

ПИСАРЧУК Олексій

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

<https://orcid.org/0000-0001-5271-0248>

e-mail: platinumPA2212@gmail.com

МІРОНОВ Юрій

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

<https://orcid.org/0000-0002-2291-5864>

e-mail: yuriymironov96@gmail.com

АЛГОРИТМ ВИЛУЧЕННЯ ОЗНАК З ЦИФРОВИХ ЗОБРАЖЕНЬ ХРОМОСОМ ЧЕРЕЗ СЕГМЕНТАЦІЮ

В статті запропоновано алгоритм вилучення ознак із об'єкту хромосоми на цифровому зображенні для задач кластеризації та ідентифікації, побудований на сегментації зображення. Особливістю даного алгоритму є незалежність результатів від форм об'єктів на зображеннях, що є перевагою через непередбачуваність геометричних ознак об'єктів розпізнавання. Проведено аналіз ефективності запропонованого алгоритму на прикладі K-means кластеризації у порівнянні з класичним вилученням ознак через нейромережу VGG16

Ключові слова: комп'ютерний зір, штучний інтелект, обробка зображень, кластеризація, класифікація, ідентифікація, машинне навчання, глибинне навчання, нейронні мережі.

PYSARCHUK Oleksii, MIRONOV Yurii

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

ALGORITHM FOR FEATURE EXTRACTION OF CHROMOSOME DIGITAL IMAGES THROUGH SEGMENTATION

This paper presents a novel algorithm for the extraction of significant features from digital images of chromosomal objects. The main goal of the algorithm is to facilitate effective clustering and identification of chromosomes based on their segmented image data. The proposed method relies on advanced image segmentation techniques that isolate chromosomal objects regardless of their geometric form, which often varies unpredictably due to biological and technical imaging factors.

A key advantage of this algorithm lies in its robustness against geometrical variability: it demonstrates consistent results even when applied to chromosome images of different shapes and contours. This adaptability makes the algorithm especially useful in real-world cytogenetic analysis, where image irregularities are common and can negatively impact the performance of classical neural networks or static feature extraction methods.

The effectiveness and precision of the developed algorithm have been rigorously evaluated through comparative analysis with the widely used convolutional neural network model VGG16. The results show that the proposed algorithm performs on par with, and in some cases even surpasses, VGG16 in terms of feature extraction quality and stability across variable datasets. This suggests that the method can be a valuable alternative or complementary approach in automated chromosome recognition systems, particularly where classical models may face limitations due to shape variability or insufficient training data.

The findings of this research contribute to the fields of digital cytogenetics, biomedical image processing, and intelligent diagnostic systems, highlighting a pathway toward more reliable chromosome analysis through tailored algorithmic approaches.

Keywords: computer vision, artificial intelligence, image processing, clustering, classification, identification, machine learning, deep learning, neural networks.

Стаття надійшла до редакції / Received 16.04.2025

Прийнята до друку / Accepted 11.05.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Застосування технологій комп'ютерного зору та штучного інтелекту в медицині є актуальною темою як з прикладної, так і з наукової точки зору. Кожна галузь медицини представляє унікальний набір особливостей та викликів при спробах подальшої автоматизації. Однією з таких галузей є репродуктивна медицина – сфера, що пов'язана з вивченням та діагностуванням хромосомних патологій. Відповідно, в її рамках присутня проблема автоматизації процесів розпізнавання хромосом та хромосомних аномалій. Вихідною точкою розпізнавання є медичні зображення хромосом пацієнтів. Їхньою особливістю є різноманітність форматів зображень та непередбачуваність геометричних форм, що робить таку автоматизацію складною науково-прикладною задачею.

Наявні програмні рішення [1-3] з обробки хромосомних зображень демонструють певний рівень автоматизації, однак на даному етапі є простір для покращення точності ідентифікації хромосом: поширеними є випадки, коли оператори змушені верифікувати результати ідентифікації і повторно проводити їх вручну. Ідентифікація хромосом у вищезгаданих програмних рішеннях проводиться на основі штучного інтелекту,

який було навченого на закритих наборах даних. Таким чином, результатом ідентифікації може стати лише факт подібності чи відмінності хромосоми пацієнту від хромосом, наявних в наборі даних.

З цього витікає наступний недолік: кінцевою метою процесу аналізу хромосом є виявлення хромосомних аномалій, тоді як програмні системи здатні ідентифікувати хромосоми за повною схожістю, але не можуть ідентифікувати конкретні аномалії, наявні в уражених хромосомах. Через відсутність автоматичної ідентифікації аномалій, оператори змушені вручну аналізувати хромосоми на наявність аномалій, і за типом та кількістю наявних аномалій приймати рішення про присутність у пацієнта хромосомних захворювань. Через це існує потреба в розробці альтернативного підходу до вилучення ознак з хромосомних зображень. Такий підхід мав би видобувати з зображень логічні ознаки, представляючи побудову кожної хромосоми як чітку структуру даних. В перспективі це дозволило б не тільки категоризувати здорові хромосоми, а також ідентифікувати уражені аномаліями зразки.

АНАЛІЗ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Задача виявлення хромосомних аномалій в даному дослідженні розглядається як задача кластеризації та ідентифікації зображень. У цьому напрямку проведено аналіз практики застосування методів машинного та глибинного навчання для автоматизації процесу виявлення хромосомних аномалій.

В якості базової основи відома математична модель процесів розпізнавання хромосомних патологій на основі генеративної змагальної мережі, запропонована Ю. Ву, І. Юе, К. Тана, В. Вангом та Т. Лу [4]. Даний підхід демонструє високу ефективність розпізнавання, але залежить від нейронної мережі, навченої на наборі даних $N=5500$ екземплярів. Також варто зазначити, що даний підхід вирішує проблему категоризації хромосом, але не вирішує проблему ідентифікації хромосомних аномалій. В публікації не згадані хромосоми з аномаліями, які становили б проблему при ідентифікації. Отже, даний підхід не вирішує проблему розпізнавання хромосомних аномалій.

В роботі Б. Важена [5] запропоновано метод класифікації хромосом за допомогою згорткової нейронної мережі та інтеграція його в робочий процес цитогенетичної лабораторії, що використовує вищезгадане ПЗ MetaSystems Ikaros. Запропонований метод показав ефективність в 98.8% успішних розпізнавань хромосом. Однак було зазначено, що тестування проходило на здорових хромосомах і розпізнавання хромосом з аномаліями становить значний виклик для систем на основі нейронних мереж.

Окрім цього, в публікації М. Мораді та С. Камаледіна Сетаредана [6] було запропоновано штучну нейронну мережу, що демонструє ефективність в 98.5% успішних розпізнавань. Однак, як і в [4], тестування обмежувалось здоровими хромосомами. Тим не менш, автори запропонували стратегію вилучення ознак через серединну вісь зображення, що була допрацьована в даному дослідженні.

Отже, пошук ефективного алгоритму вилучення ознак є частиною більшого дослідження, спрямованого на розробку підходу до ідентифікації хромосом, в тому числі уражених аномаліями. Такі хромосоми мають змінену структуру, що унеможливує ефективне застосування більшості традиційних підходів обробки цифрових зображень хромосом, як мінімум через значну варіабельність таких відхилень та обмежену доступність DataSet знімків, що їх репрезентують. Таким чином, актуальною є задача з розробки технологічного базису для аналізу хромосомних зображень, що дозволило б ідентифікувати в тому числі аномальні хромосоми. Першим кроком до розробки такого технологічного підґрунтя є алгоритм вилучення ознак, що відображає логічну структуру хромосоми.

ОПИС ЗАПРОПОНОВАНОГО АЛГОРИТМУ

Вихідними даними для алгоритму вилучення ознак були обрані цифрові зображення хромосом (рис. 1).



Рис. 1. Зображення хромосоми людини №3

В процесі неавтоматизованого аналізу хромосом, оператор порівнює їх зображення з «еталонним» зображенням хромосоми – ідеограмою. Для кожного типу хромосоми (1, 2, ... 22, X, Y) існує одна або декілька ідеограм (рис. 2).

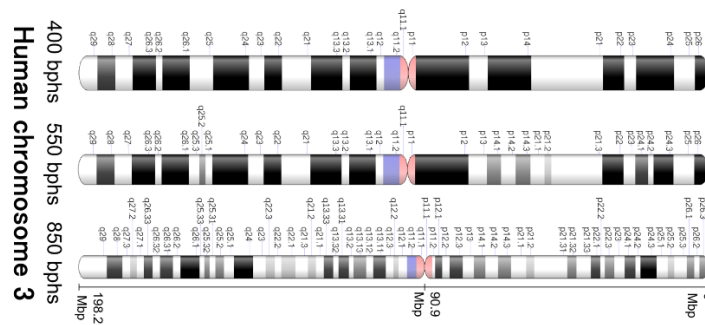


Рис. 2. Идеограми хромосоми №3

Процес аналізу та ідентифікації хромосоми полягає в порівнянні її структури зі структурою идеограми. Структура будь-якої хромосоми складається з певної послідовності сегментів, де у кожного сегмента є певна довжина та колір (чорний або білий). Таким чином, вилучення ознак з хромосомного зображення має полягати у вилученні інформації про таку структуру. Даний процес пов'язаний з певними викликами, оскільки зображення хромосом часто відрізняються низькою контрастністю.

ОПИС ЗАПРОПОНОВАНОГО АЛГОРИТМУ

Сутність алгоритму вилучення ознак полягає в сегментації візуального образу хромосоми за допомогою точок відносного екстремуму (мінімум і максимум) вздовж кольорового профілю. Межі між сегментами визначаються через медіанні значення між суміжними точками відносного екстремуму.

Для побудови кольорового профілю *COLPROF* (1) використовується використано вилучення серединної осі та подальше впорядковане отримання кольорів вздовж неї [7-8].

$$COLPROF = [px_1, \dots, px_N] \forall 0 \leq px_x \leq 255 \quad (1)$$

1. Масив *COLPROF* обробляється за допомогою фільтру Гауса з значенням $\sigma=20$;
2. За допомогою функції визначення точок відносного екстремуму [9] з урахуванням сусідніх 10 елементів вилучаються точки відносного екстремуму $EXTR_{max}$ та $EXTR_{min}$, де елементами є координати точок екстремуму вздовж *COLPROF*;
3. Точки екстремуму $EXTR_{max}$ та $EXTR_{min}$ об'єднуються в спільну колекцію *EXTR*, відсортовану по розташуванню вздовж *COLPROF*;
4. За допомогою алгоритму (2) між сусідніми точками колекції *EXTR* визначаються медіанні значення кольору;

$$MEDIAN_{x,x+1} = \frac{EXTR_x + EXTR_{x+1}}{2} - \left(\frac{EXTR_x + EXTR_{x+1}}{2} - \min(EXTR_x, EXTR_{x+1}) \right) * 0.2 \quad (2)$$

5. Вздовж серединної осі винаходяться координати точок з вищевказаними кольорами *MEDIAN*;
6. За допомогою точок відносного екстремуму *EXTR* та медіанних точок *MEDIAN* будується одновимірний масив *SEGFEAT*. За своєю структурою він є ідентичним масиву *COLPROF*, оскільки включає в себе тільки допустимі значення 0 та 255 (1). Однак, на відміну від *COLPROF*, *SEGFEAT* побудований за допомогою вилучення точок відносного екстремуму та медіанних значень. *SEGFEAT* є кінцевим результатом алгоритму – вилученими ознаками з зображення хромосоми (рис. 3).

ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ АЛГОРИТМУ СЕГМЕНТАЦІЇ

Оцінювання ефективності вищеписаних алгоритмів проводиться на датасеті $N=36$ зображень хромосом, пофарбованих GTG-методом. Хромосоми поділені на три групи: 12 хромосом №3, 12 хромосом №11 та 12 хромосом №13. Технологічно процеси моделювання та прототипування запропонованих рішень реалізовано за допомогою мови програмування Python із використанням бібліотек OpenCV, Scikit-image, Scipy, Numpy, Keras, Scikit-learn [9-13]. Оригінал коду доступний за посиланням [14].

Оцінювання ефективності проводиться в порівнянні з нейромережою для вилучення ознак VGG16, котру було обрано через її поширеність, універсальність та точність. В даному випадку вона представляє еталонний інструмент загального призначення, тому перевага запропонованого методу в результаті порівняння закликає проілюструвати їхню потенційну практичну цінність.

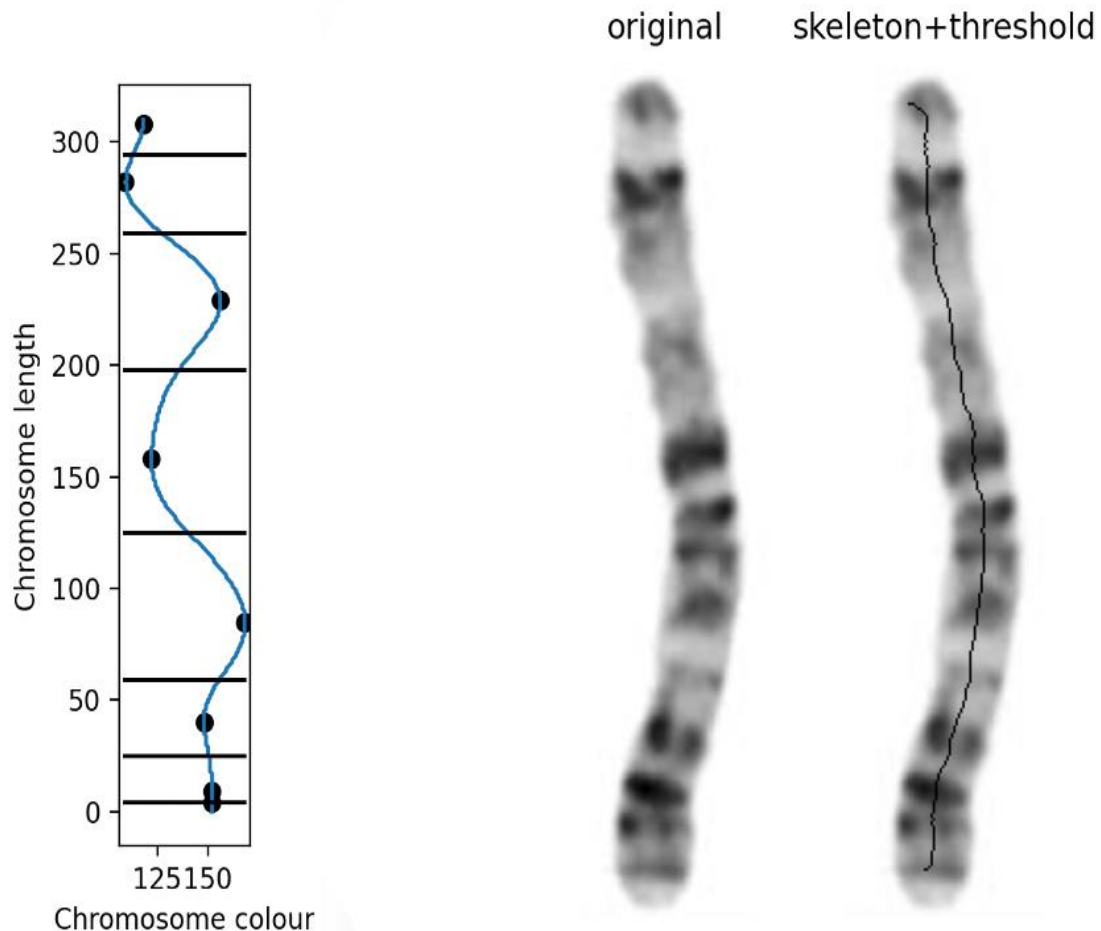


Рис. 3. Вилучення профілю чорно-білого зображення

Для оцінювання ефективності кластеризації використовуються індекс Ранда [15] та впорядкований індекс Ранда [16]. Індекс Ранда використовується для оцінки ефективності кластеризації, порівнюючи склад оригінальних кластерів (дано) і передбачуваних кластерів (результати роботи алгоритмів). Впорядкований індекс Ранда виконує ідентичну функцію, при цьому забезпечуючи менш успішний результат для менш впорядкованих кластерів. Результатом підрахунку обох індексів є числа в діапазоні $(-0.05, 1)$, де 1 – це повне співпадіння оригінальних і передбачуваних кластерів. В якості аргументів для підрахунку індексу Ранда використовуються оригінальні кластери з датасету, та результат K-means кластеризації за використанням описаного алгоритму, та VGG16 в якості референтного значення. Результати обох метрик наближаються до 1 при повному співпадінні кластерів.

Результати оцінювання ефективності зазначені в таблиці 1.

Таблиця 1

Оцінювання точності алгоритмів вилучення ознак

Алгоритм	Значення індексу Ранда	Значення впорядкованого індексу Ранда
VGG16	0.49523809523809526	0.02607630381519076
Запропонований алгоритм	0.8015873015873016	0.5548885949740564

Запропонований алгоритм показав кращий результат: +30% відносно VGG16. Це обумовлено тим, що запропонований алгоритм розроблений безпосередньо для роботи з зображеннями хромосом, та в процесі вилучення ознак спираються в тому числі на геометрію об'єктів на зображеннях, тоді як стандартний VGG16 не був здатен вилучити геометричні ознаки настільки ж успішно. Це додатково доводить актуальність подальших досліджень в сфері розпізнавання хромосомних зображень.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ

І ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ У ДАНОМУ НАПРЯМІ

В даній роботі було запропоновано алгоритм сегментації для вилучення ознак цифрових зображень хромосом з використанням у задачах кластеризації та ідентифікації. Алгоритм полягає у вилученні серединної

осі та подальший розбивці хромосоми на чорні та білі сегменти за допомогою визначення точок відносного екстремуму [9] і медіан між ними (2).

Даний алгоритм демонструє ефективність, що є вищою за стандартний алгоритм вилучення ознак: кластеризація з використанням запропонованих алгоритмів вилучення ознак дає до 30% покращення точності за індексом Ранда (таблиця 1). Таким чином, доцільним є подальші дослідження і розробка новітніх методів і моделей обробки цифрових зображень для підвищення точності ідентифікації та кластеризації хромосомних зображень.

References

1. Lucia Cytogenetics. (n.d.). Lucia Karyo. Retrieved Apr 30, 2025, from https://www.lucia.cz/en/products/lucia_karyo
2. Leica Biosystems. (n.d.). Clinical cytogenetics systems & instruments. Retrieved Apr 30, 2025, from <https://www.leicabiosystems.com/ihc-ish/fish-molecular-solutions/cytogenetics/>
3. MetaSystems. (n.d.). Ikaros karyotyping software. Retrieved Apr 30, 2025, from <https://metasystems-international.com/en/products/ikaros/>
4. Wu Y., Yue Y., Tan X., Wang W. & Lu T. (2018) End-To-End Chromosome Karyotyping with Data Augmentation Using GAN. 2018 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece. <https://doi.org/10.1109/ICIP.2018.8451041>
5. Vajen, B., Hänselmann, S., Lutterloh, F., Käfer, S., Espenkötter, J., Beening, A., Bogin, J., Schlegelberger, B., & Göhring, G. (2022). Classification of fluorescent R-Band metaphase chromosomes using a convolutional neural network is precise and fast in generating karyograms of hematologic neoplastic cells. *Cancer genetics*, 260-261, 23–29. <https://doi.org/10.1016/j.cancergen.2021.11.005>
6. Moradi, M., & Setarehdan, S.K. (2006). New features for automatic classification of human chromosomes: A feasibility study. *Pattern Recognit. Lett.*, 27, 19-28.
7. Pysarchuk, O., Mironov, Y. (2022). Chromosome Feature Extraction and Ideogram-Powered Chromosome Categorization. In: Hu, Z., Dychka, I., Petoukhov, S., He, M. (eds) *Advances in Computer Science for Engineering and Education. ICCSEEA 2022. Lecture Notes on Data Engineering and Communications Technologies*, vol 134. Springer, Cham. https://doi.org/10.1007/978-3-031-04812-8_36
8. Pysarchuk, O., Mironov, Y. (2025). Approaches to Feature Extraction for Chromosomal Data Processing. *Information society: technological, economic and technical aspects of formation*, <http://www.konferenciaonline.org.ua/us/article/id-2123/>
9. Scipy, (n.d.). Signal processing. Retrieved Apr 30, 2025, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.argrelextrema.html#argrelextrema>
10. OpenCV, (n.d.). Geometric Image Transformations. Retrieved Apr 30, 2025, from https://docs.opencv.org/4.x/d7/d1b/group_imgproc_misc.html#ga72b913f352e4a1b1b397736707afcde3
11. Numpy, (n.d.). Numpy documentation. Retrieved Apr 30, 2025, from <https://numpy.org/doc/stable/>
12. Keras, (n.d.). Keras documentation. Retrieved Apr 30, 2025, from <https://keras.io/api/>
13. Scikit-learn, (n.d.). Scikit-learn documentation. Retrieved Apr 30, 2025, from <https://scikit-learn.org/stable/api/index.html>
14. Mironov Y. (2025). Chromosome feature extraction. Retrieved Apr 30, 2025, from <https://github.com/yuriymironov96/chromosome-feature-extraction>
15. Scikit-learn, (n.d.). Rand score. Retrieved Apr 30, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html
16. Scikit-learn, (n.d.). Adjusted rand score. Retrieved Apr 30, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html#