УДК 004.932:616-076.5

В. О. Лещенко

ПОРІВНЯЛЬНИЙ АНАЛІЗ ВІДКРИТИХ АНОТОВАНИХ ДАТАСЕТІВ ДЛЯ РОЗРОБКИ СИСТЕМ АВТОМАТИЗОВАНОГО АНАЛІЗУ ЦИТОЛОГІЧНИХ ЗОБРАЖЕНЬ

Вінницький національний технічний університет

Анотація

У роботі здійснено порівняльний аналіз відкритих анотованих датасетів цитологічних зображень, які використовуються при розробці систем автоматизованої діагностики онкопатологій. Систематизовано обмеження наявних публічних ресурсів та наведено кількісні докази деградації моделей через доменний зсув. Обґрунтовано доцільність формування власної мультицентрової бази даних, репрезентативної для українського клінічного контексту.

Ключові слова: цитологічні зображення, анотовані датасети, доменний зсув, інстансна сегментація, CAD-системи, рак шийки матки.

Abstract

The paper presents a comparative analysis of publicly available annotated cytological image datasets used for developing automated diagnostic systems for oncopathologies. The limitations of existing public resources are systematized and quantitative evidence of model performance degradation due to domain shift is provided. The expediency of forming a proprietary multicentre dataset representative of the Ukrainian clinical context is substantiated.

Keywords: cytological images, annotated datasets, domain shift, instance segmentation, CAD systems, cervical cancer.

Вступ

Рак шийки матки (РШМ) залишається однією з провідних причин онкологічної смертності серед жінок в Україні. За даними Бюлетеня Національного канцер-реєстру України, частка РШМ у структурі онкозахворюваності жінок є однією з найвищих у Європі, при цьому близько половини випадків діагностуються на пізніх стадіях, а охоплення цільової групи скринінгом не перевищує 40 % [1, 2]. Перспективним інструментом масового скринінгу є системи автоматизованої комп'ютерної діагностики (CAD-системи) на основі методів глибокого навчання, ефективність яких безпосередньо залежить від якості та репрезентативності навчальних даних.

Метою роботи є порівняльний аналіз відкритих анотованих датасетів цитологічних зображень та оцінка їх придатності для розробки клінічно орієнтованих CAD-систем.

Основна частина

Класичним ядром відкритих ресурсів цервікальної цитології є набори одноклітинних зображень: Herlev (917 зображень, 7 класів) [3], SIPaKMeD (4049 ізольованих клітин, 5 класів) [4] та Mendeley LBC (963 зображення рідинної цитології у 4 класах за системою Bethesda) [5]. Набір CRIC містить 11 534 клінічно класифікованих клітин на 400 зображеннях конвенційного мазка [6]. Задачі попіксельної та інстансної сегментації забезпечують ресурси ISBI 2014/2015 (переважно синтетичні зображення) [7], CCEDD (686 зображень розміром 2048×1536) [8] та Cx22 (1320 зображень із 14 946 інстансами цитоплазми та ядер) [9]; рідкісним прикладом мультицентрового набору є CNSeg (1530 зразків від п'яти лікарень, понад 124 000 анотованих ядер) [10]. У 2024–2025 роках з'явилися ресурси, орієнтовані на роботу з повнослайдовими зображеннями: BMT (600 багатоклітинних зображень ThinPrep) [11], HiCervix із 40 229 клітин від 4496 WSI з ієрархічною трирівневою класифікацією [12] та RIVA, що публікує до чотирьох незалежних експертних анотацій [13]. Узагальнена характеристика ключових датасетів наведена в табл. 1.

Таблиця 1 — Порівняльна характеристика основних відкритих датасетів цервікальної цитології

Назва	Рік	Розмір	Тип	Рівень анотацій	Класи	К-сть лаб.
Herlev [3]	2005	917 одноклітинних	PAP конв.	pixel-level	7 (не Bethesda)	1
SIPaKMeD [4]	2018	4049 клітин / 966 кластерів	PAP	image + cell	5	1
Mendeley LBC [5]	2020	963 LBC	LBC, SurePath	image-level	4 (Bethesda)	3
CRIC [6]	2021	400 / 11 534 клітин	PAP конв.	image + cell	6 (Bethesda)	1
ISBI 2014/2015 [7]	2014–15	16 реальн. + 945 синт.	PAP, EDF	pixel-level	бінарна	1
CCEDD [8]	2022	686 / 33 614 патчів	PAP	edge / контур	бінарна	1
Cx22 [9]	2022	1320 / 14 946 інстансів	LBC	instance	бінарна	1
CNSeg [10]	2023	1530 / >124 000 ядер	LBC	nucleus instance	—	5
BMT [11]	2024	600 / 180 пац.	LBC, ThinPrep	image-level	3 (Bethesda)	1
HiCervix [12]	2024	40 229 / 4496 WSI	PAP	ієрархічна	29	мульти
RIVA [13]	2025	959 FOV / 26 158 ядер	PAP конв.	до 4 експертів	Bethesda	1

Аналіз показав, що, попри значне розширення кількості ресурсів за останні роки, відкриті датасети мають низку фундаментальних обмежень для клінічного впровадження. По-перше, доменний зсув між лабораторіями призводить до значної деградації ефективності моделей: при перенесенні навчених моделей сегментації з публічного набору на власний інституційний показник Panoptic Quality знижується з 0,82–0,88 до 0,68 [14]; модель, навчена на пулі SIPaKMeD, Herlev і CRIC, демонструє падіння F1-міри з 0,92 до 0,78 при тестуванні на реальному клінічному наборі [15]. По-друге, переважна більшість датасетів характеризується вираженим класовим дисбалансом — у наборі CRIC співвідношення нормальних клітин до плоскоклітинної карциноми сягає 42:1 [6]; високодиференційовані ураження (HSIL, ASC-H), залозисті атипії (AGC) та аденокарциномні клітини системно недопредставлені. По-третє, якість анотацій неоднорідна: значна частина датасетів анотована одним–трьома експертами без формального консенсусу, тоді як міжекспертна узгодженість при оцінці помірних дисплазій є помірною (κ = 0,45–0,58 для CIN2+) [16]; 8 з 12 розглянутих цервікальних датасетів зібрано в одній лабораторії, що додатково посилює доменний зсув. По-четверте, фіксуються артефакти власне датасетів: значна частка хибнонегативних об'єктів у CCEDD [9], майже повна синтетичність ISBI 2014 та недостатній просторовий розмір зображень Herlev для трансформерних архітектур.

Окремим обмеженням є технічна неоднорідність умов збору публічних наборів. Зображення отримані на різному обладнанні — Olympus BX53F (SIPaKMeD) [4], Zeiss AxioImager.Z2 (CRIC) [6], Nikon ELIPSE Ci (CCEDD) [8], робочий процес Hologic ThinPrep (BMT) [11]. У сукупності з регуляторними вимогами Технічного регламенту щодо медичних виробів № 753 [17], європейських MDR/IVDR та принципів FDA Good Machine Learning Practice [18] щодо репрезентативності даних цільової популяції, наявні відкриті ресурси є придатними для апробації архітектур, проте не забезпечують доказової бази для клінічної валідації CAD-системи в умовах національної системи охорони здоров'я України.

Висновки

Проведений порівняльний аналіз понад десяти відкритих анотованих датасетів цитологічних зображень засвідчив, що спільними обмеженнями є доменний зсув з падінням ключових метрик на 14–

25 п. в. при перенесенні моделей на реальні клінічні дані, виражений класовий дисбаланс та неоднорідність анотацій. З огляду на це наступним практичним етапом дослідження є формування власної бази цитологічних зображень, наближеної до реалій українських клінічних лабораторій, що забезпечить адекватну основу для розробки та валідації біотехнічної системи раннього виявлення онкопатології.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Бюлетень Національного канцер-реєстру України № 26. Рак в Україні, 2023–2024 / уклад. З. П. Федоренко та ін. К. : НКРУ, 2025. URL: http://www.ncru.inf.ua/publications/BULL_26/index.htm (дата звернення: 06.05.2026).

2. Volodko N., Chopyak V., Mazur Yu. Barriers to implementing cervical cancer screening in Ukraine: the path forward. Proceedings of the Shevchenko Scientific Society. Medical Sciences. 2025. Vol. 77, № 1. DOI: 10.25040/ntsh2025.01.01.

3. Jantzen J., Norup J., Dounias G., Bjerregaard B. Pap-smear Benchmark Data for Pattern Classification. Proc. NiSIS 2005. Albufeira : NiSIS, 2005. P. 1–9.

4. Plissiti M. E. et al. SIPaKMeD: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. 2018 IEEE ICIP. 2018. P. 3144–3148. DOI: 10.1109/ICIP.2018.8451588.

5. Hussain E., Mahanta L. B., Borah H., Das C. R. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data in Brief. 2020. Vol. 30. Art. 105589. DOI: 10.1016/j.dib.2020.105589.

6. Rezende M. T. et al. CRIC searchable image database as a public platform for conventional Pap smear cytology data. Scientific Data. 2021. Vol. 8. Art. 151. DOI: 10.1038/s41597-021-00933-8.

7. Lu Z., Carneiro G., Bradley A. P. et al. Evaluation of three algorithms for the segmentation of overlapping cervical cells. IEEE J. Biomed. Health Inform. 2017. Vol. 21, № 2. P. 441–450. DOI: 10.1109/JBHI.2016.2519686.

8. Liu J. et al. Local Label Point Correction for Edge Detection of Overlapping Cervical Cells. Frontiers in Neuroinformatics. 2022. Vol. 16. Art. 895290. DOI: 10.3389/fninf.2022.895290.

9. Liu G. et al. Cx22: A new publicly available dataset for deep learning-based segmentation of cervical cytology images. Computers in Biology and Medicine. 2022. Vol. 150. Art. 106194. DOI: 10.1016/j.compbiomed.2022.106194.

10. Zhao J. et al. CNSeg: A dataset for cervical nuclear segmentation. Computer Methods and Programs in Biomedicine. 2023. Vol. 241. Art. 107732. DOI: 10.1016/j.cmpb.2023.107732.

11. Campbell M. J. et al. BMT: A Cross-Validated ThinPrep Pap Cervical Cytology Dataset for Machine Learning Model Training and Validation. Scientific Data. 2024. DOI: 10.1038/s41597-024-04328-3.

12. Zhang X. et al. A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening. Scientific Data. 2025. Vol. 12. Art. 23. DOI: 10.1038/s41597-025-04374-5.

13. Perez Bianchi P. et al. RIVA: An Image Dataset of Conventional Pap Smear Cytology with Multiple Independent Annotations. Scientific Data. 2025. DOI: 10.1038/s41597-025-06280-2.

14. Mosquera-Zamudio A. et al. Deep-learning approaches for cervical cytology nuclei segmentation in whole slide images. Journal of Imaging. 2025. Vol. 11, № 5. Art. 137. DOI: 10.3390/jimaging11050137.

15. Ocampo-López-Escalera J. et al. Robust Cell-Level Classification for Liquid-Based Cervical Cytology Using Deep Transfer Learning: A Multi-Source Study Addressing Scanner-Induced Domain Shifts. Bioengineering. 2026. Vol. 13. Art. 289. DOI: 10.3390/bioengineering13030289.

16. Nayar R., Wilbur D. C. The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes. 3rd ed. Cham : Springer, 2015. DOI: 10.1007/978-3-319-11074-5.

17. Про затвердження Технічного регламенту щодо медичних виробів : Постанова Кабінету Міністрів України від 02.10.2013 № 753. URL: https://zakon.rada.gov.ua/laws/show/753-2013-п (дата звернення: 06.05.2026).

18. U. S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Silver Spring, MD : FDA, 2021. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device (дата звернення: 06.05.2026).

Лещенко Владислав Олександрович — аспірант групи G22-25a, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця, e-mail: lgl90pro@gmail.com

Науковий керівник: ***Заболотна Наталія Іванівна*** — д.т.н, доцент, професор кафедри біомедичної інженерії та оптикоелектронних систем, Вінницький національний технічний університет, Вінниця

Leshchenko Vladyslav O. — Faculty of Information Electronic Systems, Vinnytsia National Technical University, Vinnytsia, email: lgl90pro@gmail.com

Supervisor: ***Zabolotna Natalia I.*** — Professor of the Department of Biomedical Engineering and Optoelectronic Systems, Vinnytsia National Technical University, Vinnytsia.