DOI: 10.20998/2079-0023.2025.01.11
УДК 004.932.2:004.93'1

А. С. КОВАЛЕНКО, аспірант, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна; e-mail: anton.kovalenko@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0009-0336-871X.

ОПТИМІЗАЦІЯ ПРОЦЕСУ АНОТАЦІЇ ЗОБРАЖЕНЬ БІОЛОГІЧНИХ ОБ'ЄКТІВ МЕТОДАМИ КОМП'ЮТЕРНОГО ЗОРУ

У цьому дослідженні представлено підхід до автоматизованого формування анотованого набору даних, що містить зображення біологічних об'єктів, зокрема клітин. Запропонована методика базується на модифікованій структурі CRISP-DM, що адаптована до специфіки завдань комп'ютерного зору. Було розроблено послідовність етапів і кроків, спрямованих на ефективне виявлення та локалізацію об'єктів біологічного походження на мікроскопічних зображеннях. Процес охоплює попередню обробку зображень, яка включає бінаризацію, фільтрацію, корекцію яскравості та контрастності, а також усунення дефектів освітлення. Такі операції дають змогу покращити якість вхідних зображень і підвищити точність наступних етапів виявлення. Виявлені об'єкти автоматично локалізуються на основі морфологічного аналізу, після чого відбувається їх кластеризація за допомогою алгоритму *k*-середніх. Для групування враховуються такі ознаки, як розмір об'єкта та середнє значення кольору. Це дозволяє диференціювати різні типи клітин або структур за візуальними характеристиками. На основі локалізованих об'єктів автоматично формуються обмежувальні рамки, які зберігаються у вигляді координат у структурованому табличному форматі (.csv). Отриманий набір даних може використовуватись для навчання або тестування моделей глибокого навчання, зокрема тих, що вирішують задачі локалізації, класифікації або сегментації об'єктів на зображеннях. Запропонований підхід було апробовано на зображеннях мазків крові, що містять різні типи клітин. Усі обчислення виконано з використанням Python та таких бібліотек, як Pandas, NumPy, OpenCV і Matplotlib. Проведений аналіз точності виявлення та класифікації продемонстрував задовільні результати, що підтверджує доцільність використання розробленого конвеєра для створення анотованих наборів біологічних зображень.

Ключові слова: комп'ютерний зір, оптимізація, визначення країв, виділення контурів, сегментація об'єктів, машинне навчання, класифікація клітин крові.

Вступ. Локалізація біологічних об'єктів та структур – це процес ідентифікації та визначення їхнього просторового розташування з високою точністю. До таких структур належать клітини різних типів, білки, а також інші мікроскопічні об'єкти. Це завдання є надзвичайно важливим у сферах біології, медицини та матеріалознавства, оскільки точне позиціонування об'єктів є необхідною передумовою для проведення ефективного аналізу, діагностики та дослідницької роботи [1]. Сучасні методи локалізації охоплюють алгоритми сегментації, виявлення країв, а також підходи на основі штучного інтелекту, зокрема глибокі нейронні мережі.

Постановка задачі. Задача виявлення кількох об'єктів на зображенні полягає у знаходженні всіх об'єктів заданих класів на вхідному зображенні, визначенні їх просторових меж (у вигляді обмежувальних прямокутників), а також встановленні відповідної класової належності кожного об'єкта.

Математично задачу виявлення об'єктів на зображеннях можна формалізувати наступним чином: нехай задано набір зображень

$$X = \left\{ x_i \in \Re^{H \times W \times 3} \middle| i = 1, 2, ..., N \right\}, \quad (1)$$

де кожне зображення x_i характеризується своєю висотою H, шириною W та трьома кольоровими каналами (RGB).

Для кожного зображення x_i існує множина анотованих об'єктів:

$$Y_i = \left\{ \left(b_{ij}, c_{ij} \right) \middle| i = 1, 2, ..., M_i \right\}, \quad (2)$$

де M_i вказує на кількість об'єктів на зображенні x_i; $b_i = \left(x_{ij}, y_{ij}, w_{ij}, h_{ij} \right)$ – координати центра (або координати лівого верхнього кута – в залежності від типу представлення набора даних), ширина і висота обмежувальної рамки об'єкта j на зображенні i; $c_{ij} \{1, 2, ..., C\}$ – мітка (анотація) класу об'єкта, що належить до одного із C можливих класів.

Метою виявлення, локалізації та класифікації об'єктів на зображеннях є побудова функції

$$f : \Re^{H \times W \times 3} \to \left| \left(\hat{b}, \hat{c}, s \right) \right|, \quad (3)$$

яка для кожного зображення повертає скінченну множину кортежів $\left(\hat{b}, \hat{c}, s \right)$, де $\hat{b} \in \Re^4$ – прогнозована рамка об'єкта, $\hat{c} \in \{1, 2, ..., C\}$ – прогнозований клас об'єкта, $s_i \in [0, 1]$ – ступінь впевненості (confidence) моделі у цьому передбаченні.

Навчання моделі виконується шляхом мінімізації сумарної функції втрат

$$L = L_{cls} + \lambda \cdot L_{loc}, \quad (4)$$

де L_{cls} – функція втрат класифікації, L_{loc} – функція втрат регресії координат, $\lambda \in \Re^4$ – ваговий коефіцієнт, що регулює важливість локалізації.

Виявлення і локалізація об'єктів на зображеннях належать до сфери комп'ютерного зору та здебільшого розв'язуються із застосуванням штучних нейронних мереж. Водночас ефективне використання таких методів потребує наявності значного обсягу анотованих даних.

Створення анотованих наборів даних є основною

вимогою для розробки ефективного машинного навчання та систем ШІ. Якість і методологія анотації даних безпосередньо впливають на продуктивність моделі, що робить вибір підходу до анотації вирішальним для успіху проекту.

Підходи до анотування наборів даних. Наразі можна виділити декілька підходів до анотування наборів даних:

1) Ручне анотування. В цьому випадку анотування (розмітка) даних проводиться вручну спеціалістами в означеній області за допомогою таких інструментів як LabelImg, CVAT, LabelMe, Roboflow Annotate, VoTT, Label Studio тощо [2]. Це забезпечує високу точність і розуміння контексту. Саме такий процес анотування даних є звичайним для складних або деталізованих завдань, як розмітка медичних та біомедичних зображень. Але такий підхід вимагає багато часу, особливо для великих наборів даних. При цьому анотація вручну займає багато часу, людських та матеріальних ресурсів. Дослідження [3] показує, що близько 40% робочого часу спеціалісти з машинного навчання витрачають саме на анотування зображень, особливо для великомасштабних наборів даних. Вартість, пов'язана з маркуванням вручну, часто робить неможливими великі, повністю марковані навчальні набори, що робить їх придатним переважно для менших наборів даних або вузькоспеціалізованих доменів.

2) При автоматичному анотуванні використовуються моделі або алгоритми ШІ для створення міток без втручання людини. Такий підхід найкраще підходить для великих наборів даних, де анотування вручну не є ефективним. При цьому точність розмітки залежить від якості та відповідності використаних моделей та алгоритмів [4, 5].

3) Напівавтоматична анотація (Human-in-the-Loop – HITL) [6] поєднує попередню розмітку за допомогою методів штучного інтелекту із переглядом та виправленням людиною, збалансовуючи ефективність і якість. Цей підхід дозволяє системам ШІ пропонувати початкові мітки, які люди-анотатори потім уточнюють і підтверджують. Цей метод особливо ефективний для складних наборів даних, де однієї лише автоматизації недостатньо, а анотація вручну буде надто ресурсомісткою.

Анотація HITL гарантує, що моделі штучного інтелекту отримують точно позначені дані, що сприяє кращому прийняттю рішень і вищій точності. Це особливо корисно для медичних зображень, де помилки в анотаціях можуть призвести до критичних збоїв у роботі ШІ.

Основна частина. В цьому дослідженні пропонується дослідити можливості автоматичної розмітки неанотованих наборів даних для їх подальшого використання при детекції біологічних об'єктів на зображеннях.

Типовий робочий цикл розробки системи детекції та локалізації об'єктів на зображеннях, надано на рис. 1 [4].

На першому етапі відбувається збір зображень – без якісних і достатньо різноманітних зображень неможливо створити надійну модель. Після того, як зображення зібрано, вони проходять попередню обробку [7]. На цьому етапі дані очищуються, масштабуються, нормалізуються або іншим чином трансформуються (змінюється кольорова модель, розміри зображення тощо) з метою підготовки до подальшого аналізу.

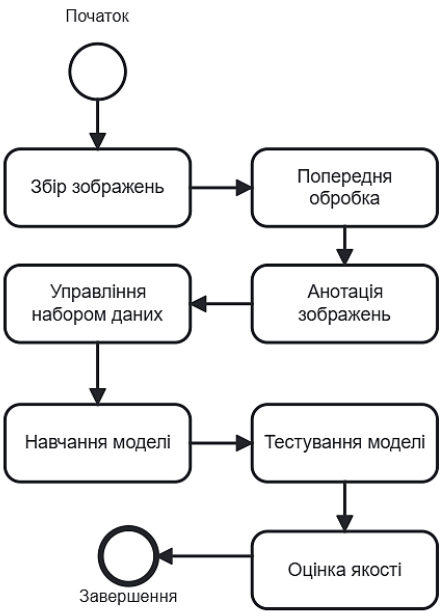

Рис. 1. Класичний пайплайн виявлення об'єктів

Наступним кроком після підготовки зображення, є їх анотація. Це трудомісткий, але надзвичайно важливий етап, адже він визначає, як саме модель буде «розуміти» вхідні дані. Анотовані зображення надходять до системи управління набором даних, де організовуються, зберігаються та контролюються з точки зору якості і доступності. Це дозволяє ефективно керувати різними підмножинами даних, як-от навчальним, валідаційним і тестовим наборами.

Далі йде навчання моделі. Тут алгоритм машинного навчання використовує підготовлені й анотовані зображення для формування здатності до розпізнавання або класифікації об'єктів. Після навчання відбувається тестування моделі – перевірка її продуктивності на нових, невідомих їй даних. Це дає змогу оцінити, наскільки добре модель узагальнює інформацію і як вона поводиться в умовах, наближених до реального застосування.

Завершальним кроком перед фінішем є оцінка якості. На цьому етапі аналізуються результати тестування, проводяться дослідження метрики точності, повноти, F1-оцінки тощо. Це дозволяє визначити, чи досягнуто бажаного рівня ефективності, чи потребує модель подальшого удосконалення.

Процес завершується тоді, коли всі етапи виконано, і модель досягла прийнятного рівня якості.

В дослідженні [8] було запропоновано підхід, що є покращенням пайплайну на рис. 1 та засновано на фреймворку CRISP-DM (рис. 2).

Процес починається з розуміння бізнес задач. На цьому етапі визначаються основні бізнес-цілі, формулюються завдання, які має вирішити модель, і встановлюються критерії успіху. Важливо не лише зрозу-

міти, що саме потрібно досягти, а й сформувати спільне бачення між аналітиками й представниками бізнесу щодо очікуваних результатів.

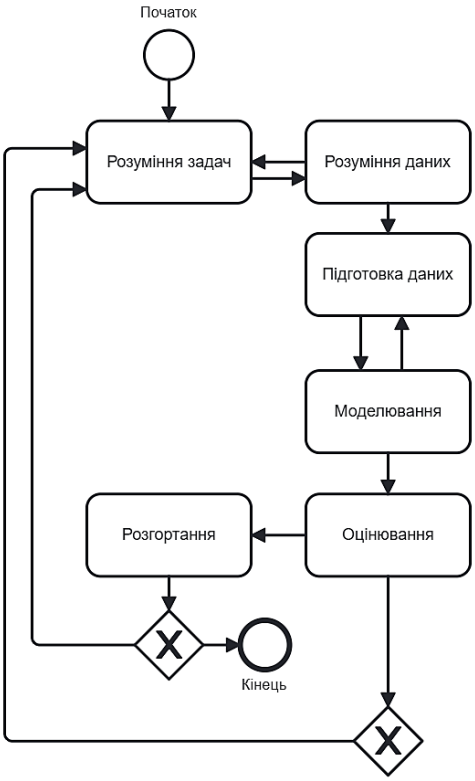

Рис. 2. Модифікований алгоритм виявлення об’єктів на зображенні

Наступним етапом є розуміння даних. Це важлива початкова фаза, на якій відбувається аналіз джерел інформації, вивчення структури, якості та змісту даних. Основна мета – усвідомити, з чим саме буде працювати аналітик або розробник, і які характеристики даних можуть вплинути на подальші етапи. Після цього розпочинається підготовка даних – процес очищення, трансформації, нормалізації та, за потреби, об’єднання даних з різних джерел.

Наступним кроком є моделювання – створення та налаштування алгоритмів машинного навчання, що будуть використовуватися для досягнення поставлених бізнес або дослідницьких цілей. Цей етап пов’язаний з експериментуванням, підбором гіперпараметрів і вибором найкращих моделей.

Далі слідує оцінювання – критично важливий етап, на якому здійснюється перевірка ефективності побудованої моделі. На основі обраних метрик (точність, F1-оцінка, середньоквадратична похибка тощо) визначається, наскільки добре модель справляється зі своїм завданням. Якщо якість моделі є незадовільною, відбувається повернення до етапу моделювання – для доопрацювання алгоритму, або навіть до підготовки даних, якщо джерело проблем криється глибше.

У разі, якщо якість моделі відповідає очікуванням і потребам, відбувається розгортання – впровадження моделі в реальне середовище, де вона починає працювати з новими даними. Після цього процес доходить до логічного завершення.

Таким чином, ця діаграма наочно демонструє не лише послідовність дій у процесі роботи з даними, а й циклічну природу моделювання – постійне вдосконалення і перевірку моделі до досягнення бажаного результату. Вона підкреслює важливість зворотного зв’язку на кожному етапі і необхідність ретельної роботи з даними на всіх рівнях.

Цей процес є ітеративним, що означає постійне повернення до попередніх етапів у разі потреби корекції або вдосконалення. Застосуємо цей підхід для виявлення, локалізації, класифікації об’єктів на зображеннях мікроскопічних знімків та створення на основі отриманих даних анотацій.

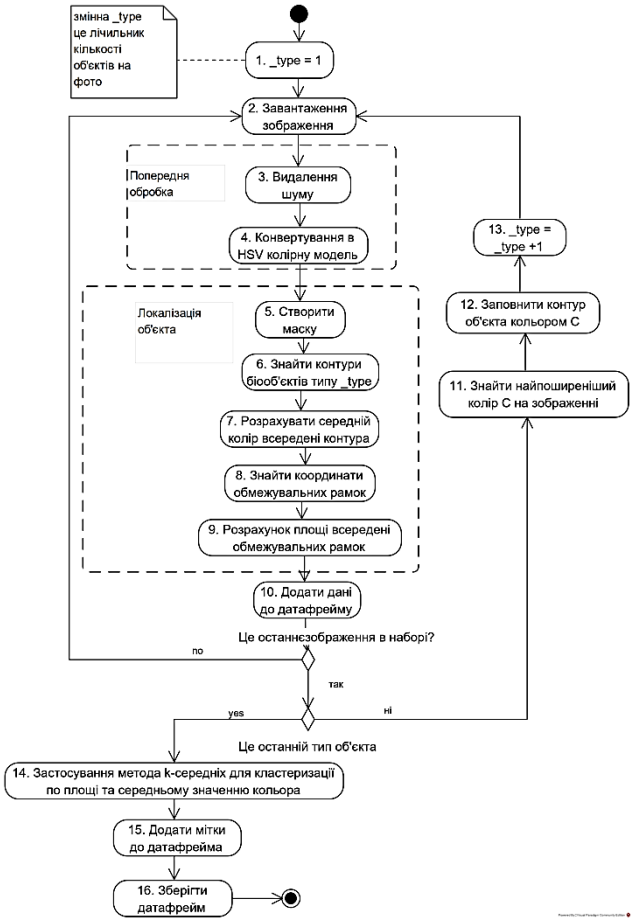

Рис. 3. Алгоритм автоматичного створення анотованого набору даних біологічних об’єктів із неанотованих зображень

Початковим етапом запропонованого підходу є ініціалізація змінної _type, яка виконує функцію лічильника для класифікації об’єктів на зображенні та задається початковим значенням 1. Далі здійснюється завантаження зображення та його попередня обробка, що включає фільтрацію шумів та перетворення зображення в колірну модель HSV з метою покращення аналізу колірних характеристик [9].

На етапі локалізації для кожного типу об’єкта створюється відповідна бінарна маска, яка дозволяє виділити контури біологічних структур заданої категорії. Після виявлення контурів обчислюється середнє значення кольору в межах кожного з них, що дає змогу більш точно диференціювати об’єкти за кольоровими

ознаками. Для кожного об'єкта також визначаються координати обмежувального прямокутника та розраховується його площа. Усі зібрані параметри вносяться до структури даних типу DataFrame.

Процедура повторюється для кожного зображення в наборі, поки не буде досягнуто останнього. Після обробки зображення перевіряється, чи завершено аналіз для всіх типів об'єктів. У разі продовження, на зображенні ідентифікується найбільш поширений колір, яким заповнюються раніше виявлені контури поточного типу. Змінна _type інкрементується і весь процес запускається знову для наступного типу біооб'єктів.

Після завершення обробки всіх типів об'єктів виконується класифікація за допомогою алгоритму *k*-means [10]. Кластеризація здійснюється на основі площі об'єктів та середнього кольору в межах їх контурів. Отримані мітки кластерів додаються до DataFrame, після чого зібрані дані зберігаються для подальшого використання.

Таким чином, описаний підхід представляє собою послідовний ітеративний процес виявлення, класифікації та аналізу об'єктів на мікроскопічних зображеннях із використанням методів обробки зображень та машинного навчання. Запропонована методика дозволяє ефективно опрацьовувати великі обсяги даних, забезпечуючи точність та узагальненість результатів.

У процесі класифікації розпізнаних об'єктів за ознаками кольору їхнього контуру та площі обмежувальної рамки була розглянута задача кластеризації методом *k*-середніх, що належить до класу алгоритмів, які мінімізують суму квадратів відстаней між об'єктами та центрами кластерів.

Нехай маємо множину об'єктів:

$$X = \{x_1, x_2, ..., x_n\}, \quad x_i \in \mathbb{R}^2, \tag{5}$$

де кожен об'єкт x_i описується двома ознаками: $x_i^{(1)}$ – числова характеристика кольору контуру; $x_i^{(2)}$ – площа обмежувальної рамки, яка обмежує розпізнаний об'єкт.

Потрібно знайти розбиття множини X на k непересічних підмножин (кластерів) $C = \{C_1, C_2, ..., C_k\}$, таких що мінімізується цільова функція втрат:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2, \tag{6}$$

де $\mu_j \in \mathbb{R}^2$ – центр (середнє значення) кластеру C_j; $\| \ \|$ – евклідова норма у двовимірному просторі ознак.

Іншими словами, мета алгоритму – знайти такі центри кластерів $\mu_1, ..., \mu_k$, які забезпечують мінімальну суму квадратів відстаней між кожним об'єктом та центром відповідного кластера.

Завдання кластеризації формулюється як задача оптимізації:

$$\min_{\{C_j\},\{\mu_j\}} \sum_{j=1}^{k} \sum_{x_i \in C_j} \left[\left(x_i^{(1)} - \mu_j^{(1)} \right)^2 + \left(x_i^{(2)} - \mu_j^{(2)} \right)^2 \right]. \tag{7}$$

Для забезпечення коректної роботи алгоритму та уникнення домінування однієї з ознак, значення даних кольору та площі перед кластеризацією нормалізуються.

Існує низка методів нормалізації даних, зокрема мінімаксне масштабування (також відоме як масштабування ознак або рескейлінг), нормалізація на основі Z-оцінки (стандартизація), десяткове масштабування та логарифмічне перетворення. У межах цього дослідження застосовано метод мінімаксної нормалізації.

Тобто для кожної ознаки $j = 1, 2$ нормалізоване значення обчислюється за формулою:

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - x_{\min}^{(j)}}{x_{\max}^{(j)} - x_{\min}^{(j)}}, \tag{8}$$

де $x_{\min}^{(j)} = \min_{1 \le i \le n} x_i^{(j)}$, $x_{\max}^{(j)} = \max_{1 \le i \le n} x_i^{(j)}$, $\tilde{x}_i^{(j)} \in [0,1]$ – нормалізоване значення ознаки.

Таким чином, всі значення кожної ознаки після перетворення потрапляють в інтервал [0,1], що гарантує однакову шкалу вимірювання та коректність розрахунку евклідової відстані в алгоритмі кластеризації.

Експеримент. Експеримент було проведено на основі набору даних [11]. Набір даних складається з 17 092 окремих зображень нормальних клітин, отриманих за допомогою аналізатора CellaVision DM96 в основній лабораторії госпітальної клініки Барселони. Усі зображення були зроблені в колірному просторі RGB. Зображення мають формат jpg і мають розмір 360×363 пікселів. Файли пов'язані з цим набором даних, ліцензованим згідно з міжнародною ліцензією Creative Commons Attribution 4.0 (CC BY 4.0).

Для фарбування мазків крові у мікроскопії використовується забарвлення Лейшмана. Зазвичай воно використовується для диференціації та ідентифікації лейкоцитів. В його основі є метанольна суміш «поліхромованого» метиленового синього та еозину, що забарвлює лейкоцити та тромбоцити у фіолетовий колір [12], тоді як еритроцити мають рожевий колір. Будемо використовувати цей факт для визначення меж та контурів об'єктів.

Для реалізації алгоритму було використано мову програмування Python і відповідні бібліотеки для обробки зображень, аналізу даних, візуалізації результатів та роботи з файловою системою: OpenCV-python, PIL, imutils, pandas, numpy, sklearn, os, glob, matplotlib, seaborn тощо.

Результат застосування алгоритму надано на рис. 4.

Аналіз результатів класифікації, зокрема вивчення розташування центрів кластерів, дає змогу інтерпретувати, які типи клітин відповідають кожному з отриманих кластерів. Було встановлено, що кластер, пов'язаний із тромбоцитами, містить значну кількість аномальних значень (викидів). У кластері, що відповідає лейкоцитам, також зафіксовано присутність викидів, проте їх частка є помітно меншою.

З метою усунення таких аномалій було повторно проведено аналіз вхідних даних. Розмір зображень, використаних у дослідженні, становить 360×363 пік-

селів. Попередній аналіз показав, що найменші з біологічно значущих об'єктів – тромбоцити мають розмір не менше 14×14 пікселів. Відтак, для підвищення точності класифікації, було вирішено не розглядати об'єкти, площа яких не перевищує 200 пікселів, оскільки вони, ймовірно, є шумами або артефактами зображення.

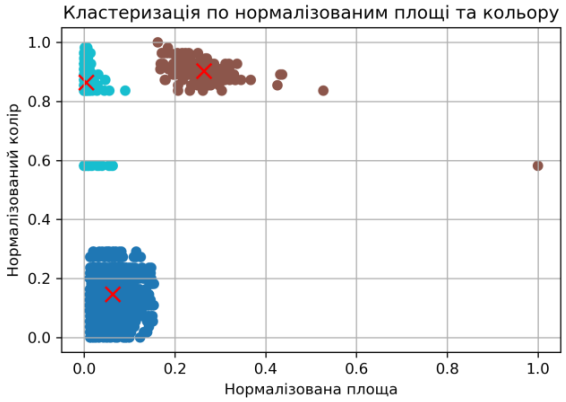

Рис. 4. Результат кластеризації клітин різних типів

Результатом виконання алгоритму став датафрейм (рис. 5) зі стовбцями file_name – назва файла без анотацій; xmin, ymin, xmax, ymax – координати верхнього лівого та правого нижнього кутів обмежувальних рамок; type – тип клітини; area – площа клітини; color – середній колір клітини; area_norm, color_norm – нормалізовані колір та площа клітини; cluster – номер кластера.

```
In [70]: 1 df_cluster.head(6)
Out[70]:
```

	file_name	xmin	ymin	xmax	ymax	type	area	color	area_norm	color_norm	cluster
0	BA_206112.jpg	227	307	248	327	platelet	420	162	0.006463	0.890909	2
1	BA_206112.jpg	141	298	165	322	platelet	576	162	0.008863	0.890909	2
2	BA_206112.jpg	87	251	103	268	platelet	272	162	0.004185	0.890909	2
3	BA_206112.jpg	123	104	239	258	WBC	17864	162	0.274877	0.890909	1
4	BA_206112.jpg	0	311	56	362	RBC	2856	122	0.043946	0.163636	0
5	BA_206112.jpg	260	291	339	362	RBC	5609	122	0.086307	0.163636	0

Рис. 5. Результуючий датафрейм

Для візуалізації та подальшої обробки результатів на зображеннях клітин наносяться обмежувальні рамки, які ілюструють просторове положення виявлених об'єктів. Також відображається кількість знайдених об'єктів на зображенні (рис. 6). Реалізація цього етапу здійснюється із використанням низки потужних бібліотек мови програмування Python, зокрема pandas, OpenCV, pillow та matplotlib.

Якість отриманих результатів була оцінена за допомогою відповідних метрик (табл. 1).

Таблиця 1 – Метрики отриманої моделі

	Точність	Влучність	Повнота	F1-оцінка
WBC	0,9814	1,0000	0,9814	0,9906
RBC	0,7895	0,8392	0,9302	0,8824
Platelets	0,8723	0,9767	0,8909	0,9318

```
D:\blood\train\BA_206112.jpg
type
RBC         10
WBC          1
platelet     3
dtype: int64
<class 'pandas.core.series.Series'>
```

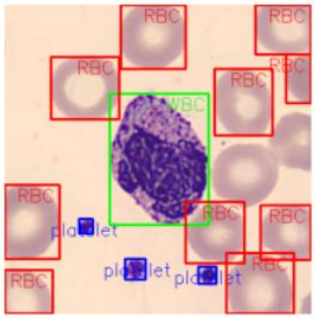

Рис. 6. Візуалізація результатів автоматичного виявлення та класифікації клітин крові

Висновки. У цій статті представлено підхід до автоматизованого створення анотованого набору даних біологічних об'єктів на зображеннях. Метод ґрунтується на адаптованій структурі CRISP-DM для задач обробки зображень і включає послідовність етапів виявлення та локалізації об'єктів. Розроблено конвеєр, що використовує методи бінаризації, фільтрації, корекції яскравості, контрастності та освітлення. Виявлені об'єкти групуються методом *k*-середніх за розміром і кольором. Підхід апробовано на прикладі клітин у мазку крові, продемонстровано його ефективність. Результатом є .csv-файл з координатами обмежувальних рамок для подальшого використання.

Список використаної літератури

1. Коваленко А. С., Северин В. П. Використання комп'ютерного зору в інтелектуальних системах, *XVI Міжнародна науково-практична конференція магістрантів та аспірантів «Теоретичні та практичні дослідження молодих науковців» (14-16 грудня 2022 року) : матеріали конференції.* Харків: НТУ «ХПІ», 2022. С. 38.
2. Lin J, Partick C. BakuFlow: A Streamlining Semi-Automatic Label Generation Tool. arXiv preprint arXiv:2506.09083, 2025. https://doi.org/10.48550/arXiv.2506.09083.
3. *2022 State of Data Science by Anaconda* [Електронний ресурс]. – URL: https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022 (дата звернення: 02.06.2025).
4. Kovalenko S. M., Kutsenko O. S., Kovalenko S. V., Kovalenko A. S. Approach to the automatic creation of an annotated dataset for the detection, localization and classification of blood cells in an image. *Radio Electronics, Computer Science, Control*. 2024. No. 1. P. 128–139. https://doi.org/10.15588/1607-3274-2024-1-12.
5. Kovalenko S., Kovalenko S., Mikhnova O., Kovalenko A., Pelikh D., Severin V., An Approach to Blood Cell Classification Based on Object Segmentation and Machine Learning *IEEE 4th KhPI Week on Advanced Technology (KhPIWeek)*. 2023. P. 1–6. DOI: 10.1109/KhPIWeek61412.2023.10312903.
6. Kutsenko A., Megel Y., Kovalenko S., Kovalenko S., Pelikh D., Rybalka A. Methods for Medical Images Contrast Measuring and Enhancement to Improve the Accuracy of Pathology Detection, *2022 XXXII International Scientific Symposium Metrology and Metrology Assurance (MMA)*, Sozopol, Bulgaria, 2022. P. 1–6. doi: 10.1109/MMA55579.2022.9993261.
7. Raman Thakur. *How Human-in-the-Loop is used in Data Annotation?* [Електронний ресурс]. – URL: https://www.labellerr.com/blog/why-is-hitl-needed-in-annotation/ (дата звернення: 02.06.2025).

8. Kovalenko S., Kovalenko S., Kutsenko A., Godlevskyi M., Severin V., Kovalenko A., Methodology for Creating Annotated Datasets of Biological Objects in Microscopic Images, *2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek)*, Kharkiv, Ukraine, 2024. P. 1–6, doi: 10.1109/KhPIWeek61434.2024.10878016.
9. Kutsenko A., Megel Y., Kovalenko S., Kovalenko S., Pelikh D., Rybalka A. Methods for Medical Images Contrast Measuring and Enhancement to Improve the Accuracy of Pathology Detection. *2022 XXXII International Scientific Symposium Metrology and Metrology Assurance (MMA)*, Sozopol, Bulgaria, 2022, pp. 1–6, doi: 10.1109/MMA55579.2022.9993261.
10. Yadav J., Sharma M., A Review of K-mean Algorithm. *Int. J. Eng. Trends Technol*. 2013. Vol. 4, iss. 7. p. 2972–2976.
11. Acevedo A., Merino A., Alférez S., Molina Á., Boldú L., Rodellar J. A dataset for microscopic peripheral blood cell images for development of automatic recognition systems. *Hospital Clinic de Barcelona*. https://doi.org/10.1016/j.dib.2020.105474.
12. Alam M. M, Islam M. T. Machine learning approach of automatic identification and counting of blood cells. *Healthcare Technology Letters*. 2019. Vol. 6, iss. 4. P. 103–108.

References (transliterated)

1. Kovalenko A. S., Severyn V. P. Vykorystannia kompiuternoho zoru v intelektualnykh systemakh [Using computer vision in intelligent systems]. *XVI Mizhnarodna naukovo-praktychna konferentsiia mahistrantiv ta aspirantiv «Teoretychni ta praktychni doslidzhennia molodykh naukovtsiv» (14-16 hrudnia 2022 roku) : materialy konferentsii* [XVI International Scientific and Practical Conference of Master's and PhD Students "Theoretical and Practical Research of Young Scientists" (December 14-16, 2022): conference materials]. Kharkiv, NTU "KhPI" Publ., 2022, p. 38. (In Ukr.).
2. Lin J, Partick C. BakuFlow: A Streamlining Semi-Automatic Label Generation Tool. arXiv preprint arXiv:2506.09083, 2025. https://doi.org/10.48550/arXiv.2506.09083.
3. *2022 State of Data Science by Anaconda.* URL: https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022 (accessed 02.06.2025).
4. Kovalenko S. M., Kutsenko O. S., Kovalenko S. V., Kovalenko A. S. Approach to the automatic creation of an annotated dataset for the detection, localization and classification of blood cells in an image. *Radio Electronics, Computer Science, Control*, 2024, no. 1, pp. 128–139. https://doi.org/10.15588/1607-3274-2024-1-12.
5. Kovalenko S., Kovalenko S., Mikhnova O., Kovalenko A., Pelikh D., Severin V., An Approach to Blood Cell Classification Based on Object Segmentation and Machine Learning *IEEE 4th KhPI Week on Advanced Technology (KhPIWeek)*. 2023. P. 1–6. DOI: 10.1109/KhPIWeek61412.2023.10312903.
6. Kutsenko A., Megel Y., Kovalenko S., Kovalenko S., Pelikh D., Rybalka A. Methods for Medical Images Contrast Measuring and Enhancement to Improve the Accuracy of Pathology Detection, *2022 XXXII International Scientific Symposium Metrology and Metrology Assurance (MMA)*, Sozopol, Bulgaria, 2022, pp. 1-6. doi: 10.1109/MMA55579.2022.9993261.
7. Raman Thakur. *How Human-in-the-Loop is used in Data Annotation?* URL: https://www.labellerr.com/blog/why-is-hitl-needed-in-annotation/ (accessed 02.06.2025).
8. Kovalenko S., Kovalenko S., Kutsenko A., Godlevskyi M., Severin V., Kovalenko A., Methodology for Creating Annotated Datasets of Biological Objects in Microscopic Images, *2024 IEEE 5th KhPI Week on Advanced Technology* (*KhPIWeek*), Kharkiv, Ukraine, 2024, pp. 1–6, doi: 10.1109/KhPIWeek61434.2024.10878016.
9. Kutsenko A., Megel Y., Kovalenko S., Kovalenko S., Pelikh D., Rybalka A. Methods for Medical Images Contrast Measuring and Enhancement to Improve the Accuracy of Pathology Detection. *2022 XXXII International Scientific Symposium Metrology and Metrology Assurance* (*MMA*), *Sozopol, Bulgaria*, 2022, pp. 1–6, doi: 10.1109/MMA55579.2022.9993261.
10. Yadav J., Sharma M., A Review of K-mean Algorithm. *Int. J. Eng. Trends Techn*ol. 2013, vol. 4, iss.7, pp. 2972–2976.
11. Acevedo A., Merino A., Alférez S., Molina Á., Boldú L., Rodellar J. A dataset for microscopic peripheral blood cell images for development of automatic recognition systems. *Hospital Clinic de Barcelona*. https://doi.org/10.1016/j.dib.2020.105474.
12. Alam M. M, Islam M. T. Machine learning approach of automatic identification and counting of blood cells. *Healthcare Technology Letters.* 2019, vol. 6, iss. 4, pp. 103–108.

Надійшла (received) 15.05.2025

UDC 004.932.2:004.93'1

A. S. KOVALENKO, National Technical University "Kharkiv Polytechnic Institute", PhD Student,
Kharkiv, Ukraine; e-mail: anton.kovalenko@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0009-0336-871X.

OPTIMIZATION OF THE ANNOTATION PROCESS FOR BIOLOGICAL OBJECT IMAGES USING COMPUTER VISION METHODS

This study presents an approach to the automated creation of an annotated dataset containing images of biological objects, particularly cells. The proposed methodology is based on a modified CRISP-DM framework, adapted to the specifics of computer vision tasks. A sequence of stages and steps has been developed to enable effective detection and localization of biological objects in microscopic images. The process involves preprocessing the images, including binarization, filtering, brightness and contrast adjustment, as well as correction of illumination artifacts. These operations help enhance the quality of the input images and improve the accuracy of subsequent detection steps. Detected objects are automatically localized based on morphological analysis, followed by clustering using the k-means algorithm. Grouping is based on features such as object size and mean color value, which allows for distinguishing between different types of cells or structures based on visual characteristics. Bounding boxes are automatically generated for the localized objects, and their coordinates are stored in a structured tabular format (.csv). The resulting dataset can be used to train or test deep learning models, particularly for tasks such as object localization, classification, or segmentation. The proposed approach was validated using images of blood smears containing various types of cells. All computations were carried out using the Python programming language and libraries such as Pandas, NumPy, OpenCV, and Matplotlib. The analysis of detection and classification accuracy demonstrated satisfactory results, confirming the feasibility of using the developed pipeline for automated generation of annotated biological image datasets.

Keywords: computer vision, optimization, edge detection, contour detection, object segmentation, machine learning, blood cell classification.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Коваленко Антон Сергійович / Kovalenko Anton Serhiyovych