Середня	Повна	Немає
Unity Perception	Синтетика	PNG, EXR, JSON	Середня	Висока	Висока	Висока	Немає
NVIDIA Omniverse	Синтетика	Різні 3D та 2D	Складна	Висока	Дуже висока	Висока	Часткова
Blender	Синтетика	Усі базові 2D/3D	Середня	Середня	Висока	Середня	Немає

Розмічувальні системи відрізняються за рівнем автоматизації: від повністю ручної анотації (LabelImg) до напівавтоматичних інструментів, що використовують моделі попереднього навчання для прискорення маркування (CVAT, Supervisely). У великих проєктах команда анотації може складатися з десятків людей, тому важливими є контроль якості, журналювання, можливість відкату змін та аудит міток - усі ці функції підтримують сучасні хмарні платформи.

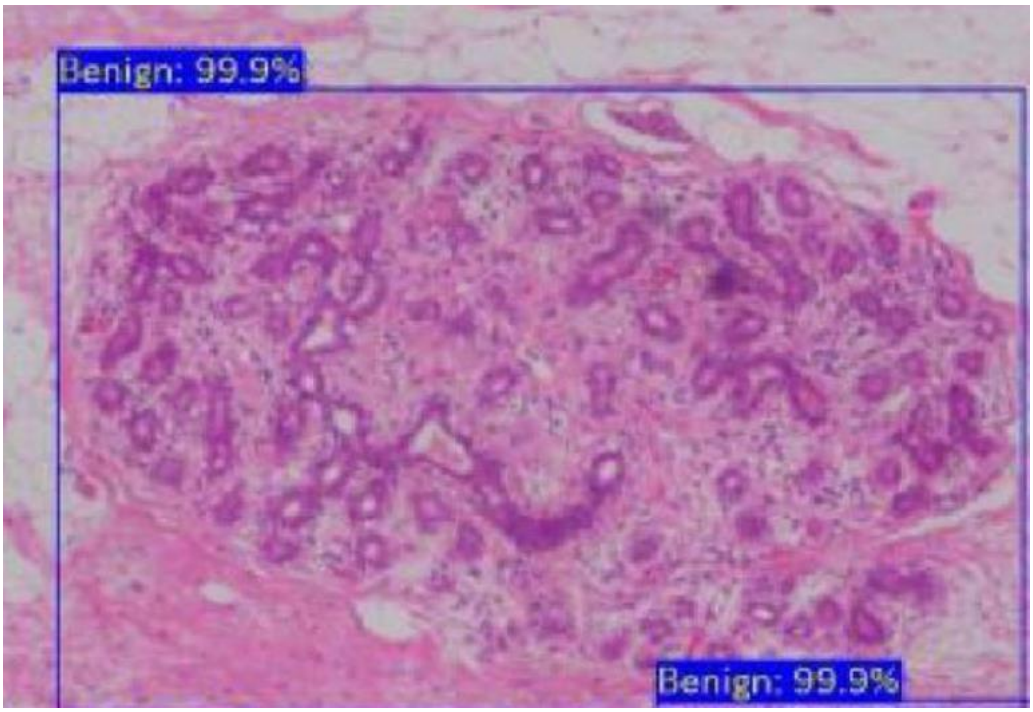

Рисунок 1.2 - Робота ручної анотації LabelImg

У цьому зображені LabelImg виступає як мінімалістичний, але достатньо гнучкий інструмент для побудови базової розмітки «область пухлини / фон». Лікар працює з окремими обрізками (patches) або зменшеними WSI, послідовно відмічаючи прямокутниками вогнища карциноми.Це демонструє, що інструмент сам по собі не гарантує якості; критичною є домовленість про протокол розмітки і міжекспертна узгодженість (рисунок 1.2).

З точки зору формування набору даних, використання LabelImg чітко задає структуру: кожне зображення має список bounding box'ів із класами (наприклад, "invasive carcinoma", "benign/normal"), що добре підходить для задач детекції та класифікації на рівні вогнищ. Недолік підходу - обмежена підтримка складних масок і масштабних WSI: інструмент оптимізований під 2D-зображення помірного розміру, а при роботі з гігапіксельними слайдами дослідники змушені попередньо різати їх на тайли [35].

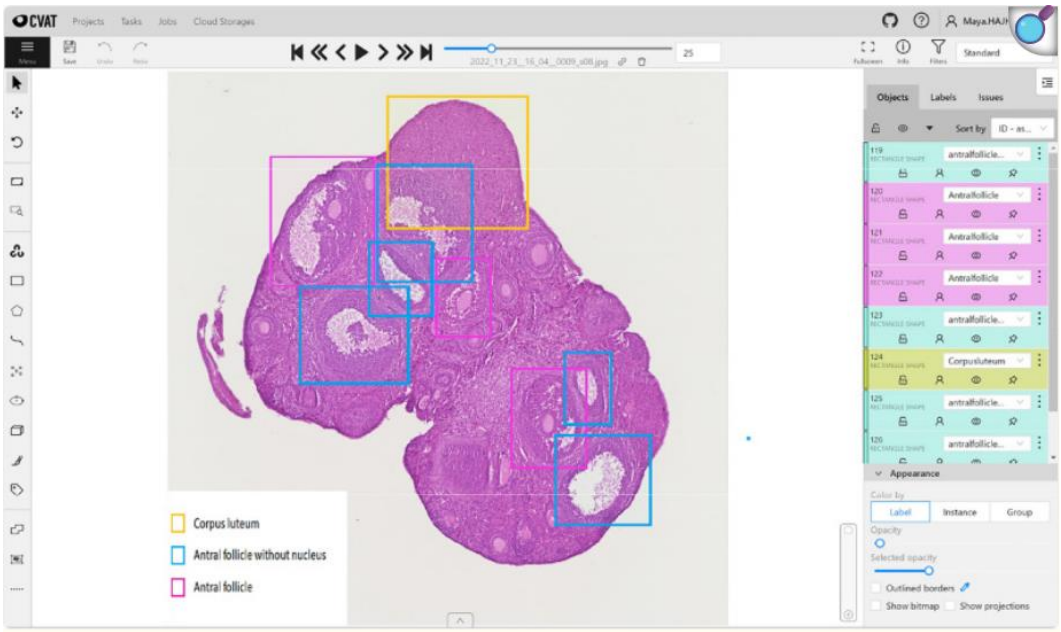

Рисунок 1.3 - Приклад роботи у CVAT

Робота з автоматичного виявлення овариальних фолікулів на гістологічних зрізах, де було виконано супервізовану детекцію об'єктів на H&E-забарвлених зрізах яєчника. Для створення навчальної вибірки зображення WSI спершу обрізають у тайли, а потім у CVAT проводять розмітку кожного тайлу:

створюють bounding box’и та, за потреби, полігони для різних типів фолікулів (з ядром, без ядра) і структур, таких як corpus luteum (рисунок 1.3) [36].

З погляду формування датасету, робота демонструє декілька важливих переваг CVAT:

1. він підтримує роботу з великими наборами тайлів, організованих у проєкти й завдання;
2. дозволяє централізовано керувати списком класів (типи фолікулів, інші структури);
3. зберігає анотації у форматах, сумісних із популярними фреймворками (COCO, YOLO тощо), що спрощує подальший експорт даних у пайплайн глибокого навчання.

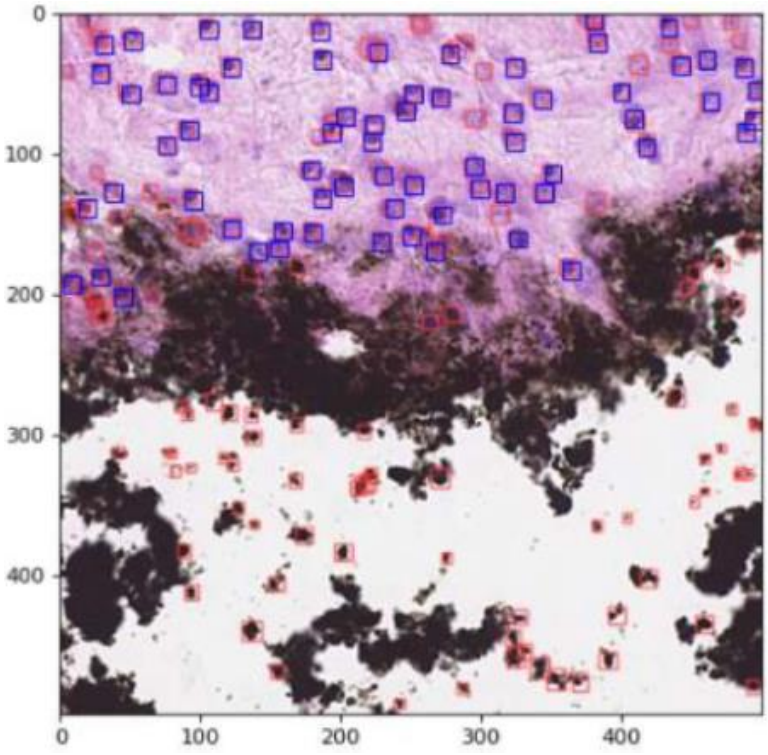

Рисунок 1.4 - Аугументація гістологічного забраження за допомогою YOLO

Особливість Albumentations у цьому зображені - підтримка одночасної аугментації зображення та відповідної маски або координат об’єктів, тобто забезпечується узгодженість між зміненим зображенням і ground truth (рисунок 1.4). Це важливо для задач сегментації ядер, де навіть невелика невідповідність

між маскою та зображенням може зіпсувати навчання. Автори використовують бібліотеку як модуль у пайплайні PyTorch та TensorFlow, налаштовуючи ланцюжок перетворень, що зберігають морфологічну інформативність ядер, але збільшують варіативність зовнішнього вигляду.

З точки зору формування набору даних, програма демонструє перехід від «статичного» датасету до «динамічного», де реальний обсяг тренувальних прикладів визначається не кількістю вихідних зображень, а комбінацією аугментацій. Це особливо актуально для гістології, де отримання великої кількості розмічених зображень дорого й трудомістко, а відмінності у протоколах фарбування, сканерах і лабораторіях призводять до сильних зсувів розподілу даних. Albumentations дає спосіб моделювати частину цих варіацій на етапі формування тренувальних батчів [37].

Таблиця 1.3 - Порівняння програмних засобів формування наборів даних

Генерація синтетичних даних	Основі ознаки та можливості	Переваги та застосування
Unity Perception	Створення 3D-сцен Автоматична генерація даних з розміткою	Використовується якщо реальні дані дорогі або недоступні
NVIDIA Omniverse	Реалістичний рендеринг Складні сценарії	Використовується для досліджень і тренування моделей з високими вимогами
Blender	Генерація зображень та анімацій підтримка різних форматів	Гнучкий та доступний інструмент для генерації синтетичних даних

Окрему групу становлять програмні засоби для генерації синтетичних даних, які останніми роками набули особливої актуальності. Їх застосування виправдане в тих випадках, коли збір реальних даних є дорогим, трудомістким

або навіть неможливим. Прикладом таких інструментів є Unity Perception - модуль у середовищі Unity, який дозволяє створювати тривимірні сцени та автоматично генерувати дані з розміткою. Подібні можливості пропонують NVIDIA Omniverse та популярний редактор Blender, які підтримують фізично коректний рендеринг та створення реалістичних сценаріїв (таблиця 1.3). Завдяки синтетичним даним можна отримати майже необмежену кількість прикладів із різноманітними варіаціями умов освітлення, положення об’єктів та шумів [33].

У багатьох випадках синтетичні дані дозволяють імітувати рідкісні або небезпечні події (аварії, відмови обладнання, екстремальні погодні умови), які неможливо отримати в реальності. Такі інструменти також забезпечують повну автоматичну розмітку, що значно скорочує витрати на створення навчальних наборів.

1.4 Висновки до розділу 1

Формування навчальних наборів даних є багатокомпонентним процесом, у якому важливу роль відіграє правильний вибір джерел інформації та способів їхньої підготовки. Реальні та синтетичні дані мають різні властивості, що визначають їх придатність для конкретних задач: реальні зображення забезпечують високу достовірність, тоді як синтетичні дозволяють отримувати контрольовані та варіативні приклади. Якість навчальної вибірки значною мірою залежить від репрезентативності, відсутності упередженості, різноманітності сцен і точності метаданих. Важливим аспектом є також використання спеціалізованих інструментів для анотації, структуризації та збереження даних, оскільки вони забезпечують системність та відтворюваність під час побудови моделей глибокого навчання.

2 АЛГОРИТМИ ФОРМУВАННЯ НАБОРІВ ДАНИХ

2.1 Основні вимоги до навчальних наборів даних для глибокого навчання

У процесі розроблення та навчання глибоких нейронних мереж надзвичайно важливу роль відіграє якість навчальних наборів даних. Саме від повноти, репрезентативності та структурованості даних залежить здатність моделі ефективно засвоювати закономірності та узагальнювати отримані знання на нові, раніше невідомі приклади. Формування якісного й збалансованого набору даних є ключовою передумовою досягнення високої точності, стійкості та узагальнювальної здатності глибоких моделей [20]. Недостатня різноманітність, наявність шумів або помилкових міток у даних можуть призвести до перенавчання або втрати здатності моделі робити коректні прогнози. Тому визначення основних вимог до навчальних наборів даних є важливим етапом, що забезпечує надійність і ефективність процесу глибокого навчання.

Формально залежність ефективності моделі від якості даних можна подати через функцію втрат, яка визначає різницю між прогнозованим результатом і реальним значенням:

$$L = f(x_i, y_i, \theta)$$

де:

L - значення функції витрат;

x_i - вхідні дані;

y_i - цільові мітки;

θ - параметри (ваги) моделі.

Обсяг і різноманітність навчальних даних є одними з ключових факторів, що визначають ефективність і стійкість глибоких нейронних мереж. У процесі

навчання велика кількість прикладів дозволяє моделі більш точно оцінювати параметри, формувати складні представлення ознак і знижувати ризик перенавчання. Недостатній обсяг даних призводить до того, що модель може добре запам’ятовувати конкретні приклади з навчальної вибірки, але втрачає здатність робити точні прогнози на нових, раніше невідомих даних. Водночас різноманітність прикладів відіграє критичну роль у формуванні стійкої узагальнювальної здатності. Вона включає варіації об’єктів, умов зйомки або запису, фонів, освітлення, руху та навіть характеристик сенсорних пристроїв, що дозволяє моделі навчатися розпізнавати суттєві ознаки незалежно від зовнішніх змін і шуму [21].

У сфері комп’ютерного зору роль різноманітності даних особливо помітна: наприклад, для задачі класифікації об’єктів необхідно, щоб модель бачила об’єкти у різних ракурсах, з різними розмірами, на різних фонах та при змінному освітленні. Відсутність таких варіацій може призвести до того, що мережа буде добре впізнавати об’єкти лише в обмежених умовах, але помилково класифікуватиме ті ж об’єкти у нових ситуаціях. Аналогічно, у розпізнаванні мовлення велике значення мають записи різних мовців, з різними тембрами голосу, швидкістю мовлення та фоновими шумами. Мережа, навчена лише на стандартних записах у контрольованих умовах, буде недостатньо стійкою при роботі з живими голосами у реальному середовищі, де фонова акустика, шум та індивідуальні особливості мовлення суттєво варіюють. У біомедичних зображеннях ситуація стає ще складнішою: медичні знімки можуть отримуватися різними типами апаратів, при різних протоколах сканування та в умовах мінливого фізіологічного стану пацієнтів. Якщо навчальний набір даних не відображає цих варіацій, мережа може демонструвати високу точність на тестових зображеннях, отриманих у тих же умовах, але істотно помилятися на даних з інших клінік або від інших сенсорів.

Таким чином, обсяг та різноманітність навчальних даних формують основу для розвитку узагальнювальної здатності глибоких нейронних мереж. Чим більше репрезентативних та варіативних прикладів доступно під час

тренування, тим більш надійною, стійкою до змін зовнішніх умов і здатною до коректних прогнозів у реальному середовищі стає модель. Це підкреслює необхідність системного підходу до формування наборів даних, де кількість і якість прикладів взаємопов’язані та взаємно підсилюють ефект навчання.

Таблиця 2.1 - Порівняння вимог до даних у різних сферах

Сфери застосування	Тип даних	Джерело вірацій	Мета розширення
Комп’ютерний зір	Зображення, відео	Освітлення, фон, кут огляду, тип камери	Підвищення точності розпізнавання об’єктів у реальних умовах
Розпізнавання мовлення	Аудіосигнали	Акценти, інтонація, якість мікрофону	Підвищення стійкості до шумів і різних голосів
Біометричні зображення	Мікроскопічні або рентгенівські знімки	Тип тканини, параметри сканера, колірне забарвлення	Узагальнення між лабораторіями й апаратними системами

Дисбаланс класів у навчальних наборах даних - це ситуація, коли одна категорія (клас) представлена великою кількістю прикладів, а інші класи - значно меншою. У практичних завданнях глибокого навчання це трапляється дуже часто. Наприклад, у медичних дослідженнях переважають нормальні зразки, тоді як патологічні зустрічаються рідко. У фінансових системах більшість транзакцій є легітимними, а випадки шахрайства складають лише малу частку. У задачах виявлення дефектів на виробництві нормальні об’єкти значно переважають дефектні.

Щоб компенсувати дисбаланс, у функції втрат вводять ваги класів:

$$L = -\sum_{i=1}^{N} w_{y_i} \log p(y_i|x_i)$$

де:

L - загальна функція втрат (Loss);

N - кількість прикладів у наборі даних;

x_i - вхідні дані (ознаки або зображення);

y_i - правильна мітка класу для прикладу;

$p(y_i \mid x_i)$ - ймовірність, яку модель присвоює правильному класу;

w_{y_i} - ваговий коефіцієнт для балансування класів (використовується при дисбалансі);

знак “–” гарантує, що функція втрат набуває менших значень при вищій ймовірності правильного класу.

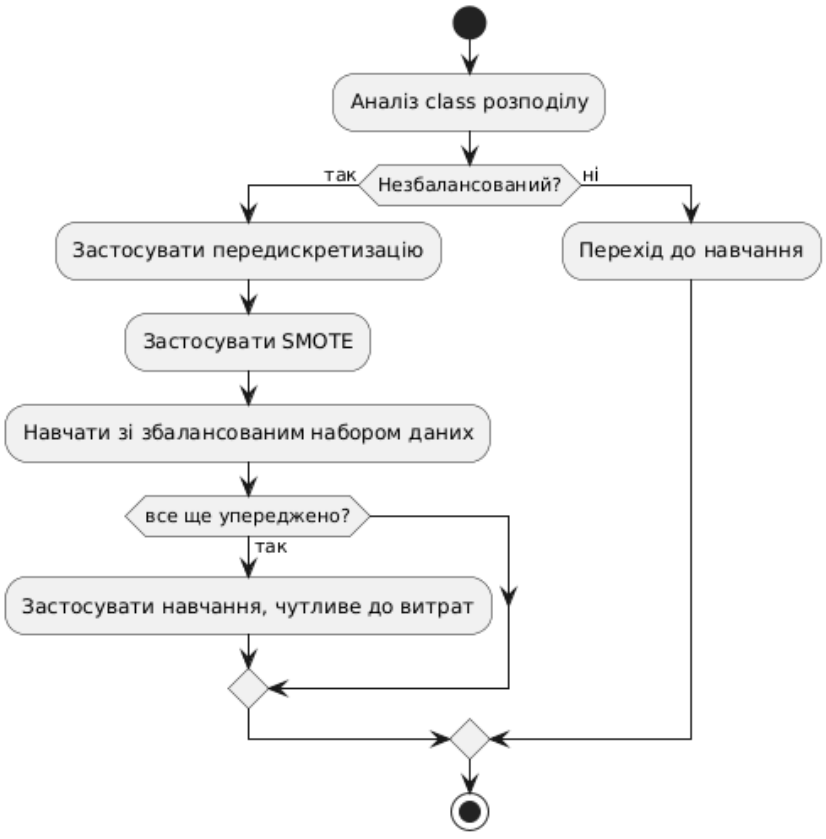

Рисунок 2.1 - Діаграма робочого процесу балансування класів

Для моделей глибокого навчання дисбаланс класів є критичною проблемою. Нейронні мережі, оптимізуючи функцію втрат, прагнуть мінімізувати загальну помилку на всіх даних. У результаті вони переважно «вчаться» на прикладах домінуючого класу, практично ігноруючи рідкісні категорії. Це призводить до того, що модель демонструє високу точність у передбаченні частого класу, але майже не здатна правильно класифікувати рідкісні приклади.

Наслідками дисбалансу класів можна вважати такі пункти як:

1) Зміщення моделі у бік більш представленого класу, якщо один клас складає більшість даних, модель адаптується до нього. Це означає, що вона починає «бачити» переважно закономірності домінуючого класу і нехтує специфікою рідкісних прикладів.

2) Упередженість моделі, якщо навчена на дисбалансованому наборі даних мережа має тенденцію завжди передбачати більш частий клас, модель може ігнорувати патології, що призводить до потенційно критичних помилок.

3)Зниження точності для рідкісних класів, тобто стандартні метрики точності втрачають сенс у дисбалансованих наборах, оскільки велика кількість правильних передбачень для домінуючого класу маскує невдачі на рідкісних класах. Це особливо важливо у медичних, фінансових та безпекових застосуваннях, де помилка рідкісного класу може мати серйозні наслідки.

До методів боротьби з дисбалансом можна вважати такі підходи як:

1) Oversampling - збільшення рідкісних класів, метод полягає у збільшенні кількості прикладів рідкісного класу шляхом їх дублювання або створення варіацій існуючих прикладів.

Перевагами даного підходу є проста реалізація, що дозволяє використовувати існуючі дані, а недоліками - ризик перенавчання, оскільки мережа бачить одні й ті ж приклади багато разів.

2) Undersampling - зменшення домінуючого класу, підхід полягає у випадковому видаленні прикладів з більш представленого класу, щоб вирівняти кількість прикладів.

Переваги підходу дозволяє ефективно зменшувати упередженість і прискорює навчання, недоліки - втрачається інформація, що може погіршити здатність моделі загально узагальнювати.

3) SMOTE (Synthetic Minority Over-sampling Technique), метод штучного збільшення рідкісного класу шляхом генерації нових синтетичних прикладів. Нові приклади створюються на основі інтерполяції між існуючими зразками рідкісного класу.

Перевага - зменшує ризик перенавчання, оскільки додає різноманітні приклади, а недолік - можливість створення неприродних або невластивих даних.

4) Cost-sensitive learning - навчання з урахуванням вартості помилки, тобто у цьому підході помилки рідкісного класу важать більше при обчисленні функції втрат, що змушує модель звертати більше уваги на рідкісні приклади, не змінюючи самі дані.

Переваги: не вимагає збільшення або скорочення набору даних, дозволяє адаптувати модель до реальної важливості помилок.

Недоліки: потребує правильного налаштування ваг, інакше може виникнути перенавчання або навпаки недостатня увага до рідкісних класів.

Для прикладу дисбалансу класів розглянемо задачу автоматичного аналізу гістологічних зображень тканин для виявлення патологій. У більшості наборів даних нормальна тканина зустрічається значно частіше, ніж патологічна - інколи в 20–30 разів. Якщо застосовувати стандартне навчання без балансування, модель майже завжди буде класифікувати зображення як нормальні, ігноруючи патологічні ділянки. Це робить її непридатною для практичного застосування, адже в медичній сфері важлива точна і своєчасна ідентифікація патологій. Використання методів oversampling, SMOTE або cost-sensitive learning дозволяє вирівняти вплив класів і підвищити точність класифікації рідкісних патологічних випадків [30].

Якість та узгодженість анотацій у навчальних наборах даних є визначальними чинниками достовірності та ефективності глибокого навчання.

Коректна розмітка забезпечує правильне співвідношення між вхідними прикладами та їхніми цільовими мітками, що безпосередньо впливає на формування функції втрат (loss) і стабільність навчального процесу. Одним із ключових показників є міжекспертна узгодженість (inter-annotator agreement) - ступінь згоди між різними анотаторами щодо присвоєних міток. Для її кількісної оцінки використовують статистичні показники, зокрема коефіцієнт Каппи (Cohen's Kappa), який враховує випадкову згоду, або коефіцієнт Флейсса (Fleiss' Kappa) для більш ніж двох експертів [22].

Типові проблеми розмітки включають неоднозначність інтерпретації об'єктів, появу шуму через людські помилки, різний рівень досвіду анотаторів та варіативність критеріїв оцінки. Для підвищення якості застосовують методи перевірки, зокрема подвійну розмітку (коли один приклад маркують кілька експертів), контрольні вибірки із перевіреними мітками та валідаційні тести для оцінки узгодженості.

Метадані відіграють ключову роль у сучасних наборах даних для глибокого навчання, оскільки вони описують контекст, походження та умови отримання даних, забезпечуючи їх правильне тлумачення та повторне використання. Наявність детальних метаданих робить набір даних більш структурованим, прозорим і корисним для подальших досліджень. У цьому контексті важливими є принципи FAIR - Findable, Accessible, Interoperable, Reusable, що означають: знахідність (дані повинні бути ідентифікованими через унікальні DOI або індекси), доступність (чітко визначені умови доступу), сумісність (використання узгоджених форматів і стандартів) та повторне використання (наявність ліцензій і повного опису)[23].

Приклад структури метаданих для біомедичного набору зображень може включати поля: автор, ліцензія, DOI, опис, змінні (тип тканини, метод сканування, розмір зображення), формат (наприклад, TIFF або DICOM).

Отже, стандартизовані метадані є фундаментом для відтворюваності наукових експериментів, оскільки вони забезпечують можливість перевірки,

порівняння та повторного використання результатів у незалежних дослідженнях.

2.2 Алгоритми попередньої обробки зображень

Попередня обробка зображень є важливим етапом формування якісних наборів даних для глибокого навчання, оскільки саме на цьому рівні забезпечується коректність і узгодженість вхідної інформації. Вона спрямована на усунення шумів, артефактів та варіацій у яскравості чи кольоровій гамі, що можуть негативно впливати на результати навчання моделей. Завдяки попередній обробці досягається стандартизація зображень, що сприяє підвищенню стабільності та швидкості збіжності нейронних мереж під час тренування. У цьому підрозділі розглядаються основні алгоритми попередньої обробки, зокрема методи нормалізації кольору, масштабування, обрізання зображень та видалення артефактів, які забезпечують однорідність даних і створюють оптимальні умови для подальшого навчання глибоких моделей [26].

Нормалізація кольору є однією з ключових задач попередньої обробки зображень, що використовуються для навчання нейронних мереж. Її метою є усунення небажаної варіативності кольорів, спричиненої різними умовами освітлення, характеристиками камер або сканерів, а також відмінностями у процесах забарвлення чи цифрового перетворення (рисунок 2.2). Такі відмінності можуть змінювати розподіл кольорових каналів, що призводить до некоректного навчання моделі та зниження її узагальнювальної здатності. Особливо критичною колірна узгодженість є у сферах, де колір несе важливе семантичне навантаження, наприклад у гістологічних або мікроскопічних зображеннях, де навіть незначні колірні відхилення можуть вплинути на класифікацію клітинних структур. Забезпечення єдиного колірного простору

підвищує стабільність тренування та точність розпізнавання об’єктів нейронними мережами.

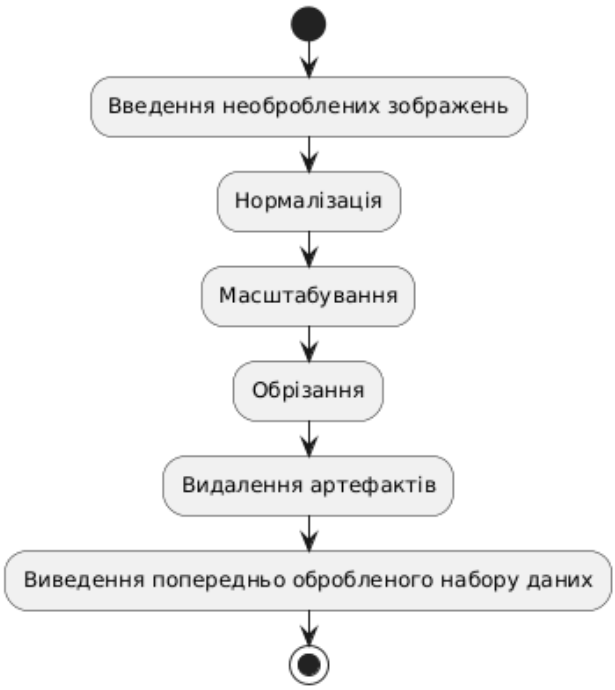

Рисунок 2.2 - Діаграма Конвеєра попередньої обробки зображень

Алгоритм нормалізації кольору Macenko є одним із найпоширеніших методів стандартизації забарвлення біомедичних зображень, зокрема у гістології (рисунок 2.3). Його принцип ґрунтується на переході від колірного простору RGB до простору оптичної щільності (Optical Density, OD), де інтенсивність пікселів описується логарифмічною залежністю від кількості світла, що проходить крізь зразок. Далі методом сингулярного розкладу (SVD) оцінюються основні вектори забарвлення, які відповідають різним барвникам. Отримані вектори використовуються для нормалізації відносно еталонного зображення, що дозволяє реконструювати колірну композицію з узгодженими характеристиками.

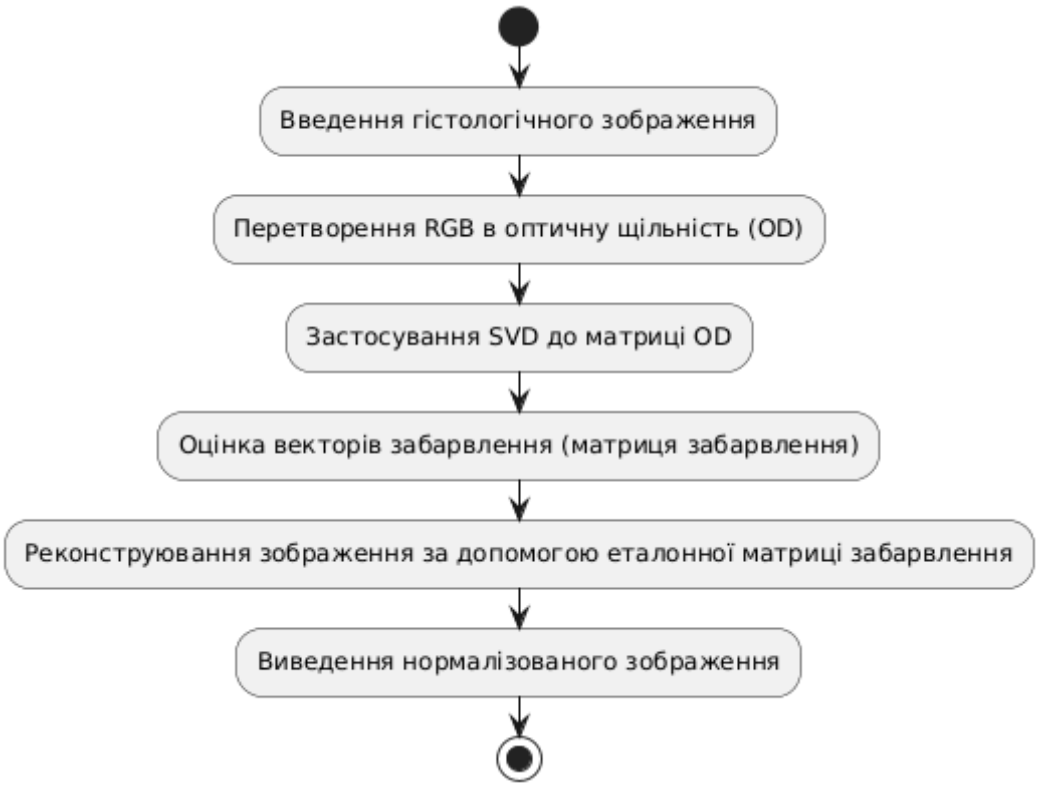

Рисунок 2.3 - Діаграма Алгоритму Macenko

Метод Macenko особливо ефективний при аналізі гістологічних зображень із гематоксилін-еозиновим забарвленням, де потрібно мінімізувати вплив варіацій лабораторного протоколу. Основними обмеженнями є чутливість до шуму, залежність від вибору стабільного еталона та зниження точності при наявності сильних артефактів.

Першим кроком є перехід з простору інтенсивностей до оптичної щільності (OD), що характеризує ступінь поглинання світла барвником. Обчислення виконується за формулою:

$$OD = -\log\left(\frac{I+1}{I_0}\right)$$

де:

I - інтенсивність пікселя;

I_0 - максимальна інтенсивність (біла точка).

Попри це, використання алгоритму Macenko значно підвищує рівень стандартизації біомедичних наборів даних, забезпечуючи узгодженість кольорів і покращуючи якість навчання глибоких нейронних мереж.

Алгоритм Reinhard є одним із класичних методів нормалізації кольору, який забезпечує узгодження кольорових характеристик між зображеннями. Його принцип полягає у перетворенні колірного простору з RGB до LAB, де компоненти L (яскравість), a і b (кольорові координати) мають більш лінійний зв'язок із людським сприйняттям кольору. Для кожного каналу обчислюються середнє значення та стандартне відхилення, після чого значення пікселів нормалізуються так, щоб статистики поточного зображення відповідали статистикам еталонного. Таким чином досягається узгодженість кольорового балансу без складного моделювання забарвлення.

Після перетворення у LAB-простір для кожного каналу виконується нормалізація через усереднення та масштабування відносно еталонного зображення. Формула для нормалізації яскравості (аналогічно застосовується до A та B каналів) має вигляд:

$$L' = \frac{L - \mu_L}{\sigma_L}\sigma_{L_r} + \mu_{L_r}$$

де:

L- значення яскравості в оригінальному зображенні;

μ_L, σ_L - середнє значення та стандартне відхилення яскравості вихідного зображення;

σ_{L_r}, μ_{L_r} - відповідні статистики для еталонного зображення.

У порівнянні з методом Macenko, підхід Reinhard є простішим у реалізації та менш чутливим до параметрів, проте демонструє меншу стабільність у складних сценах із високим рівнем шуму або неоднорідним освітленням (таблиця 2.2). Водночас він добре підходить для нормалізації макро- та мікрофотографій, зокрема при створенні навчальних наборів даних для класифікації гістологічних зрізів, де важливо зберегти загальну колірну узгодженість між зображеннями.

Таблиця 2.2 - Порівняння алгоритмів Macenko та Reinhard

Параметр	Macenko	Reinhard
Точність відновлення кольору	Висока при однорідному забарвленні	Середня, залежить від вибору еталону
Швидкість виконання	Повільніший через SVD-декомпозицію	Дуже швидке для статичних операцій
Стабільність при варіаціях освітлення	Висока, якщо стабільний еталон	Помірна - чутливий до шуму
Вимоги до еталонного зображення	Потрібне якісне, стандартизоване еталонне зображення	Еталон може бути довільним, але бажано з близькою колірною гамою
Типові сфери застосування	Біометричні зображення	Макро- і мікрофотографії, кольорові сцени з варіативним освітленням

Масштабування, зміна роздільності та обрізання (cropping) є базовими етапами попередньої обробки зображень, що забезпечують узгодженість вхідних даних і впливають на стійкість глибоких нейронних мереж. Масштабування дозволяє привести всі зображення до єдиного розміру, що необхідно для формування однакових тензорів під час навчання. Зміна роздільної здатності використовується для адаптації даних до обчислювальних обмежень і архітектури. Масштабування, зміна роздільності та обрізання (cropping) є базовими етапами попередньої обробки зображень, що забезпечують узгодженість текстури моделі. При цьому застосовуються різні методи інтерполяції - nearest-neighbor (найближчого сусіда), bilinear (білінійна) та bicubic (бікубічна), які визначають спосіб обчислення нових піксельних значень.

Обрізання (cropping) допомагає виділити ключові області інтересу, зменшити фонові шуми та підвищити інваріантність моделі до положення

об’єкта. Водночас надмірне зменшення масштабу може призвести до втрати дрібних деталей, некоректне обрізання - до перекривання важливих елементів, а зміна aspect ratio (співвідношення сторін) - до спотворення об’єктів.

Тому оптимальний вибір параметрів масштабування, інтерполяції та обрізання є необхідною умовою забезпечення стабільності навчання та узагальнювальної здатності нейронних мереж.

Такі дефекти можуть ускладнювати аналіз зображення та знижувати точність навчання моделей.

Для їх усунення застосовуються алгоритми фільтрації: median filter для зменшення імпульсного шуму, Gaussian filter для згладжування інтенсивності та bilateral filter для збереження контурів (таблиця 2.3). Корекція фону використовується для вирівнювання освітлення, усунення меж областей, для згладження переходів між сегментами скану, а морфологічні операції (ерозія, дилатація, відкриття, закриття) допомагають прибрати дрібні артефакти структури.

Таблиця 2.3 - Типи фільтрів для усунення шумів

Тип фільтра	Переваги	Недоліки
Median Filter	Добре видаляє імпульсний шум, зберігає краї об'єктів	Повільна обробка для великих зображень
Gaussian Filter	Ефективно приглушує випадкові шуми, плавно згладжує зображення	Розмиває краї та дрібні деталі
Bilateral Filter	Поєднує згладжування та збереження контурів	Висока обчислювальна складність, повільна роботи

Під час сканування зображень, особливо у біомедичних або мікроскопічних дослідженнях, часто виникають різноманітні артефакти, що

спотворюють дані. До них належать шуми сенсорів, відблиски від освітлення, пил або подряпини на поверхні зразка, межі між сканованими ділянками (overlap-області) та неоднорідність освітлення, яка створює варіації яскравості.

Видалення таких артефактів суттєво покращує якість сегментації тканин або клітин, зменшує хибні контури та підвищує точність класифікації. Отже, ефективна фільтрація є невід'ємною складовою формування чистого та надійного навчального набору даних.

2.3 Алгоритми сегментації та анотації

Сегментація та анотація зображень є ключовими етапами у формуванні навчальних наборів даних для глибокого навчання, оскільки саме вони забезпечують створення точних і структурованих міток, необхідних для навчання моделей.

Сегментація дозволяє виділити об'єкти або області інтересу на зображенні, що формує основу для подальшої класифікації, детекції чи кількісного аналізу. Анотація, своєю чергою, встановлює відповідність між зображенням і його семантичним змістом, надаючи мережі інформацію про категорії, межі та просторові взаємозв'язки об'єктів. Висока якість сегментації безпосередньо визначає точність і надійність глибоких моделей, особливо у задачах медичної діагностики, комп'ютерного зору чи автономних систем. Таким чином, правильна сегментація та узгоджена анотація є фундаментом ефективного та достовірного процесу навчання нейронних мереж (рисунок 2.4).

Ручна анотація зображень є базовим етапом створення навчальних наборів для глибокого навчання, що забезпечує точне визначення об'єктів і їхніх характеристик на рівні пікселів або регіонів. Цей процес виконується експертами, які вручну виділяють об'єкти інтересу та створюють відповідні мітки, що надалі використовуються для навчання та валідації моделей. Існують

три основні типи розмітки: маски (pixel-wise segmentation), контури (boundary annotation) та області інтересу (ROI, region of interest). Маски передбачають позначення кожного пікселя відповідно до класу і застосовуються у задачах семантичної сегментації, де важлива точність локалізації об’єктів.

Рисунок 2.4 - Діаграма процесу сегментації та анотації зображення

Для ручної анотації застосовуються спеціалізовані інструменти, такі як Labelbox, CVAT, Supervisely, LabelMe, що забезпечують зручний інтерфейс, можливість колективної роботи та інтеграцію з навчальними пайплайнами. Ручна розмітка, незважаючи на трудомісткість, є критично важливою, оскільки формує якісну основу для подальшої автоматизації, використання напівавтоматичних або повністю автоматичних методів сегментації та підвищує точність і надійність моделей глибокого навчання.

Контури використовуються для окреслення меж об’єктів, що дозволяє проводити морфометричні вимірювання, оцінювати форму та розміри структур (таблиця 2.4). ROI-розмітка виділяє ключові області на зображенні без деталізації на рівні пікселів і є доцільною у діагностичних дослідженнях або при швидкій візуальній оцінці даних.

Таблиця 2.4 - Порівняння типів розмітки

Тип розмітки	Мета	Точність	Інструменти
Маски	Семантична сегментація піксель за пікселем	Висока	CVAT, Supervisely, LabelMe
Контури	Морфологічні дослідження об'єктів	Середня-висока	LabelMe, supervisely
ROI	Виділення ключових областей	Середня	Labelbox, CVAT, QuPath

Напівавтоматичні методи анотації зображень поєднують ручне виділення об’єктів із автоматизованими алгоритмами, що дозволяє значно прискорити процес розмітки та підвищити її точність. Принцип роботи таких методів полягає у використанні автоматичних підходів, наприклад порогових методів, контурного виділення або сегментації за інтенсивністю, які пропонують попередні мітки. Користувач при цьому може вручну коригувати контури, уточнювати межі та видаляти помилково виділені області. Інструменти на приклад Paint.net (з плагінами), Adobe Photoshop, QuPath та ImageJ забезпечують інтерактивні інтерфейси для комбінованої роботи: користувач може швидко застосувати алгоритм, переглянути результат і внести необхідні корективи.

Такі підходи особливо корисні при аналізі гістологічних зрізів, об’єктів із нерівномірним забарвленням або неоднорідним фоном, де повністю автоматична розмітка часто дає помилки. Перевагами напівавтоматичних методів є скорочення часу анотації, зменшення людських помилок та можливість досягти високої точності при ручному уточненні.

Водночас існує компроміс між швидкістю і точністю: автоматизація дозволяє обробляти великі обсяги даних швидко, але остаточний контроль за контуром і якістю все ще потребує експертного втручання для досягнення максимальної достовірності розмітки.

Автоматизовані алгоритми попередньої сегментації широко застосовуються для підготовки зображень перед експертною розміткою, оскільки вони дозволяють прискорити процес і зменшити трудомісткість ручної анотації. Один із класичних методів - Watershed, який використовує градієнти інтенсивності для поділу зображення на області. Він добре справляється із чіткими границями об’єктів, але чутливий до шуму та часто потребує попередньої фільтрації або маркерної ініціалізації для запобігання пере-сегментації. GrabCut базується на мінімізації енергетичної функції через Graph Cuts, що дозволяє ефективно виділяти об’єкти на неоднорідному фоні при мінімальному ручному введенні (початковому ROI), забезпечуючи високу точність, але із значними обчислювальними затратами. Морфологічні операції (dilation, erosion, opening, closing) виділяють структурні особливості об’єктів на основі форми та розмірів, швидко працюють і ефективні для очищення шуму, проте мають обмежену здатність до сегментації складних форм.

Порівняно з точністю, GrabCut забезпечує найкращу сегментацію на складних фонах, Watershed - середню, а морфологія - базову. За обчислювальною складністю Watershed і морфологічні операції є швидкими, тоді як GrabCut значно вимогливіший. Потреба у ручному втручанні мінімальна у морфології, середня у Watershed (маркери) і початкова у GrabCut.

Експертна перевірка (валідація) анотацій є критичним етапом формування якісних навчальних наборів даних для глибокого навчання. Вона забезпечує точність розмітки об’єктів, коректність контурів, меж і класифікаційних ознак, що безпосередньо впливає на ефективність моделей у задачах класифікації, сегментації та детекції. Експерти оцінюють правильність анотацій шляхом візуальної перевірки, порівняння виділених областей із фактичними структурами та перевірки відповідності заданим категоріям. Для

оцінки міжекспертної узгодженості (inter-annotator agreement) застосовують статистичні показники, такі як коефіцієнт Каппи або Fleiss' Kappa, що дозволяють кількісно виміряти ступінь згоди між різними анотаторами.

Особливо важливою експертна валідація є у медичних і біологічних даних, де некоректна розмітка може призвести до помилкових висновків і знизити клінічну або наукову значущість результатів. Для контролю якості застосовуються різні підходи: подвійна перевірка, коли кожне зображення оцінюють кілька експертів; незалежна розмітка, що дозволяє виявляти систематичні помилки; а також аналіз метрик IoU (Intersection over Union) та Dice coefficient, які кількісно оцінюють збіг між автоматичною або первинною розміткою та експертною перевіркою.

Для кількісної оцінки якості розмітки та результатів автоматичної сегментації найчастіше використовуються метрики IoU (Intersection over Union - перетин над об'єднанням) та Dice Coefficient.

$$IoU = \frac{TP}{TP + FP + FN}$$

$$Dice = \frac{2TP}{2TP + FP + FN}$$

TP - кількість пікселів, правильно віднесених до об'єкта;

FP - пікселі, помилково віднесені до об'єкта;

FN - пікселі об'єкта, які не були розпізнані алгоритмом.

IoU (Коефіцієнт перетину) показує, наскільки область, визначена алгоритмом, перекривається з реальною областю об'єкта. Значення IoU→1 означає ідеальну сегментацію.

Dice (Коефіцієнт схожості) акцентує увагу на збалансуванні між правильними і помилковими пікселями, і є більш чутливим до малих об'єктів.

Таким чином, експертна валідація гарантує достовірність та стабільність навчальних наборів, забезпечуючи надійну основу для тренування глибоких нейронних мереж.

2.4 Алгоритми аугментації та генерації даних

Аугментація даних (data augmentation) є важливою складовою підготовки навчальних наборів для глибоких нейронних мереж, оскільки дозволяє штучно збільшити обсяг даних та підвищити їх різноманітність без необхідності додаткового збору нових зразків. Вона включає застосування різноманітних трансформацій до існуючих зображень - таких як обертання, масштабування, дзеркальне відображення, зміни освітленості чи кольору - що дозволяє моделі навчатися більш стійко до варіацій реальних умов.

Аугментація допомагає зменшити ризик перенавчання (overfitting), покращує здатність нейронної мережі узагальнювати знання на нові приклади та забезпечує більш стабільні й надійні результати класифікації, детекції та сегментації. У цьому підрозділі розглядаються основні методи аугментації та генерації даних, які застосовуються для підвищення якості навчальних наборів та ефективності навчання глибоких моделей.

Геометрична аугментація зображень включає набір трансформацій, які змінюють просторову організацію пікселів, створюючи нові варіанти навчальних даних, зберігаючи при цьому семантичну інформацію об’єктів (рисунок 2.5). Основні методи включають:

− Поворот (Rotation) - обертання зображення на заданий кут, що дозволяє мережі стати інваріантною до орієнтації об’єктів.

− Відзеркалення (Flip/Mirroring) - горизонтальне або вертикальне дзеркальне відображення, корисне для симетричних об’єктів.

− Масштабування (Scaling) - зміна розміру зображення або об’єкта, що підвищує інваріантність моделі до масштабу.

− Зсув (Translation) - переміщення об’єкта по горизонталі або вертикалі, що допомагає моделі розпізнавати об’єкти у різних положеннях.

− Обрізання (Crop) - виділення частини зображення, яке підсилює фокус на регіоні інтересу.

− Афінні перетворення (Affine transformations) - комбінації поворотів, масштабування, зсувів та віддзеркалень із збереженням паралельності прямих.

− Перспективні перетворення (Perspective/Projective transformations) - моделюють зміну точки зору, що корисно для сцен із різними кутами огляду.

Рисунок 2.5 - Діаграма конвеєра доповнення даних

Ці операції дозволяють створювати численні варіації навчальних прикладів без втрати семантики, підвищуючи здатність нейронної мережі узагальнювати знання. Водночас надмірне застосування аугментації може призвести до спотворення об’єктів, втрати контексту або некоректного відображення структури, що негативно впливає на точність моделі (таблиця 2.5).

Методи зміни кольору та контрасту є важливими для аугментації даних, оскільки дозволяють нейронним мережам навчатися стійко до варіацій освітлення, сканування та забарвлення зразків. До основних підходів належать:

− Зміна яскравості, насиченості та контрастності - коригування цих параметрів у зображенні дозволяє моделі адаптуватися до різних рівнів освітленості та інтенсивності кольорів;

− Перетворення в інші колірні простори (HSV, LAB) - розділення компонентів освітленості та кольору дає можливість змінювати інтенсивність або тон окремо, зберігаючи структурну інформацію;

− Додавання шуму та варіації освітлення - симулює реальні дефекти сканування або відмінності між обладнанням, допомагаючи моделі ігнорувати незначні артефакти;

− Випадкові фільтри та спотворення - додають непередбачувані зміни, що підвищують різноманітність тренувальних даних.

Таблиця 2.5 - Вплив колірних варіацій на точність моделі

Домен	Тип колірної варіації	Вплив на точність	Приклад застосування
Біомедицина	Відмінності лабораторних забарвлень	Сильний	Гістологічні зрізи з різних лабораторій
Промисловість	Освітлення на конвеєрі	Помірний	Виявлення дефектів на виробах
Транспорт	Різні погодні умови, час доби	Помірний	Розпізнавання дорожніх знаків

Одним із ефективних методів підвищення стійкості моделей глибокого навчання є додавання випадкового шуму. Така техніка дозволяє моделі навчитися ігнорувати незначні варіації у зображеннях, що виникають через

різницю сенсорів, умови освітлення або цифрові артефакти. Зазвичай використовується гаусовий шум, який моделює природні флуктуації інтенсивності пікселів.

$$I' = I + \mathcal{N}(0, \sigma^2)$$

де;

I - початкове зображення;

I' - зображення після додавання шуму;

$\mathcal{N}(0, \sigma^2)$ - нормально розподілений шум із середнім значенням 0 та дисперсією σ^2.

Такі операції дозволяють створювати нові варіанти навчальних прикладів, зберігаючи семантичний зміст зображення. У біомедичних дослідженнях це особливо актуально: наприклад, гістологічні зображення або мікроскопічні знімки можуть відрізнятися кольором і насиченістю залежно від лабораторії, протоколів фарбування та налаштувань сканера. Аугментація кольору дозволяє навчити модель розпізнавати об’єкти незалежно від таких варіацій, підвищуючи її узагальнювальну здатність та точність у реальних умовах.

2.5 Висновки до розділу 2

Формування навчальних наборів даних базується на застосуванні послідовних алгоритмів, які забезпечують структурну цілісність та інформативність вибірки. Попередня обробка зображень, включаючи фільтрацію шумів, корекцію кольору та нормалізацію, дозволяє підвищити якість даних для подальшого аналізу. Методи сегментації та анотації

створюють точні маски об’єктів, які є критично важливими для навчання моделей комп’ютерного зору. Аугментація та синтетичне генерування зображень забезпечують варіативність вибірки та допомагають уникнути проблеми дисбалансу класів. Ефективність застосованих алгоритмів визначається балансом між точністю, швидкістю та здатністю адаптуватися до різних типів зображень, що у підсумку формує стабільну та придатну до навчання вибірку.

3 ФОРМУВАННЯ НАБОРІВ ДАНИХ В ПРОГРАМНОМУ СЕРЕДОВИЩІ PAINT.NET

3.1 Алгоритми структурування та організації датасету

Правильна організація та структурування навчальних даних є критично важливими для ефективного навчання глибоких нейронних мереж та забезпечення відтворюваності експериментів. Добре структурований датасет дозволяє легко інтегрувати нові зразки, контролювати якість анотацій, відслідковувати джерела даних і забезпечує прозорість у всьому процесі підготовки та обробки інформації. Крім того, чітка структура сприяє автоматизації обробки, застосуванню аугментації та інших методів попередньої обробки, зменшуючи ризик помилок і непослідовності в тренуванні моделей. У цьому підрозділі розглядаються основні алгоритми та підходи до організації датасетів, включаючи стандартизацію файлових структур, ієрархію категорій, ведення метаданих та механізми контролю якості, що забезпечують стабільну та надійну основу для подальшого навчання глибоких моделей.

У сучасних навчальних наборах для глибокого навчання використовуються різні формати збереження зображень, які відрізняються якістю, підтримкою кольорової глибини та сумісністю з медичними системами.

- PNG - формат без втрат, що забезпечує точне відтворення пікселів і підтримує альфа-канал для прозорості. Його часто використовують для збереження масок і анотацій, де критична точність пікселів;
- JPEG - формат із втратами, який забезпечує значне стиснення файлів за рахунок деякої деградації якості. Його застосовують для загальних зображень, де важливий баланс між розміром файлу та візуальною якістю, наприклад у комп’ютерному зорі чи класифікації об’єктів;
- TIFF - універсальний формат із підтримкою багатошарових зображень і високої глибини кольору, що зберігає точність даних без втрат. Його часто

використовують у біомедичних задачах, особливо для гістопатології, де зображення можуть бути великими і потребують високої деталізації;

− WSI (Whole Slide Image) - спеціалізований формат для сканованих гістологічних зрізів, сумісний із медичними стандартами DICOM або OME-TIFF, дозволяє зберігати величезні зображення з багаторівневою структурою та підтримкою метаданих.

Таблиця 3.1 - Вибір формати за задачею

Задача	Рекомендований формат	Пояснення вибору
Анотація масок	PNG	Точна передача пікселів без втрат
Загальні зображення	JPEG	Компактність, прийнятна якість
Біометричні високо-деталізовані	TIFF	Підтримка багатошаровості без втрат
Цифрові гістологічні зрізи	WSI	Великий розмір, сумісність з DICOM

Рекомендації щодо вибору формату:

− PNG - для анотацій і масок;

− JPEG - для загальних зображень із обмеженим обсягом;

− TIFF/WSI - для гістопатології та високодеталізованих медичних зображень, де важлива точність і сумісність із стандартами.

Правильне структурування директорій навчального набору є ключовим для організованої роботи з даними, забезпечення відтворюваності експериментів та автоматизації процесу навчання моделей. Стандартна архітектура передбачає поділ набору на train, validation та test, що дозволяє відокремити дані для навчання, налаштування гіперпараметрів та оцінки продуктивності моделі. Крім того, можуть додаватися каталоги augmented для

збереження аугментованих прикладів, metadata для зберігання інформації про зображення та анотації, а також masks для сегментаційних масок або ROI.

Приклад стандартної структури директорій:

```
dataset/
│
├─ train/
│   ├─ images/
│   └─ masks/
│
├─ validation/
│   ├─ images/
│   └─ masks/
│
├─ test/
│   ├─ images/
│   └─ masks/
│
├─ augmented/
│   ├─ images/
│   └─ masks/
│
└─ metadata/
    └─ annotations.csv
```

Така організація полегшує інтеграцію набору даних із фреймворками PyTorch, TensorFlow та Keras, оскільки більшість бібліотек очікують однакову структуру для автоматичного завантаження зображень та масок. Крім того, вона спрощує застосування пайплайнів попередньої обробки та аугментації, дозволяє легко додавати нові дані та повторно використовувати набори у різних експериментах, забезпечуючи зручність та ефективність роботи з великими датасетами.

Анотаційні маски та метадані є невід’ємною частиною навчальних наборів для глибокого навчання, оскільки вони визначають локалізацію об’єктів, їхні категорії та додаткові властивості. Для збереження анотацій використовуються різні формати, залежно від задачі та сумісності з інструментами. JSON часто застосовується у сучасних фреймворках і дозволяє зберігати координати об’єктів, класи та шляхи до масок у структурованому

вигляді. XML, зокрема у форматі Pascal VOC, описує bounding box, клас об’єкта та метадані для кожного зображення, що забезпечує сумісність із багатьма бібліотеками для детекції об’єктів. COCO — більш гнучкий формат JSON, який дозволяє зберігати сегментаційні маски, ключові точки, категорії та додаткову інформацію про зображення, використовується у великих датасетах комп’ютерного зору.

Приклад JSON для одного об’єкта:

```
{
  "id": 101,
  "class": "cell",
  "bbox": [50, 30, 120, 100],
  "mask_path": "masks/image_101_mask.png"
}
```

Збереження таких метаданих дозволяє узгоджувати різні набори даних, полегшує інтеграцію нових прикладів, а також сприяє дотриманню FAIR-принципів (Findable, Accessible, Interoperable, Reusable). Стандартизовані анотації забезпечують повторюваність експериментів, автоматизовану обробку та спільне використання наборів даних у науковій спільноті, мінімізуючи ризик помилок і невідповідностей між різними джерелами інформації.

Стандартизована структура директорій та уніфіковані формати збереження зображень, анотацій і метаданих є критично важливими для ефективного управління навчальними наборами даних у глибокому навчанні. Вони забезпечують сумісність із популярними фреймворками, такими як PyTorch, TensorFlow та Keras, дозволяють легко інтегрувати нові дані та застосовувати пайплайни попередньої обробки й аугментації. Крім того, стандартизація сприяє масштабованості: великі обсяги зображень і масок можуть бути організовані так, щоб їх можна було швидко обробляти та повторно використовувати в різних експериментах. Водночас правильне структурування та формати збереження гарантують відтворюваність досліджень, оскільки інші дослідники можуть відтворити навчання моделей,

використовуючи однакові правила організації та метадані. Таким чином, стандартизована організація та збереження даних є фундаментальною передумовою для побудови надійних, ефективних та узагальнювальних моделей глибокого навчання.

3.2 Структура та інтерфейс програмного засобу

Інтерфейс ПЗ виглядає так, при запуску програми відкривається основне вікно (рисунок 3.1), на якому зображено компоненти для швидкого доступу які відкриваються самостійно при запуску а саме: Інструменти, Історія, Палітра, Шари.

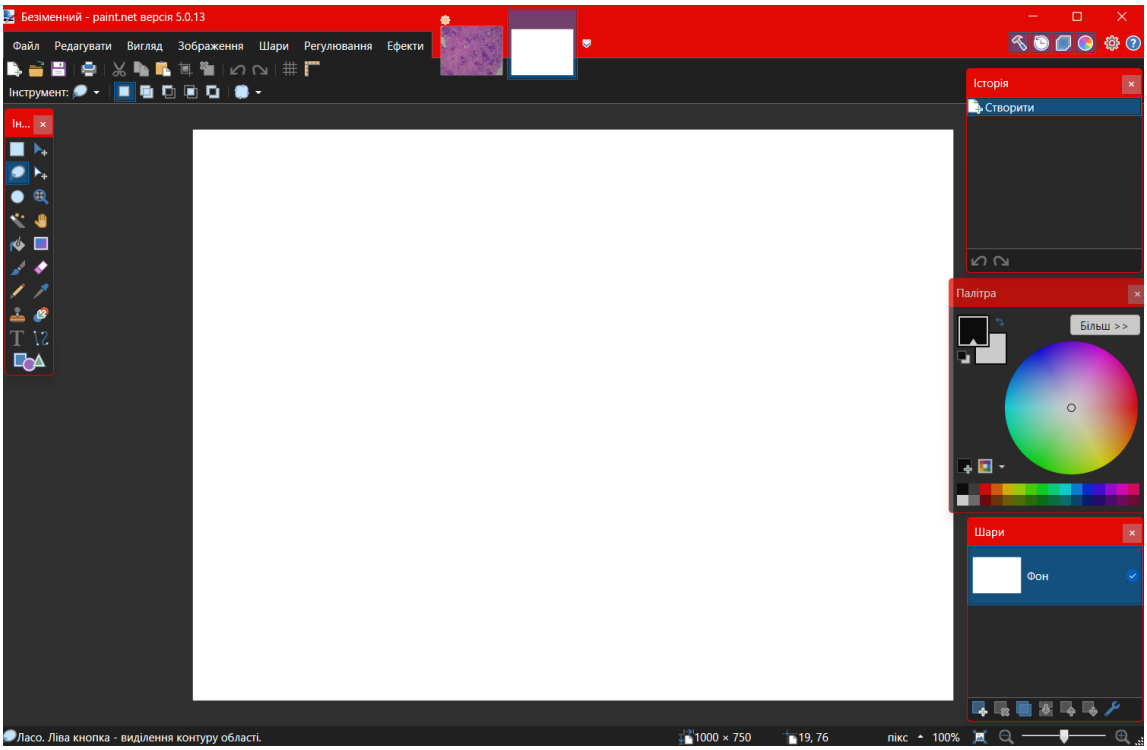

Рисунок 3.1 - Основне вікно ПЗ

Панель інструментів дозволяє виконувати різні маніпуляції з зображенням такі як:

- перемістити виділені частини, дозволяє перемістити виділені частини автоматом виділяє всі тому потрібно вручну вибрати частину яку саме потрібно перемістити;

- ласо, дозволяє обвести певну частину максимально коректно;

- масштаб, дозволяє приближувати або віддаляти зображення за допомогою лівої та правої кнопки миші;

- чарівна паличка, дозволяє виділяти область зображення яка має схожий між собою колір, також можна налаштувати чутливість виділення;

Компонент історія роботи з зображенням показує яку дію було виконано в даний момент або раніше (рисунок 3.2), що дозволяє переглядати виконану роботу та повертати виконану в даний момент дію до минулої якщо була зроблена помилка.

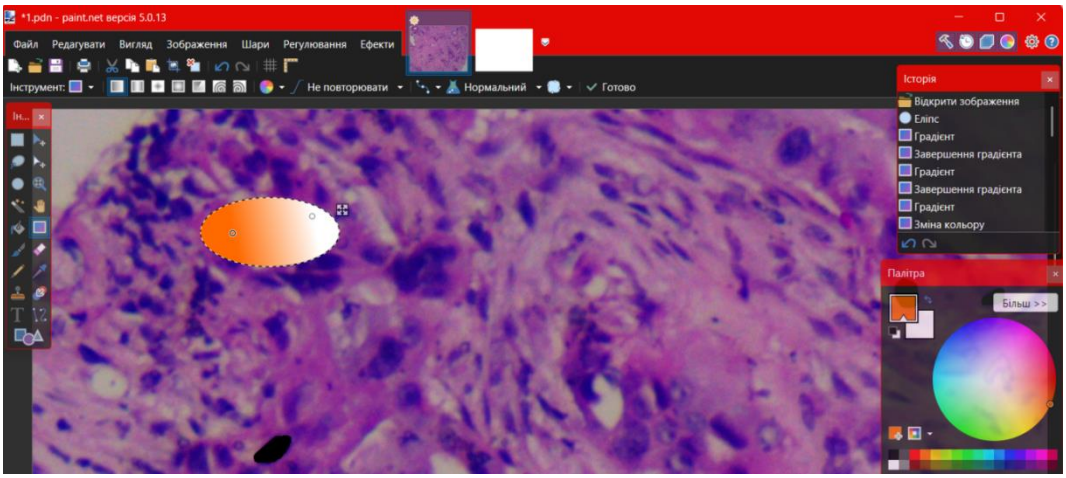

Рисунок 3.2 - Історія виконаної роботи

Компонент шари показує скільки шарів є на зображені і дозволяє додавати або видаляти шари за потреби (рисунок 3.3). Шари можна зробити під кожну дію виконану з зображенням, а також робити його видимим або не видимим для перевірки.

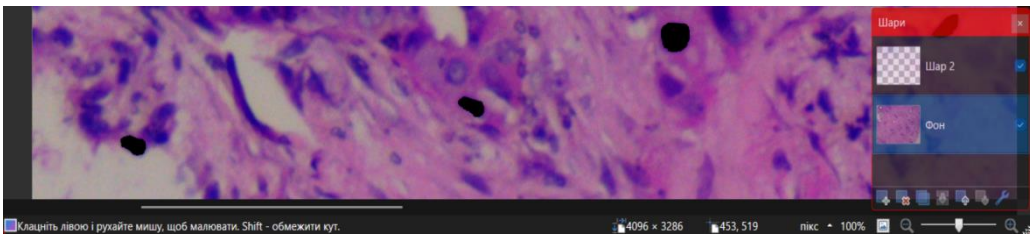

Рисунок 3.3 - Шари

Компонент палітра показує який саме колі вибраний для заливки виділених об’єктів, також дозволяє коригувати колір і його прозорість (рисунок 3.4).

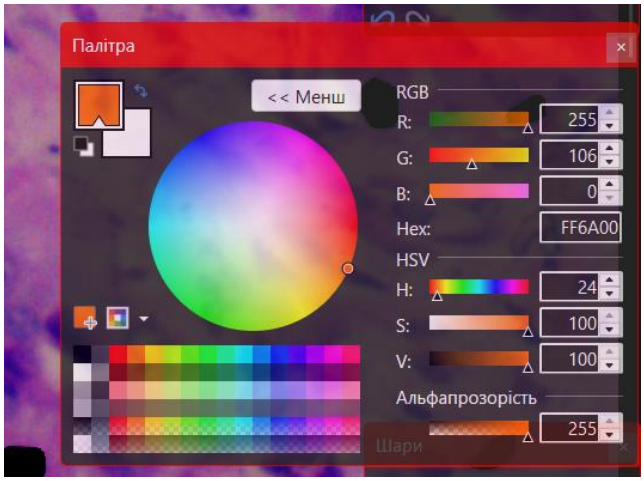

Рисунок 3.4 - Можливості палітри

Параметр контрастність який можна вибрати в верхній допоміжній панелі дозволяє робити зображення більш або менш контрастним (рисунок 3.5), що дозволяє більш коректно виділяти потрібні об’єкти (рисунок 3.6).

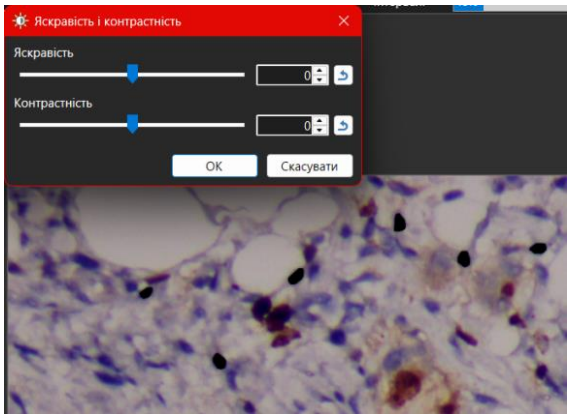

Рисунок 3.5 - Зображення без контрастності

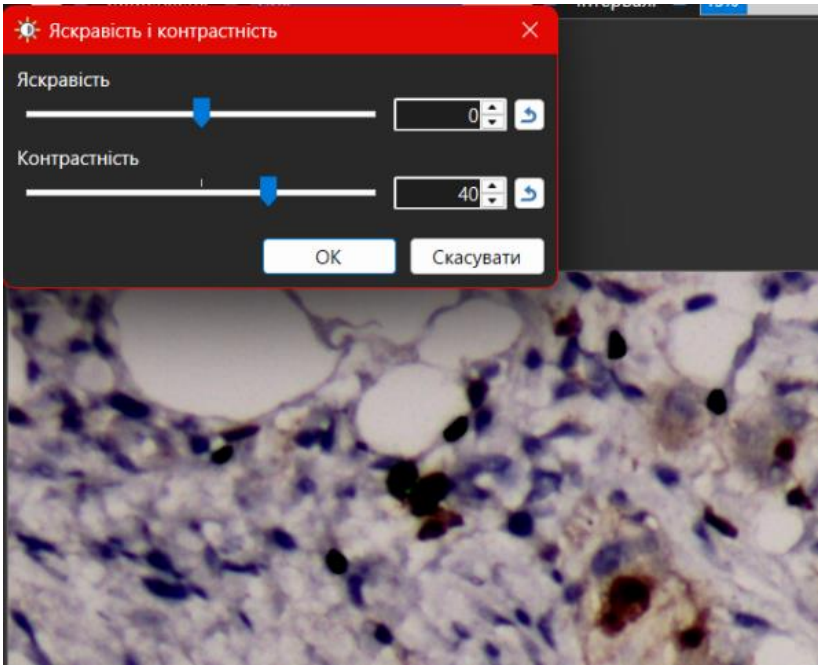

Рисунок 3.6 - Зображення з викрученою контрастністю на +40%

Параметр яскравість дозволяє робити зображення більш або менш яскравим для більш чіткого і правильного виявлення потрібних об’єктів (рисунок 3.7).

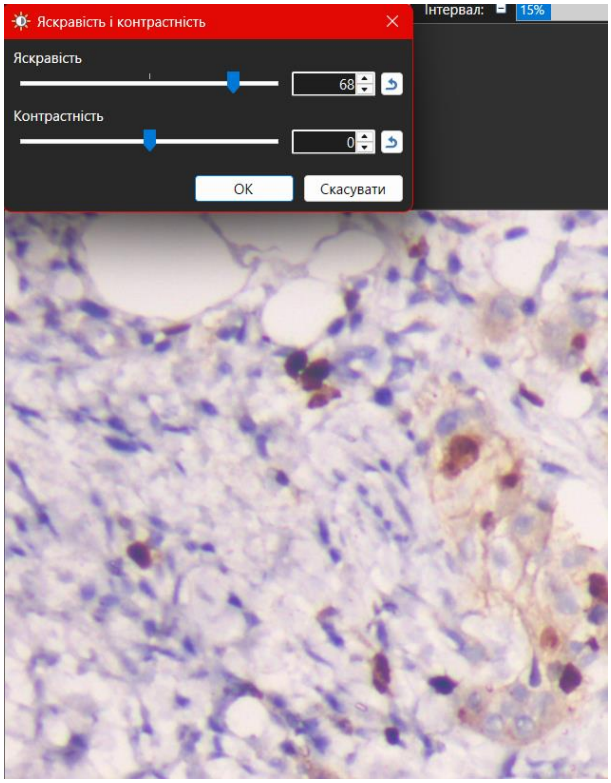

Рисунок 3.7 - Зображення з викрученою яскравістю на +68%

Параметр медіанного розмиття дозволяє розмивати зображення для чіткого виявлення об’єктів (рисунок 3.8).

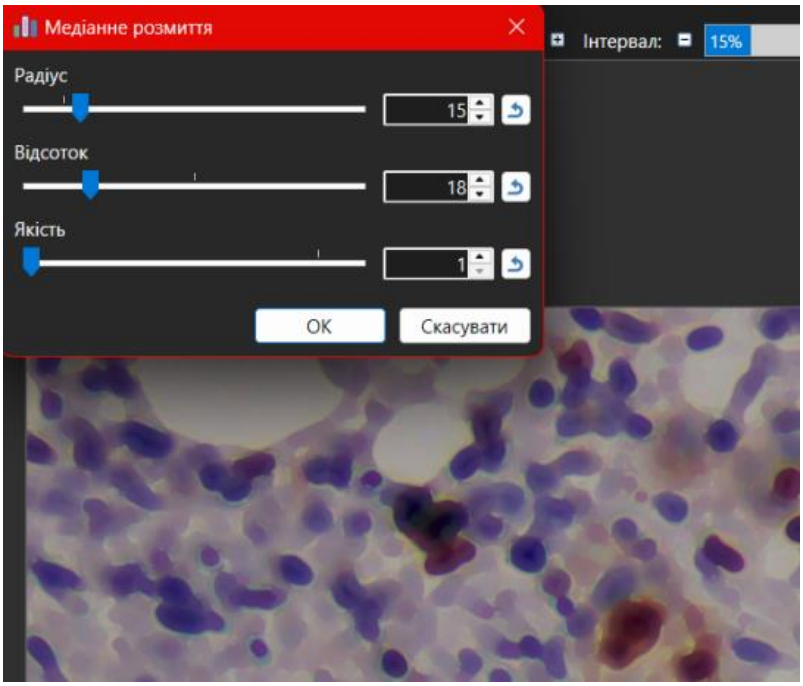

Рисунок 3.8 - Зображення з використанням медіанного розмиття

Порівняємо програму Paint.NET з схожою програмою для редагування зображення QuPath.

При відкриті програми зразу видно панель інструментів, докладний опис вибраного зображення, панель виділених об'єктів (рисунок 3.9).

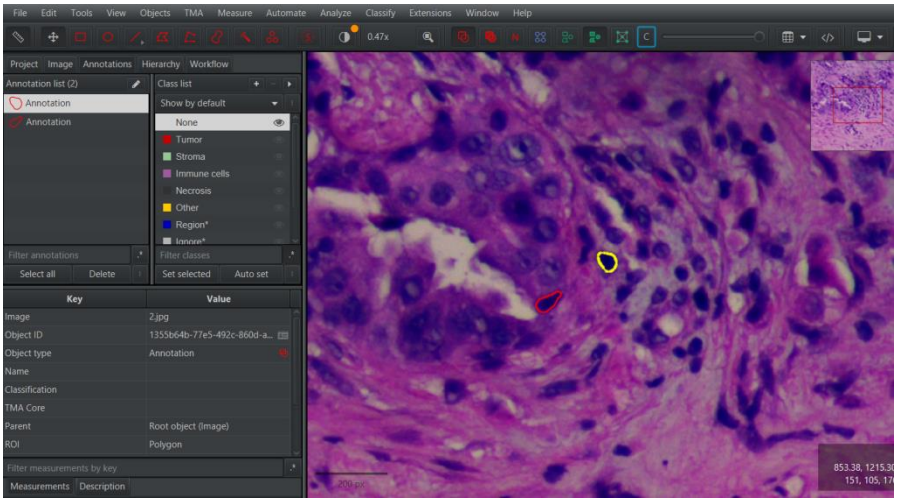

Рисунок 3.9 - Основне вікно ПЗ

В панелі інструментів з потрібних нам інструментів є чарівна паличка, редагування яскравості та контрастності, але немає інших потрібних компонентів таких як: Палітри для заливання виділених об'єктів, інструмента Ласо, компонента для створення додаткового шару.

Компонент яскравість та контрастність в даній програмі є більш детальним з зміною RGB діапазону (рисунок 3.10, 3.11), що дозволяє біль коректно виявляти об'єкти.

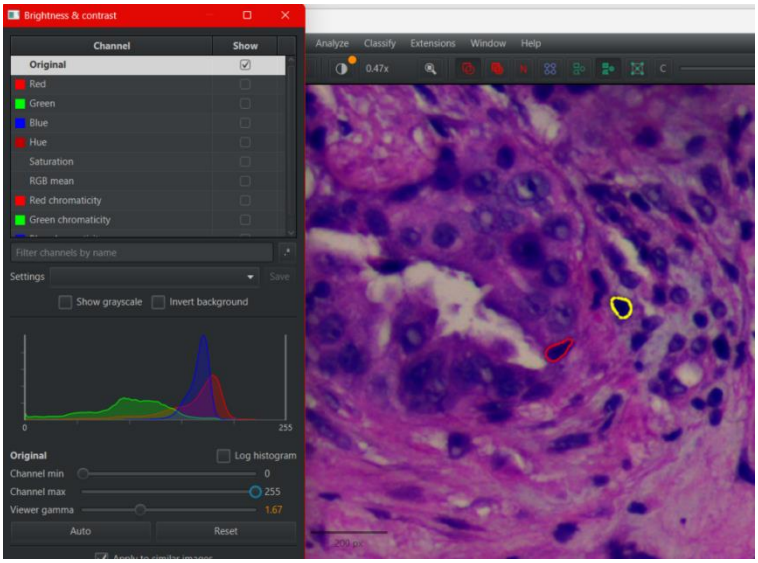

Рисунок 3.10 - Не редаговане зображення

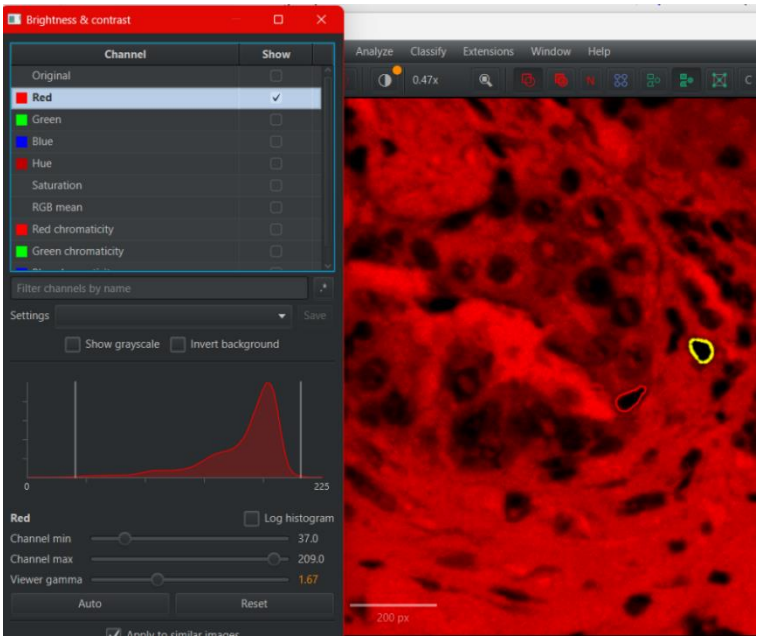

Рисунок 3.11 - Змінення фону на червоний колі для виявлення об'єктів

3.3 Експерименти

Проведемо експеримент з зображенням для виявлення клітин, для початку візьмемо нове зображення (рисунок 3.12).

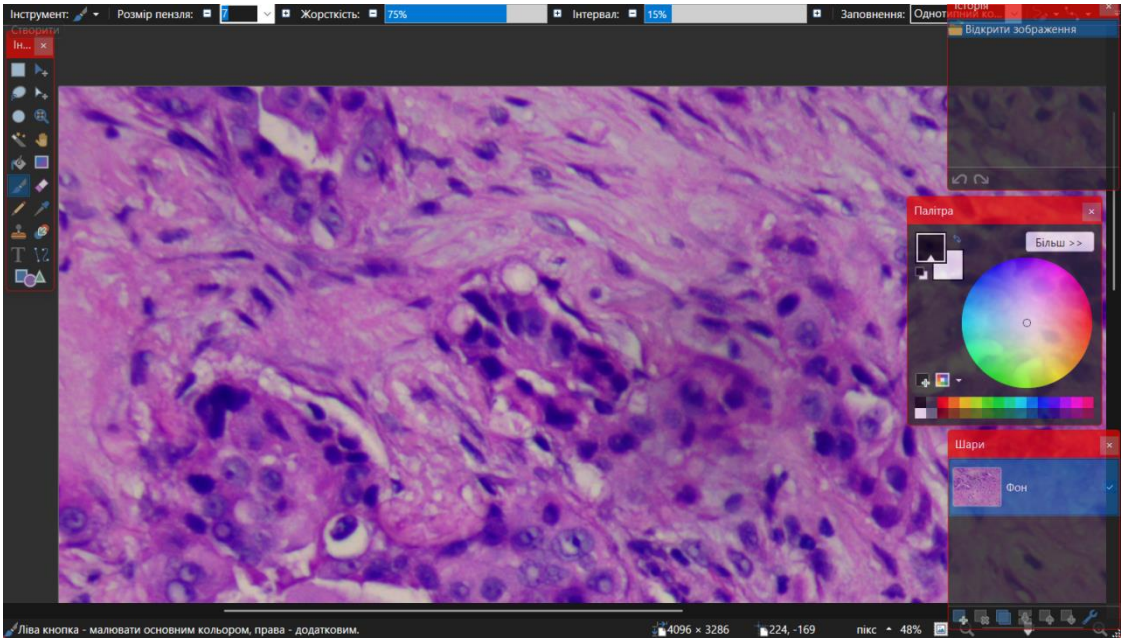

Рисунок 3.12 - Нове зображення

Добавимо для даного зображення новий шар для роботи з ним (рисунок 3.13).

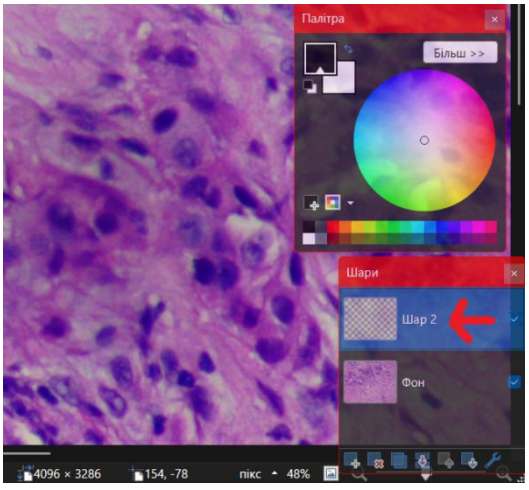

Рисунок 3.13 - Додавання нового шару

Далі вибираємо з панелі інструментів інструмент “Ласо” і виділяємо необхідний об’єкт на зображенні (рисунок 3.14).

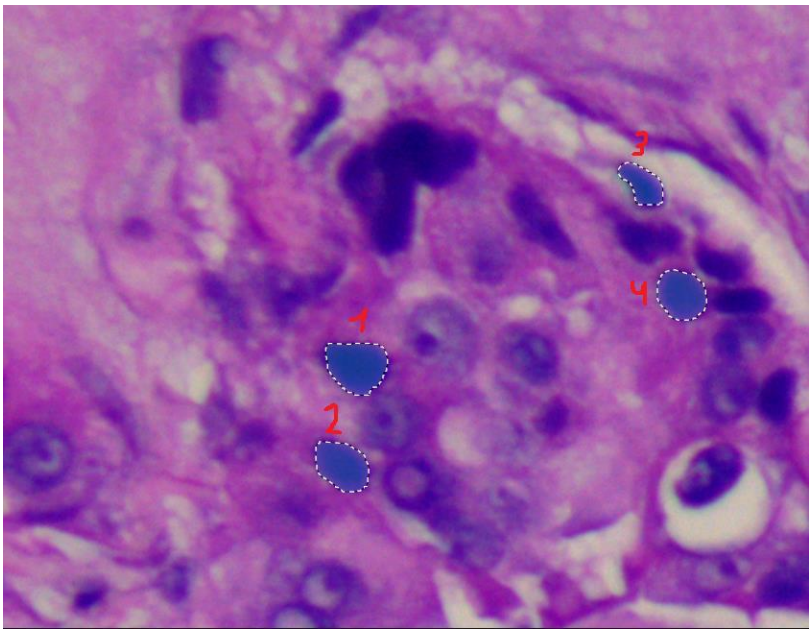

Рисунок 3.14 - Виділені об’єкти за допомогою інструмента Ласо

Далі заливаємо його чорним кольором для чіткого розуміння заходження об’єкта за допомогою палітри (рисунок 3.15).

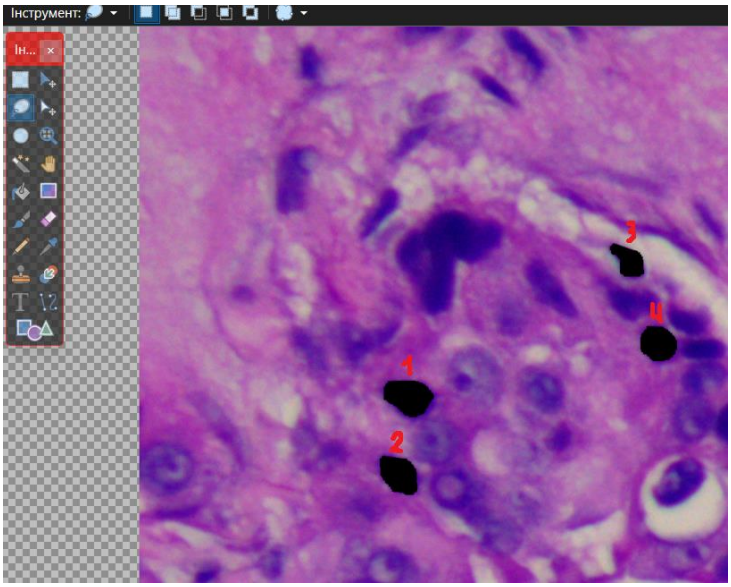

Рисунок 3.15 - Робота з об’єктами

Також змінимо параметри яскравість та контрастність для біль чіткого виявлення потрібних об’єктів (рисунок 3.17) та порівняємо з звичайним зображенням (рисунок 3.16).

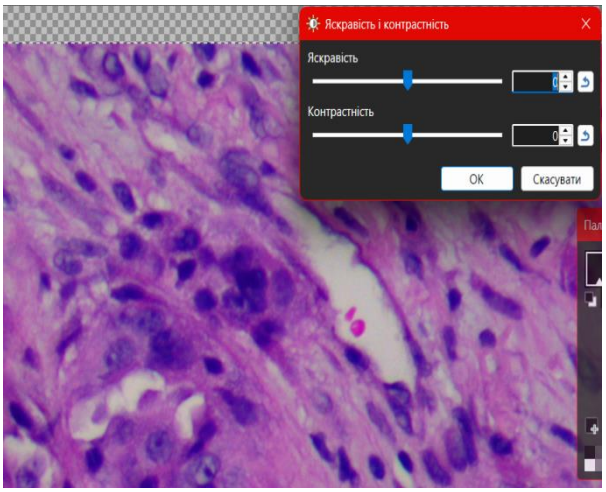

Рисунок 3.16 - Початкові значення

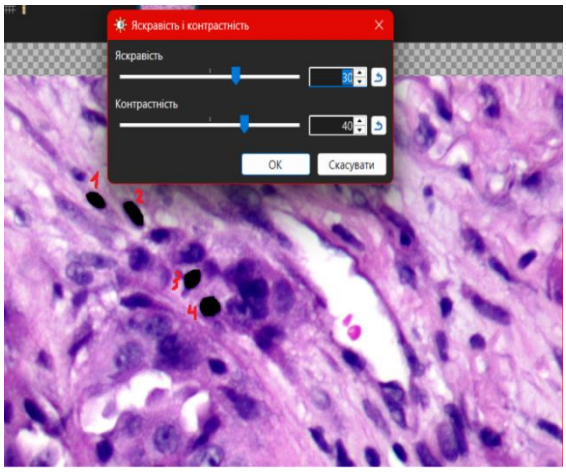

Рисунок 3.17 - Вибрані значення

Нижче представлено результати експерименту, спрямованого на визначення впливу різних варіантів попереднього оброблення зображень у програмному середовищі Paint.NET на якість навчального набору даних, призначеного для тренування глибокої нейронної мережі. Експеримент має на меті виявити залежності між якістю підготовлених зображень, складністю анотації та підсумковими показниками роботи моделі.

Метою експерименту є оцінювання того, як різні алгоритми попереднього оброблення впливають на формування структурованого, чистого та узгодженого датасету. Основна увага приділяється таким аспектам: зменшення кількості зображень з артефактами, скорочення часу анотації, підвищення узгодженості між розмітниками та покращення точності моделі, навченої на сформованих наборах даних. Таким чином, якість датасету розглядається як інтегральний показник, що охоплює як властивості самих зображень, так і результати роботи моделі.

Для експерименту використовувався вихідний набір із 1500 зображень, різних за роздільною здатністю, рівнем шуму та контрастності. На їх основі

сформовано три незалежні піднабори, у яких застосовано різні підходи до попереднього оброблення:

– Варіант A - відсутність оброблення. До зображень застосовувалося лише масштабування до одного розміру та базове кадрування без додаткових корекцій.

– Варіант B - базове попереднє оброблення. Виконувалася корекція яскравості та контрасту, видалення найочевидніших артефактів, а також легке шумозаглушення.

– Варіант C - розширене попереднє оброблення. Крім дій, описаних у варіанті B, до зображень застосовувалася нормалізація кольору, ширше шумозаглушення та підсилення різкості, а також уніфікація фону там, де це було можливим.

Усі три піднабори анотувалися двома незалежними розмітниками, що дозволило оцінити не лише якість зображень, але й суб'єктивний аспект інтерпретації даних.

Для кожного варіанта попереднього оброблення було виконано послідовні етапи оцінювання:

1. Анотація зображень. Фіксувався час розмітки одного зображення, а також частка випадків, коли розмітник повідомляв про ускладнення через низьку якість зображення.

2. Оцінка узгодженості між розмітниками. За допомогою коефіцієнта κ(Коена) оцінювалась ступінь узгодженості розмітки між двома учасниками анотації.

3. Навчання моделі. На основі кожного піднабору незалежно навчалась одна й та сама архітектура нейронної мережі. Умови навчання, включаючи кількість епох і параметри оптимізатора, залишалися незмінними.

4. Оцінювання точності. Для кожного з трьох варіантів було визначено точність моделі на однаковій тестовій вибірці, незалежній від навчальних наборів.

У межах експерименту використано такі метрики:

- P_{art} - частка зображень, що містять артефакти та ускладнюють анотацію;
- T_{ann} - середній час анотації одного зображення, с;
- κ - коефіцієнт узгодженості між розмітниками;
- Acc - точність моделі на валідаційній вибірці.

Ці метрики дозволяють кількісно порівняти варіанти попереднього оброблення та визначити, які саме підходи забезпечили найвищу якість датасету. Результати наведено у таблиці 3.2.

Таблиця 3.2 - Порівняльні результати оцінювання якості датасету

Варіант опрацювання	P_{art}, %	T_{ann}, с	κ	Acc, %
A – без оброблення	18	42	0,71	84,5
B – базове	9	35	0,81	88,2
C – розширене	4	31	0,87	90,1

Отримані дані демонструють зменшення частки проблемних зображень майже вдвічі при застосуванні базового оброблення та у чотири рази - при розширеному. Водночас скорочується час анотації, що свідчить про покращення візуальної якості та зрозумілості зображень. Зростання коефіцієнта узгодженості від 0,71 до 0,87 свідчить, що попереднє оброблення зменшує неоднозначність інтерпретації об’єктів. Підсумкові значення точності моделі підтверджують ці висновки: розширене оброблення забезпечило приріст точності майже на 6 відсоткових пунктів порівняно з варіантом без будь-яких корекцій.

Графік демонструє чітку тенденцію зростання якості моделі разом із підвищенням якості підготовки зображень (рисунок 3.18). Найбільший приріст спостерігається при переході від необробленого датасету до базово нормалізованого, що свідчить про вагомий вплив навіть простих операцій попереднього оброблення на ефективність навчання.

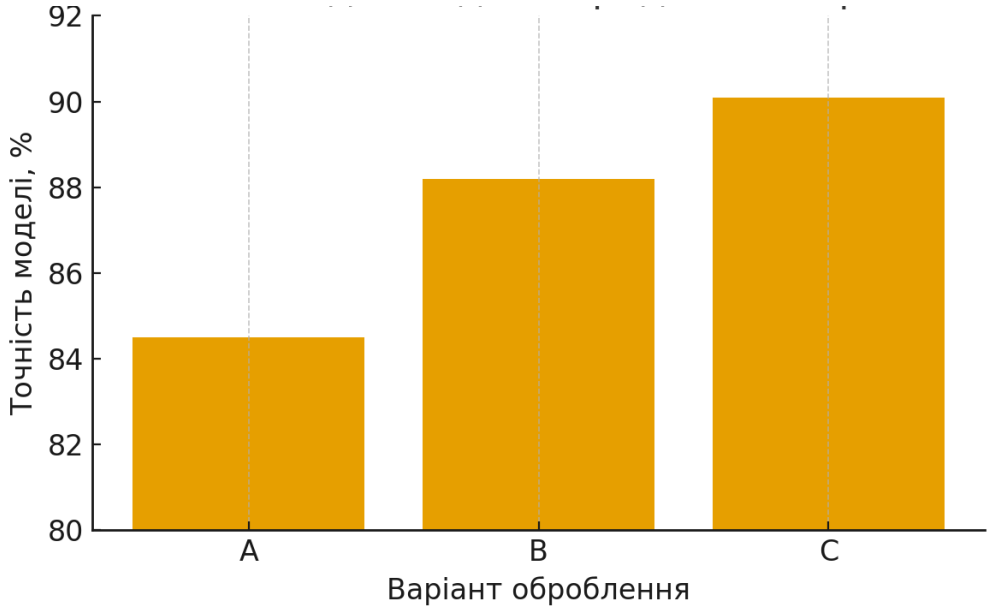

Рисунок 3.18 – Попереднє оброблення

Висновки до розділу 3

Проведений експеримент підтверджує суттєвий вплив попереднього оброблення зображень у середовищі Paint.NET на якість сформованих навчальних наборів даних. Застосування алгоритмів нормалізації, зменшення шумів, корекції контрасту, уніфікації фону та підсилення різкості дозволило:

- зменшити кількість артефактів, що ускладнюють анотацію;
- підвищити швидкість роботи розмітників;
- покращити узгодженість анотацій;
- підвищити точність моделі, навченої на підготовлених зображеннях.

ВИСНОВКИ

1. Проведено комплексний аналіз джерел формування навчальних наборів даних для задач глибокого навчання, зокрема відкритих датасетів, спеціалізованих колекцій. Проаналізовано програмні засоби, що застосовуються для збору, анотації та структурування даних, включаючи сучасні інструменти сегментації, аугментації та попередньої обробки.

2. Розроблено структуровану методику формування датасетів, що охоплює попередню обробку, сегментацію, анотацію, аугментацію та генерацію синтетичних даних.

3. Розроблено алгоритми до інтеграції методів сегментації, анотації який дозволяє формувати різноманітні та збалансовані навчальні вибірки.

4. Реалізовано алгоритми обробки зображень у середовищі Paint.NET, що включають роботу з шарами, виділення об’єктів, корекцію параметрів зображення й підготовку масок.

5. Проведені експерименти з формування та обробки навчального набору даних підтвердили ефективність розроблених алгоритмів. Результати показали, що застосовані методи забезпечують покращення виявлення та сегментації об’єктів та забезпечують відповідність даних вимогам моделей глибокого навчання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Копія А.В., Пошук наборів біомедичних зображень за метаданими. Збірник тез доповідей III Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Інтелектуальні комп'ютерні системи та мережі» 25 листопада 2025 р. Тернопіль. Україна. С. 78-79

2. Копія А.В., Алгоритми анотації гістологічних зображень III Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Інтелектуальні комп'ютерні системи та мережі» 25 листопада 2025 р. Тернопіль. Україна. С. 79-81

3. Березький О.М., Мельник Г.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітнього ступеня "Магістр". Спеціальність: 123 - Комп'ютерна інженерія. Магістерська програма - Комп'ютерна інженерія". Тернопіль: ЗУНУ, 2024. 32 с.

4. Методичні вказівки до оформлення курсових проектів, звітів про проходження практики, випускних кваліфікаційних робіт для студентів спеціальності «Комп'ютерна інженерія» / І.В. Гураль, Л.О. Дубчак / Під ред. О.М. Березького. Тернопіль: ТНЕУ, 2019. 33 с.

5. Березький О. М., Березька К. М., Попіна С. Ю. Статистичне оброблення цитологічних зображень. Вісник Хмельницького національного університету: зб. наук.-техн. праць. Сер.: Технічні науки. 2012. № 5. С. 161–164.

6. Березький О. М., Березька К. М., Батько Ю. М., Мельник Г. М. Синтез альтернативних рішень при структурному проектуванні систем автоматизованої мікроскопії. Науковий вісник НЛТУ України: зб. наук.-техн. праць. Львів: РВВ НЛТУ України. 2009. Вип. 19.5. С. 258-268.

7. Березький О. М., Мельник Г.М., Батько Ю.М., Дацко Т. В. Інтелектуальна система для діагностування різних форм раку молочної залози на основі аналізу гістологічних і цитологічних зображень. Науковий вісник НЛТУ України: зб. наук.-техн. праць. Львів: РВВ НЛТУ України. 2013. Вип. 23.13. С. 357-367.

8. Березький О. М., Батько Ю.М., Мельник Г.М. Інформаційно-аналітична система дослідження та діагностування пухлинних клітин на основі аналізу їх зображень. Вісник Хмельницького національного університету. Технічні науки. 2008. №4. С.33-41.

9. Березький О. М., Батько Ю.М., Мельник Г.М. Комп'ютерна система аналізу біомедичних зображень. Вісник Національного університету «Львівська політехніка». Комп'ютерні науки та інформаційні технології. 2009. № 570. С. 84-89.

10. Березький О. М., Батько Ю.М., Мельник Г.М. Методи сегментації біомедичних зображень. Вісник Хмельницького національного університету. Технічні науки. 2010. №1. С.189- 197.

11. Березький О. М. Методи та алгоритми перетворення контурів зображень в афінному просторі. Вісник Національного університету «Львівська політехніка». Комп'ютерні науки та інформаційні технології. 2009. № 638. С. 185-189.

12. Грицик В.В., Березька К.М., Березький О. М. Моделювання та синтез складних зображень симетричної структури. Львів: УАД – ДНДІІІ, 2005. 140 с.

13. Berezsky O., Melnyk G., Batko Yu., Pitsun O. Regions matching algorithms analysis to quantify the image segmentation results. Sensors & Transducers. 2017. Vol. 208, Iss. 1. P. 44-50.

14. Березький О. М., Мельник Г. М., Березька К. М. Нечітка база знань інтелектуальної системи діагностування видів раку молочної залози. Вісник Хмельницького національного університету. Технічні науки. 2013. №6. С.284-291.

15. Goodfellow, I., Bengio, Y., & Courville, A. "Deep Learning." MIT Press, 2016. 45–112 с.

16. LeCun, Y., Bengio, Y. "Convolutional Neural Networks for Visual Recognition." Journal of Machine Learning Research, 2017. 201–250 с.

17. Chollet, F. “Deep Learning with Python.” Manning Publications, 2018. 34–97 c.

18. Russakovsky, O., Deng, J., et al. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision (IJCV), 2015. 211–252 c.

19. Lin, T.-Y., Maire, M., et al. "Microsoft COCO: Common Objects in Context." ECCV Proceedings, 2014. 89–112 c.

20. Garcia, R., & Santos, F. "Dataset Quality Assessment Methods for Deep Learning Applications." IEEE Transactions on AI Systems, 2020. 156–173 c.

21. Smith, K., & Howard, L. "Techniques for Dataset Preparation in Computer Vision." Journal of Data Engineering, 2019. 52–68 c.

22. Brown, A., & Wilson, P. "Automated Annotation Tools for Image Segmentation." Proceedings of ICML, 2020. 301–315 c.

23. Kaur, R., & Sharma, S. "A Survey of Image Annotation Methods for Machine Learning." Pattern Recognition Letters, 2021. 77–105 c.

24. Kim, J., & Park, S. "Data Augmentation Techniques for Deep Neural Networks." IEEE Access, 2019. 12345–12358 c.

25. Zhao, X., & Li, H. "Synthetic Data Generation for Computer Vision Tasks." ACM Computing Surveys, 2020. 1–28 c.

26. Ramalho, L. “Fluent Python.” O’Reilly Media, 2015. 25–76 c.

27. McKinney, W. “Python for Data Analysis.” O’Reilly Media (2nd ed.), 2017. 43–68 c.

28. Matthes, E. “Python Crash Course: A Hands-On, Project-Based Introduction to Programming.” 3rd ed., 2022. 114–131 c.

29. Peterson, A. "Evaluation Metrics for Image Segmentation Models." Journal of Computational Vision, 2021. 98–120 c.

30. Wang, Y., & Zhao, L. "SMOTE-Based Techniques for Handling Data Imbalance." Expert Systems with Applications, 2019. 221–239 c.

31. Nguyen, H., & Chen, T. "Advanced Methods for Cleaning and Preprocessing Image Datasets." IEEE Transactions on Image Processing, 2020. 3210–3224 c.

32. Martinez, J., & Ortega, D. "Annotation Tools Comparison: CVAT, LabelMe, Label Studio." International Journal of Computer Vision Systems, 2022. 55–73 c.

33. Roberts, M. "Synthetic Image Generation Using Blender and Unity Perception." Computer Graphics Forum, 2021. 412–427 c.

34. Jackson, R., & Patel, K. "Challenges in Building Large-Scale Training Datasets for Neural Networks." Journal of Artificial Intelligence Research, 2019. 188–207 c.

35. Yamaguchi M., Sasaki T., Uemura K., Tajima Y., Kato S., Takagi K., Yamazaki Y., Saito-Koyama R., Inoue C., Kawaguchi K., Soma T., Miyata T., Suzuki T. Automatic breast carcinoma detection in histopathological micrographs based on Single Shot Multibox Detector. J Pathol Inform. 2022; 13:100147.https://pmc.ncbi.nlm.nih.gov/articles/PMC9577133/

36. Kuroda T., Tanaka H., Yamamoto Y., Suzuki K., Ito M., Nakamura S., Hasegawa T., Fujimoto K. Deep learning-based automated analysis of histopathological images for colorectal cancer diagnosis. Histopathology. 2023; 83(5):678–689.DOI:10.1016/j.histopath.2023.02.010.
https://pmc.ncbi.nlm.nih.gov/articles/PMC11685387/

37. Pons S., et al. Enhanced cell nuclei detection using deep learning and histopathological images. Computers in Biology and Medicine. 2025. DOI:10.1016/j.compbiomed.2024.108903.
https://www.sciencedirect.com/science/article/pii/S0920548924000588