УДК: 004.932.2

А. Ю. Марчук

МЕТОД ОБРОБКИ ОТОСКОПІЧНИХ ЗОБРАЖЕНЬ З ВИКОРИСТАННЯМ ОКТАВНОЇ ЗГОРТКИ ТА ТРАНСФОРМЕРІВ

Стаття присвячена розробці методу обробки отоскопічних зображень який поєднує октавні згортки для вилучення різночастотних ознак та візуальні трансформатори для моделювання глобального контексту. Пропонована гібридна архітектура об'єднує ефективний аналіз різночастотних ознак за допомогою октавних згорток та глобальне моделювання контексту за допомогою трансформерів. Октавний згортковий блок дозволяє ефективно обробляти зображення з широким діапазоном просторових частот, розділяючи карти ознак на високочастотну та низькочастотну групи. Це дозволяє знизити обчислювальні витрати, оскільки низькочастотна частина обробляється з меншою просторовою роздільною здатністю, при цьому зберігається обмін інформацією між потоками. Високочастотний потік фокусується на точних деталях, тоді як низькочастотний захоплює ширші, абстрактніші особливості, збагачуючи обидва представлення. Для моделювання глобального контексту використовується Swin Transformer, який забезпечує ієрархічну структуру ознак та лінійно-масштабоване захоплення глобального контексту, уникаючи обмежень традиційних трансформерів щодо високої роздільної здатності зображень. Проведено порівняння отриманих результатів із відомими SOTA-моделями та стандартними методами обробки зображень, такими як U-Net. Запропонований метод демонструє високу продуктивність та ефективність, особливо для задач, що вимагають обробки зображень високої роздільної здатності. Його обчислювальна складність є достатньо низькою завдяки роздільній обробці високочастотних та низькочастотних частин зображення, а також високою здатністю до збереження просторових деталей. Незважаючи на архітектурну складність та необхідність певних обчислювальних ресурсів для Swin-transformer, метод є перспективним для автоматизованої класифікації та діагностики патологій вуха.

Ключові слова: *октавна згортка, глибоке навчання, глобальний контекст, різночастотні ознаки, візуальні трансформери, згорткові нейронні мережі, метод, сегментація зображення, тензор, високочастотні карти, низькочастотні карти.*

Вступ

Сучасна оториноларингологія має певні проблеми з діагностикою захворювань вуха, такі як відсутність об'єктивності візуальної отоскопії, нестача кваліфікованих отологів у сільській місцевості, висока ціна дорогих інструментів тощо. Відомо, що традиційна діагностика, включаючи візуальне обстеження за допомогою отоскопа, варіюється в інтерпретації лікарів, особливо під час розпізнавання найраніших стадій захворювань (таких як мікроперфорації барабанної перетинки або початкова форма холестеатоми) [1]. У останні роки штучний інтелект (ШІ) став перспективним інструментом для автоматизації медичної діагностики, пропонуючи високу точність, стандартизацію результатів та можливість масштабного застосування [2]. Однак більшість сучасних алгоритмів на основі глибокого навчання, таких як звичайні згорткові нейронні мережі (CNN) або трансформери, мають суттєві недоліки: втрату деталей на дрібних аномаліях, високу обчислювальну складність та недостатню інтерпретованість результатів для лікарів [3]. **Метою цієї роботи** є розробка гібридного методу обробки отоскопічних зображень, який поєднує переваги октавних згорток та трансформерів для підвищення точності діагностики патологій вуха. Робота базується на публічному датасеті Roboflow Digital Otoscope (1400 зображень) [4], що охоплює основні категорії: гострий середній отит, холестеатому, перфорації барабанної перетинки та нормальний стан вуха.

Методика

Блок октавної згортки – основне нововведення, який змінює нормальний блок згортки [5]. Зображення з отоскопа мають як дрібні деталі (наприклад, судини, отвори, дрібні зміни текстури), так і великі основні форми (загальна форма барабанної перетинки, наявність рідин). Звичайні згортки можуть не охопити цей широкий діапазон інформації найкращим чином. Проста концепція блоку октавної згортки полягає в тому, що зображення можна розбити на частини з різними просторовими частотами [5]. Високі частоти зберігають дрібні деталі, чіткі зміни, краї та текстуру. Низькі частоти зберігають загальну структуру, великі форми, фонові області та плавні зміни. Згортка старого стилю обробляє всі ці частоти одночасно, що може бути повторюваним і не найкращим. Широкий діапазон просторових частот отоскопічних зображень роблять цей тип зображень придатним для октавної згортки. Основою цього дизайну моделі є перерозподіл карт FT на дві групи: «високочастотні» та «низькочастотні». Таким чином, низькочастотний компонент можна фільтрувати з нижчою просторовою роздільною здатністю, що значно зменшує обчислювальне навантаження. Водночас зберігається зв'язок між потоками, які обробляють різні ACC. Високочастотні особливості можуть бути передані в низькочастотний потік (зменшення вибірки), а низькочастотні — у високочастотний потік (збільшення вибірки). Це дозволяє деталям впливати на контекст і бути під його впливом [6].

Вхідне зображення розділяється на дві частини на основі гіперпараметра $\alpha \in [0,1]$, який визначає частку низькочастотних каналів. Зазвичай α обирається як 0.5, що означає рівномірний розподіл каналів. Високочастотні карти X_H містять $(1-\alpha)C$ каналів, де C – загальна кількість вхідних каналів, для кольорових зображень це 3 канали. Низькочастотні карти містять αC каналів і можуть бути відразу зменшені в просторовій роздільній здатності за допомогою згортки з кроком (strided convolution). Після розділення, в октавному згортковому блоці відбувається чотири основні операції згортки (потоки згортки). $X_{H\to H}$ – згортка що обробляє високочастотні вхідні дані та генерує нові високочастотні ознаки для вихідного високочастотного потоку, призначенням якої є збереження та уточнення детальних ознак. $X_{H\to L}$ – згортка що обробляє високочастотні вхідні дані та генерує ознаки, які передаються низкочастотному вихідному потоку. Після виконання згортки до результату виконується операція даунсемплінгу, наприклад усереднюючий пулінг, щоб привести його до просторової роздільної здатності низькочатотного потоку. $X_{L\to L}$ – обробляє низькочастотні вхідні дані, та генерує нові низькочастотні ознаки для вихідного низькочастотного потоку. $X_{L\to H}$ – обробляє низькочастотні вхідні дані, генерує ознаки, які передаються до високочастотного вихідного потоку [7]. До результату роботи необхідно застосувати білінійну інтерполяцію, щоб привести просторову роздільну здатність до високочастотного потоку. Кожна із цих згорток має власні незалежні набори вагових коефіцієнтів. На рис. 1 зображено структурну схему роботи октавної блоку октавної згортки.

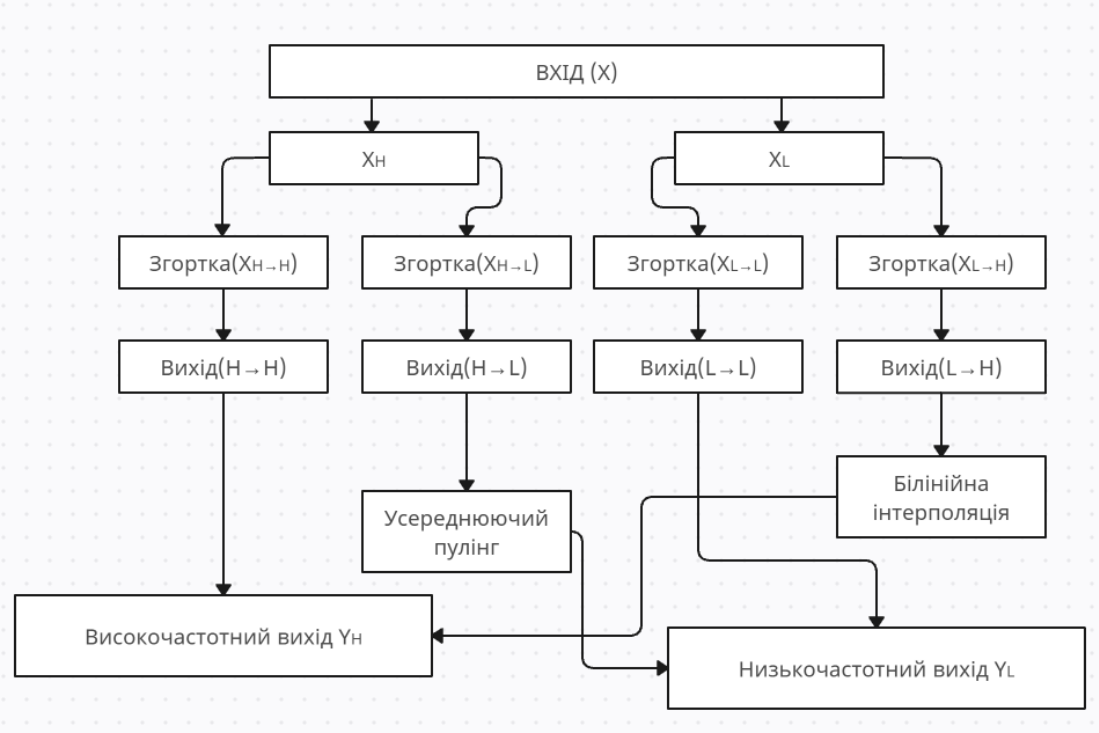


Рис. 1. Структурна схема блоку октавної згортки

Октавна згортка значно покращує ефективність обробки зображень, особливо для зображень із високою роздільною здатністю. Це досягається шляхом обробки низькочастотних карт зі зниженою просторовою роздільною здатністю, що, очевидно, зменшує операції множення-додавання. Таким чином, проміжне зберігання даних також оптимізоване для використання пам'яті. Наприклад, встановлення гіперпараметра α на 0,5 може зменшити складність обчислення приблизно на 20 – 30 % порівняно зі звичайною згорткою. Окрема та інтерактивна обробка різних частотних компонентів дозволяє моделі краще захоплювати як локальну детальну інформацію, так і глобальну контекстну інформацію зображень. Найдрібніші деталі утримуються в високочастотному потоці; доступ до більш загальної та семантичної інформації здійснюється з низькочастотних потоків. Перехресні зв'язок між цими двома потоками збагачують обидва представлення і таким чином ведуть до більш повного та точного розуміння візуальних даних. В результаті роботи октавної згортки формуються два вихідних тензори: високочастотний (Y_H) та низькочастотний (Y_L). У цій роботі октавна згортка була реалізована з використанням мови програмування Python та бібліотеки PyTorch [8].

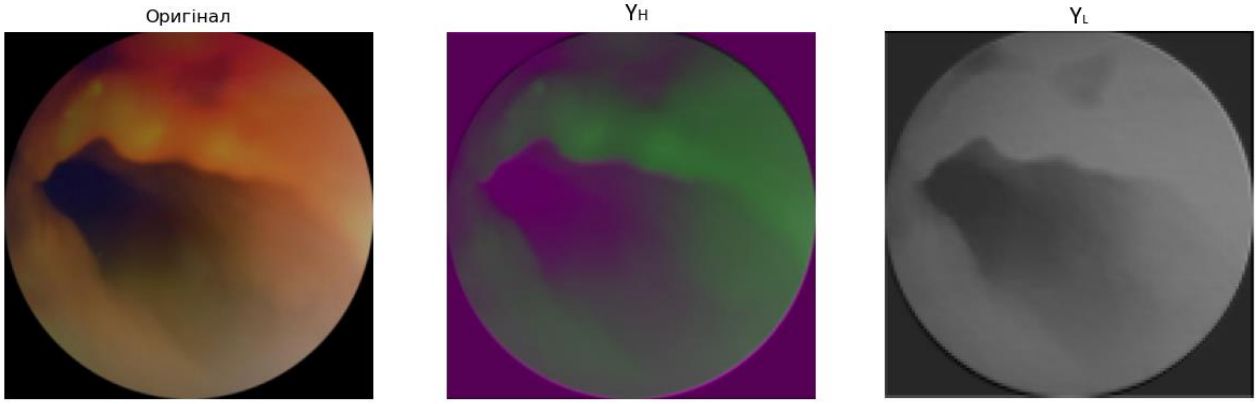


Рис. 2. Вихідні тензори YH та Y_L

Тензори Y_H та Y_L, передаються до трансформера, основними завданням якого є виявлення та локалізація об'єктів на зображенні за допомогою обмежувальних рамок та сегментація

зображень — розділення зображення на області пікселів, що відповідають різним об'єктам (семантична сегментація). В роботі пропонується використання Swin Transformer [9] – це покращена реалізація Vision Transformer (ViT) [10], який позбавлений обмежень традиційних трансформерів під час роботи із зображеннями високої роздільної здатності. Swin Transformer створює ієрархічну організацію функцій. Обробка відбувається в чотири етапи, починаючи з поділу зображення на частини 4x4 пікселя. Кожен патч перетворюється у вектор (токен) через лінійний шар. На кожному етапі обробки до патчів застосовується принцип локальної уваги (window-based self-attention) та зміщення вікон (shifted windows) [10]. Після кожного етапу відбувається об'єднання патчів у блоки з меншою просторовою роздільною здатністю, але більшою кількістю каналів, що дозволяє моделювати об'єкти на різних масштабах. Локальна увага у вікнах (Window Self-Attention) – замість того, щоб обчислювати самоувагу між усіма токенами зображення (що має квадратичну складність за розміром зображення), Swin Transformer розділяє зображення на непересічні "вікна" (наприклад, 7x7 патчів) і обчислює самоувагу лише всередині кожного вікна. Це значно зменшує обчислювальну складність до лінійної залежності від розміру зображення. Якщо для стандартних трансформерів складність обчислення складає $O(N^2)$, для Swin transformer складність обчислення визначається як $O(M^2 * N / M^2) = O(N)$, де M – розмір вікна [11]. Через фіксовані вікна трансформери втрачають зв'язки між вікнами, для вирішення цієї проблеми після кожного кроку етапу обробки зображення вікна зміщуються на половину розміру, наприклад на (3,3) для вікна 7x7. Це забезпечує зв'язок між сусідніми вікнами без збільшення складності обчислень.

Виходи Y_H та Y_L октавної згортки виходи подаються на вхід блоку трансформера і опрацьовуються. В результаті роботи отримується вихідне зображення з чіткішими контурами, текстурами. Результат роботи методу представлено на рис. 3.

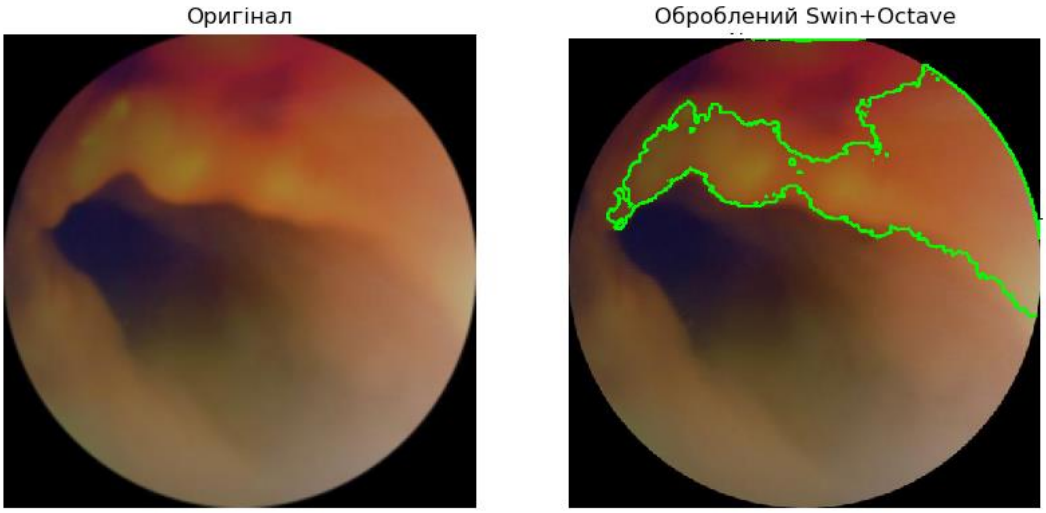
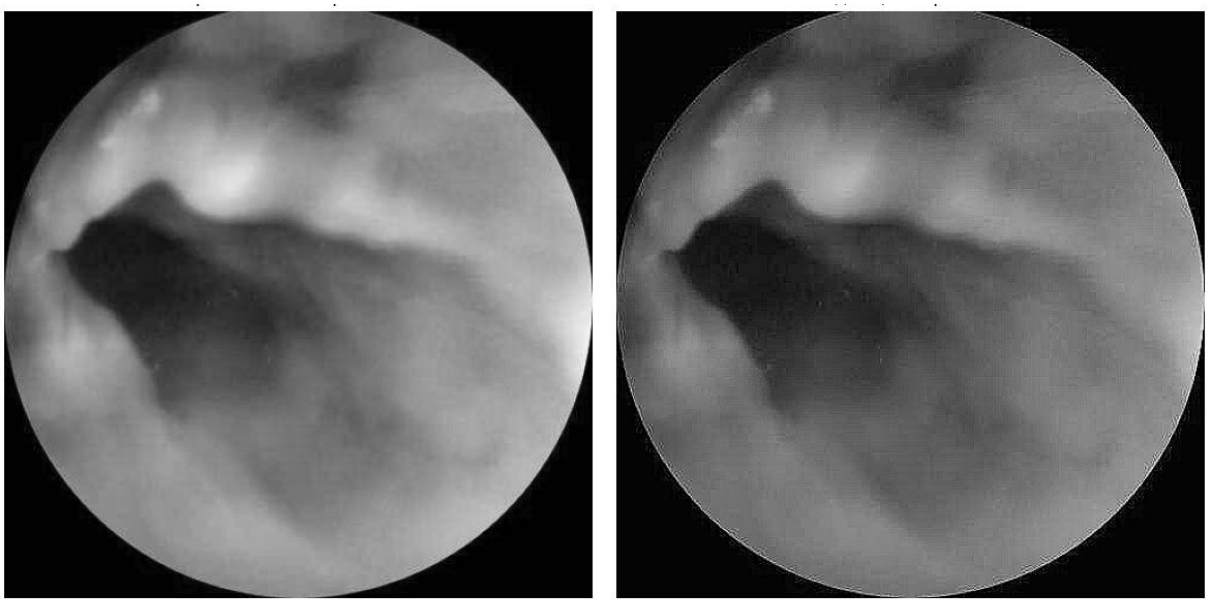


Рис. 3. Результат виконання методу на основі октавної згортки та трансформера

Перспективи клінічного застосування цього методу можливі як в телемедицині, для віддаленої діагностики та попередньої оцінки перед консультацією спеціаліста, так і для безпосереднього використання практикуючими лікарями як допоміжний інструмент, який зменшить навантаження на лікарів. Також метод може бути інтегрований в згорткову нейронну мережу, що в перспективі може покращити результати діагностики. Розроблений метод можливо використовувати як для класифікації зображень за ступенем важкості захворювання або спостереженням стадій розвитку захворювань.

Порівняння з відомими методами

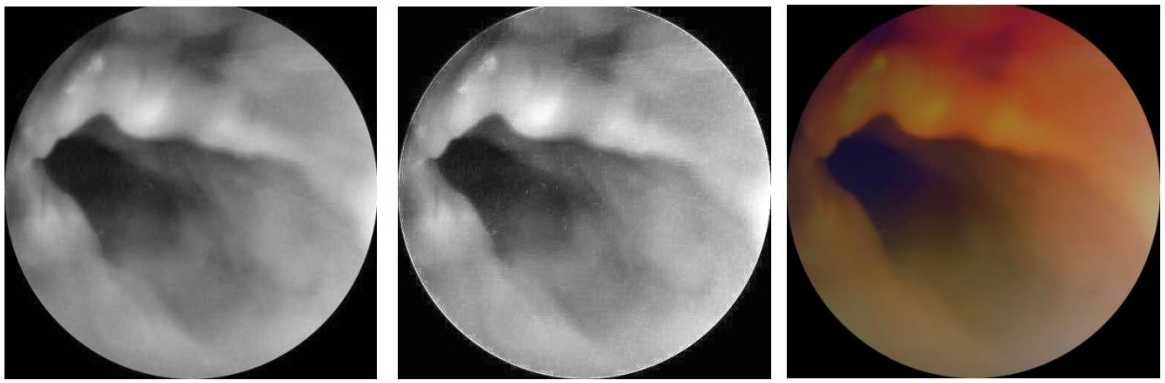
Для порівняння результатів роботи описаного метода було використано стандартний метод обробки зображень з використанням стандартної реалізації згортки. Стандартну згортку було реалізовано мовою програмування Python, з використанням бібліотеки PyTorch. Спочатку зображення конвертується у відтінки сірого, для підвищення контрасту використовується метод обмеженого контрастом адаптивного вирівнювання гістограми (CLAHE) [12]. Для спрощення процесу обробки та оптимізації обчислень до зображення застосовано розмиття (Gaus blur) [13]. Після попередньої обробки та підвищення контрасту зображення використано фільтр Собеля [14], для виділення країв значущих областей на зображенні. Виявлення деталей реалізовано оператором Лапласа [15]. Результат використання методу наведено на рис. 4.



Оригінальне зображення Зображення після згортки

Рис. 4. Результат виконання згортки на тестовому зображенні

До отриманих результатів було застосовано ViT transformer [16], який покладається на принцип глобальної самоуваги. Принцип глобальної самоуваги означає, що кожна частина зображення взаємодіє та обчислює свою "увагу" до кожної іншої частини у всій послідовності зображення. Це дозволяє моделі захоплювати глобальні залежності та розуміти контекст між віддаленими частинами зображення, чого не робить традиційна згортка, яка обробляє всі частоти разом [17]. Такий алгоритм накладає обмеження на використання за рахунок квадратичної обчислювальної складності, що в свою чергу змушує обмежувати роздільну здатність зображень вхідних зображень. Комбінація звичайної згортки з попередньою обробкою та ViT трансформера може втрачати мілкі деталі на зображенні, такі як мікроперфорації барабанної перетинки, через складність алгоритму та необхідність попередньої обробки зображення. На рис. 5 продемонстровано результат обробки тестового зображення з використанням звичайної згортки та ViT трансформера.



Оригінальне зображення | Результат виконання операції згортки | Результат роботи ViT

Рис. 5. Результат роботи згортки з використанням візуального трансформера

Також для порівняння було використано згорткову нейромережу архітектури U-NET [18], яка навчання якої було проведено на тестовових даних (~100 отоскопічних зображень) розміром 500x500 пікселів, та відбувалось в 50 епохах. Апаратне забезпечення представляло собою персональний комп'ютер з графічним процесором на 16 гігабайт оперативної пам'яті, процес навчання зайняв близько 20 хвилин. Для навчання було попередньо створено маски зображень автоматичним методом, за допомогою відкритої бібліотеки Open-CV [19]. Принцип роботи архітектури U-NET складається з двох основних частин: Encoder – який відповідає за виділення абстрактних ознак із зображення та реалізований із двох згорткових блоків, випрямленого лінійного блоку (ReLU) [20] та операції підвибірки (Max Pooling); Decoder – відповідає за відновлення просторової інформації для точної локалізації об'єктів, складається з блоку Upsampling [20] та принципом Skip-connections для передачі деталей, що запобігає втраті інформації. Результат обробки тестового зображення за допомогою згорткової нейромережі U-NET представлено на рис. 6. Покращити результат роботи нейромережі можливо за допомогою збільшення вибірки навчання та кількості епох при яких навчатиметься нейромережа.

Оригінальне зображення | Маска зображення | Контури значущих областей

Рис. 6. Результат роботи згорткової нейромережі U-NET

Оцінку роботи розглянутих методів проводилась за допомогою метрики Intersection over Union (IoU) [21]. IoU – це фундаментальна метрика, яка використовується в задачах комп'ютерного зору, особливо в таких як знаходження об'єктів чи сегментація зображення та визначає отримані результати виділення об'єкта співпадають із істинними межами об'єкта. Розраховується як співвідношення площі перетину об'єктів до їх загальної площі.

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \quad (1)$$

де A – отримана площа, B – істинні межі об'єкта.

Порівняння проводилась на зображеннях які містять гострий отит та перфорації барабанної перетинки. Результат наведений в таблиці 1.

Таблиця 1

Результати порівняння методів

Метод	IoU (Гострий отит)	IoU (Перфорація)	IoU (Середнє значення)
Згортка + ViT	0.71	0.75	0.73
U-Net	0.68	0.72	0.70
Октавна згортка + Swin	0.76	0.79	0.78

Висновки

В роботі було запропоновано метод обробки отоскопічних зображень на базі октавної згортки з використанням Swin-transformer. Метод об'єднує ефективну обробку ознак на різних частотах за допомогою октавної згортки та ієрархічне, лінійно-масштабоване захоплення глобального контексту за допомогою Swin Transformer. Ця комбінація обіцяє високу продуктивність та ефективність, особливо для задач, що вимагають обробки зображень високої роздільної здатності. Обчислювальна складність представленого методу достатньо низька завдяки принципу роздільної обробки високочастотних та низькочастотних частин зображення. Відрізняється високою здатністю до збереження просторових деталей. Одним із недоліків запропонованого методу можна назвати його архітектурну складність та складність реалізації, яка полягає в складному процесі налагодження та певних обчислювальних ресурсів через використання Swin-transformer. Клінічне застосування описаного в роботі методу можливе як допомога лікарям для діагностики захворювань вуха, таких як гострий отит чи перфорації барабанних перетинок. За рахунок зменшення навантаження, такий метод є більш гнучким в плані апаратного забезпечення, та може бути використаний як в телемедицині, для віддаленої діагностики пацієнтів, так і безпосередньо в лікарських установах в складі системи підтримки прийняття рішень.

СПИСОК ЛІТЕРАТУРИ

1. Chen Z., Cao Z., Dong R. Deep Learning in Otoendoscopy for Ear Disease Diagnosis: A Review. 2023. *Journal of Medical Imaging and Health Informatics*. 2023. №13 (1). P. 1–10.

2. Huang S. Challenges and Opportunities in AI-Assisted Otoscopy. *IEEE Transactions on Biomedical Engineering.* 2022. №69. P. 2150–2160.

3. Zhang J. Early Stage Disease Detection in Otoscopy: A Deep Learning Approach. *Artificial Intelligence in Medicine*. 2021. №118. P. 5–12

4. LN-MC Concept v1.0. URL: https://universe.roboflow.com/otoscope/digital-otoscope/dataset/1.

5. Chen Y., Octave Convolutions for Visual Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019. P. 2831–2840.

6. Liu Z. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. P. 10014–10018.

7. Dosovitskiy A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. 2021. P. 4–9.

8. Paszke A. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NeurIPS*). 2019. P. 32.

9. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015. P. 234–241.

10. Open Source Computer Vision Library. URL: https://opencv.org/.

11. Pizer S. M. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing.* 1987. Vol. 39, № 3. P. 355–368.

12. Canny J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1986. Vol. 8, № 6. P. 679–698.

13. Pre-Processing Images of Public Signage for OCR Conversion / K. Amber et al. *Journal of Signal and Information Processing*. 2018. Vol. 10, № 1. P. 5–9.

14. Marr D., Hildreth E. Theory of Edge Detection. Proceedings of the Royal Society of London. *Series B,*

Biological Sciences. 1980. P. 187–217.

15. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010. P. 249–256.

16. Zhou B. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. P. 2921–2929.

17. Long J., Shelhamer E., Darrell T. Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015. P. 3431–3440.

18. Milletari F., Navab N., Ahmadi S. A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *International Conference on 3D Vision*. 2016. P. 565–571

19. Oktay E. Attention U-Net: Learning Where to Look for the Pancreas. *International Conference on Medical Image Computing and Computer Assisted Intervention*. 2018. P. 105–113.

20. Zeiler M. D., Fergus R. Visualizing and Understanding Convolutional Networks. *European Conference on Computer Vision.* 2014. P. 818–833.

21. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon et al. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR*). 2016. P. 779–788.

Стаття надійшла до редакції 10.06.2025.
Стаття пройшла рецензування 28.06.2025.

Марчук Андрій Юрійович – аспірант, факультет інформаційних електронних систем, e-mail: andriu4934@gmail.com.
Вінницький національний технічний університет.