УДК 004.048+004.94

О.Р. ЯРЕМА
Львівський національний університет імені Івана Франка
С.А. БАБІЧЕВ
Університет Яна Евангелиста Пуркінє в Усті на Лабі, Чехія;
Херсонський державний університет

ГІБРИДНА МОДЕЛЬ АНАЛІЗУ ЕКСПРЕСІЇ ГЕНІВ ІЗ ЗАСТОСУВАННЯМ АНСАМБЛЕВИХ СТРАТЕГІЙ КЛАСТЕРИЗАЦІЇ ТА КЛАСИФІКАЦІЇ ДЛЯ ДІАГНОСТИКИ СТАНУ СКЛАДНИХ СИСТЕМ

Аналіз даних експресії генів є одним із ключових інструментів сучасної біоінформатики та комп'ютерних наук, оскільки дає змогу ідентифікувати біомаркери, формувати молекулярні профілі та підтримувати процеси діагностики онкологічних захворювань. Актуальність цього напряму пояснюється потребою у методах, здатних адекватно обробляти великі транскриптомні дані з високою розмірністю, різнорідністю та шумами. Такі особливості суттєво ускладнюють використання традиційних методів кластеризації та класифікації, унаслідок чого вони втрачають інтерпретованість і точність. У статті запропоновано гібридну модель, що поєднує алгоритм самоорганізованих дерев (Self-Organizing Tree Algorithm, SOTA) з консенсусними стратегіями агломеративної та спектральної кластеризації. Запропонований ансамблевий підхід дає змогу формувати узгоджені та інформативні кластери профілів експресії генів, що зменшує вплив локальних аномалій і підвищує надійність отриманих результатів. Побудовані кластери використовувалися як нові ознаки для класифікаційної моделі на основі алгоритму Random Forest, гіперпараметри якого було оптимізовано за допомогою байєсівських методів у поєднанні зі стекінгом. Моделювання виконано на великомасштабній матриці експресії, що включала понад 6 тис біологічних зразків і 18 тис генів, охоплюючи 14 класів зразків. Результати показали, що спектральний консенсусний варіант SOTA стабільно забезпечує найкращі значення внутрішніх індексів якості кластеризації та найвищу точність класифікації. Зокрема, у конфігураціях із трьома-п'ятьма кластерами досягнуто 100% точності та F1-міри, що підтверджує діагностичну значущість виділених груп генів. Розроблений ансамблевий конвеєр являє собою масштабований і стандартизований інструмент аналізу транскриптомних даних, який може бути інтегрований у системи підтримки прийняття рішень для ранньої діагностики стану складних систем.

Ключові слова: *дані експресії генів, метрики близькості, гібридна модель, кластеризація, класифікація, персоналізована медицина, гібридизація.*

O.R. YAREMA
Ivan Franko National University of Lviv
S.A. BABICHEV
Evangelista Purkyně University in Ústi nad Labem, Czech Republic;
Kherson State University

HYBRID MODEL FOR GENE EXPRESSION ANALYSIS USING ENSEMBLE CLUSTERING AND CLASSIFICATION STRATEGIES FOR DIAGNOSING COMPLEX SYSTEMS

Gene expression data analysis is one of the key tools of modern bioinformatics and computer science, as it enables the identification of biomarkers, the formation of molecular profiles, and the support of cancer diagnostics. The relevance of this field is explained by the need for methods capable of adequately processing large-scale transcriptomic data characterized by high dimensionality, heterogeneity, and noise. Such features significantly complicate the use of traditional clustering and classification methods, leading to reduced interpretability and accuracy. This article proposes a hybrid model that combines the Self-Organizing Tree Algorithm (SOTA) with consensus strategies of agglomerative and spectral clustering. The proposed ensemble approach enables the formation of consistent and informative clusters of gene expression profiles, which reduces the impact of local anomalies and increases the reliability of the results obtained. The constructed clusters were used as new features for a classification model based on the Random Forest algorithm, whose hyperparameters were optimized using Bayesian methods in combination with stacking. The modeling was carried out on a large-scale expression matrix including more than 6,000 biological samples and 18,000 genes, covering 14 sample classes. The results demonstrated that the spectral consensus version of SOTA consistently provided the best values of internal clustering quality indices and the highest classification accuracy. In particular, in configurations with three to five clusters, 100% accuracy and F1-score were achieved, confirming the diagnostic significance of the identified gene groups. The developed ensemble pipeline represents a scalable and standardized tool for transcriptomic data analysis that can be integrated into decision-support systems for early diagnosis of complex systems.

Key words: *gene expression data, similarity metrics, hybrid model, clustering, classification, personalized medicine, hybridization.*

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

Постановка проблеми

Онкологічні захворювання залишаються однією з провідних причин захворюваності та смертності у світі; рання й точна діагностика критично впливають на результати лікування. Профілювання експресії генів утвердилося як один із ключових інструментів для відкриття біомаркерів і молекулярних підписів, що розрізняють підтипи раку, прогнозують прогресування та підтримують вибір терапії [1]. Водночас характерна багатовимірність, неоднорідність і зашумленість транскриптомних даних ускладнюють застосування класичних підходів машинного навчання. Традиційні алгоритми кластеризації і класифікації втрачають стійкість та інтерпретованість у таких умовах, що знижує їхню практичну цінність для медичних застосувань. Це створює наукову проблему, яка полягає у відсутності універсальних методів аналізу, здатних одночасно:

– забезпечити структуризацію великомасштабних транскриптомних даних із високим рівнем шуму;

– формувати стабільні та інформативні кластери, релевантні біологічному контексту;

– інтегрувати результати неконтрольованого групування з контрольованою класифікацією для підвищення діагностичної точності.

Подолати ці виклики доцільно за допомогою ансамблевих та гібридних методів (bagging, boosting, stacking), які поєднують сильні сторони кількох класифікаторів і кластеризаторів, зменшують дисперсію результатів і підвищують їх узагальнюваність [2].

Аналіз останніх досліджень і публікацій

Гібридні конвеєри, які поєднують неконтрольовану кластеризацію з подальшою контрольованою класифікацією, продемонстрували здатність виокремлювати біологічно змістовні ознаки та досягати високої точності на складних наборах даних [3]. Зазначені методи інтегрують доменні знання (онтології, бази шляхів), що забезпечує кращу інтерпретованість і клінічну релевантність одержаних результатів [4]. Ансамблеві підходи до кластеризації та класифікації дають змогу виявляти патерни експресії за умов гетерогенності пухлин, забезпечуючи одночасно високу точність і здатність до узагальнення, що є необхідним для діагностичних застосувань [5].

Огляд сучасних рішень для класифікації раку за даними експресії охоплює як класичні моделі (SVM, Random Forest), так і глибинні підходи – згорткові та графові нейромережі, а також трансформери, які моделюють ієрархічні/довгодальні залежності та мережевий контекст омікс-даних [6]. Водночас глибинні моделі потребують великих обсягів розмічених даних, відзначаються обмеженою прозорістю процесу прийняття рішень та високою чутливістю до налаштування численних гіперпараметрів, що особливо проявляється у разі зашумлених і дисбалансних вибірок.

Для підвищення стійкості в аналізі омікс-даних запропоновано низку ансамблевих рішень: інтелектуальний конвеєр на базі MLP для діагностики раку молочної залози; консенсусну кластеризацію для scRNA-seq із метою стабілізації ознак; рангові ансамблі дерев, що покращують мультикласову точність у bulk-RNA-seq/мікромасивах. Додаткові дослідження підтверджують переваги мультиомної консенсусної кластеризації у задачах стратифікації пацієнтів із раком шлунка [7], а також ефективність ансамблевих мереж співекспресії для ідентифікації релевантних біомаркерів у пухлинах молочної залози та простати [8].

Більшість досліджень зосереджується або на ансамблевій кластеризації (ensemble clustering), або на ансамблевій класифікації (ensemble classification), рідко поєднуючи їх у єдиний узгоджений конвеєр. Зокрема, консенсусна кластеризація для стратифікації пацієнтів із недрібноклітинним раком легень (НДРЛ) підвищує точність виділення молекулярних підгруп, проте зазвичай не включає контрольованого компонента для діагностичного прогнозу.

Аналіз функціонального збагачення біологічних шляхів часто виконується як окрема *post hoc* процедура, що базується на зовнішніх базах знань (Kyoto Encyclopedia of Genes and

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

Genomes (KEGG), Gene Ontology (GO)). Відсутність інтеграції цього етапу у класифікаційний конвеєр зменшує його вплив на процес відбору ознак і обмежує інтерпретаційні можливості моделей [9].

Мета дослідження

Метою дослідження є розроблення гібридного конвеєра (modular pipeline), що поєднує ансамблеву кластеризацію (SOTA з агломеративним і спектральним консенсусом) та ансамблеву класифікацію, забезпечуючи стійкі, інтерпретовані й узагальнювані діагностичні моделі для різних типів раку на гетерогенних транскриптомних даних.

Виклад основного матеріалу дослідження

Методологія класифікаційного конвеєра (classification pipeline) передбачає поетапну інтеграцію результатів кластеризації з ансамблевими методами навчання. Це підвищує точність класифікації та зберігає інтерпретованість структур даних експресії генів. Загальну схему подано на рис. 1.

Запропонований конвеєр поєднує ансамбль кластеризацій та класифікацій для підвищення діагностичної точності та біологічної інтерпретованості даних експресії генів.

Метод включає такі етапи (рис. 1):

1. Бутстреп-ресемплінг (bootstrap resampling) і базова кластеризація з формуванням N розбиттів (partitions).
2. Консенсусна кластеризація з побудовою фінального розбиття та вибором оптимального числа кластерів K.
3. Формування кластерів, що відображають однорідні групи профілів експресії.
4. Підготовка даних кластерів і поділ на тренувальну (70%) та тестову (30%) вибірки.
5. Класифікація: для кожного кластера будується окремий Random Forest, результати яких агрегуються логістичною регресією у стекінг-моделі з оптимізацією гіперпараметрів.
6. Підсумкова оцінка точності, повноти та F1-міри для вибору оптимальної конфігурації.

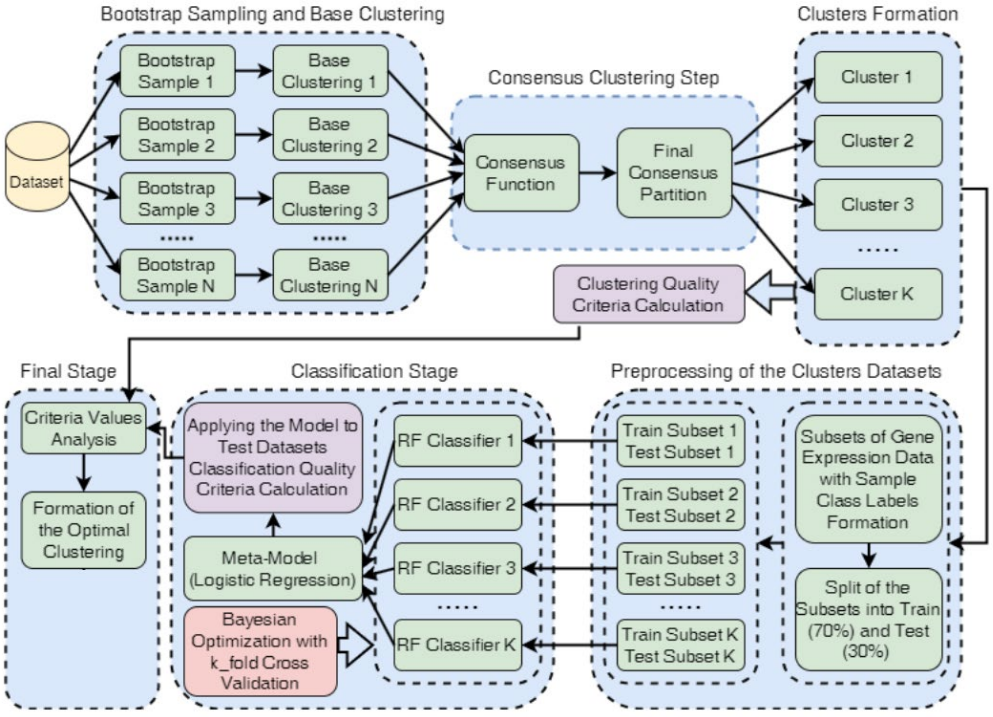

Рис. 1. Блок-схема класифікаційного конвеєра на основі кластеризації профілів експресії генів

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

Ансамблеві методи кластеризації з використанням баггінгу (bagging) [10] поєднують результати кількох розбиттів одного й того ж набору даних із метою отримання більш узагальненої та стабільної структури. Принцип подібний до ансамблевої класифікації: об'єднання «слабких» моделей підвищує стійкість та якість підсумкового результату.

Загальний процес включає два етапи:

1.Генерація ансамблю алгоритмів кластеризацій. Для цього використовують різні підмножини об'єктів або ознак, різні ініціалізації, параметри алгоритмів чи навіть різні методи кластеризації.

2. Агрегація результатів. Будується матриця співналежності, яка характеризує частоту спільної належності пари об'єктів до одного кластера. На її основі формується фінальне консенсусне розбиття.

Одним із найпоширеніших підходів є консенсусна кластеризація, яку широко застосовують у біоінформатиці [11]. Вона дає змогу знизити нестабільність окремих алгоритмів і виявити більш узагальнену структуру даних. Методи на основі баггінгу реалізують принцип бутстепової агрегації (bootstrap aggregating): із початкового набору даних створюються випадкові підвибірки, на яких незалежно виконується кластеризація. Далі отримані розбиття об'єднуються в єдину структуру, що зменшує чутливість до випадкової ініціалізації та враховує варіативність у вхідних даних. Це особливо важливо для задач аналізу експресії генів, де дані мають високу розмірність і значний рівень шуму. На основі матриці співналежності застосовують різні стратегії:

– Агломеративна кластеризація – інтерпретація матриці як матриці подібності й побудова дендрограми.

– Спектральний підхід – перетворення матриці на зважений граф і кластеризація на основі його власних векторів.

– Методи голосування (Voting) – визначення належності об'єктів до кластерів на основі більшості або повторної кластеризації матриці подібності.

Таким чином, застосування баггінг-орієнтованої кластеризації сприяє зниженню варіативності результатів, підвищує здатність моделей до узагальнення та забезпечує формування більш стійких кластерів у високовимірних біоінформаційних даних.

Самоорганізуючий алгоритм кластеризації SOTA (Self-Organizing Tree Algorithm) поєднує принципи самоорганізованих карт Кохонена (SOM) та зростаючих клітинних структур, формуючи бінарну ієрархію, яка природно відображає таксономічні або еволюційні відносини [6]. На відміну від класичних методів кластеризації (k-means, ієрархічної кластеризації чи DBSCAN) SOTA має низку важливих переваг:

– Динамічна адаптація структури до локальної щільності даних.

– Автоматичне визначення оптимальної кількості кластерів.

– Ефективне моделювання складних кореляцій у високовимірних профілях експресії генів.

Завдяки ітеративному процесу навчання та поділу вузлів алгоритм здатний будувати інтерпретовані ієрархічні структури без попереднього задання числа кластерів. Це робить SOTA особливо корисним у задачах аналізу транскриптомних даних, де спостерігаються значний шум і висока розмірність. Для підвищення надійності також застосовувався алгоритм k-medoids, що формує кластери на основі реальних об'єктів (медоїдів), на відміну від k-means, який оперує середніми значеннями. Такий підхід забезпечує більшу стійкість до шуму та кращу інтерпретованість результатів у біологічному контексті [12; 13].

У цьому дослідженні k-medoids реалізовано з використанням кореляційної відстані як метрики подібності. Хоча метод поступається SOTA у здатності відображати складні топологічні структури, його простота та сумісність із матрицями відстаней роблять його практичним інструментом для початкового аналізу високовимірних біологічних даних.

Таким чином, комбінація SOTA та k-medoids у межах ансамблевого підходу дає змогу поєднати інтерпретованість, стійкість і здатність до роботи з гетерогенними наборами даних, що критично важливо для подальших етапів класифікації та біологічної інтерпретації.

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

Для перевірки ефективності кластеризації використано два взаємодоповнюючі складника: внутрішньокластерну компактність та міжкластерну віддаленість.

Внутрішньокластерна компактність (QC_{in}) визначається як середнє значення відстаней між кожним профілем експресії та медоїдом відповідного кластера:

$$QC_{in} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{N_i}\sum_{j=i}^{N_i} dist\left(x_{ij}, edoid_i\right)\right), \tag{1}$$

де k – кількість кластерів, N_i – кількість профілів у кластері i, $dist(\cdot,\cdot)$ – кореляційна відстань.

Міжкластерна відстань (QC_{out}) визначається як середня відстань між медоїдами всіх пар кластерів:

$$QC_{out} = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} dist\left(medoid_i, medoid_j\right) \tag{2}$$

Для узагальненої оцінки використано інтегральний критерій якості:

$$QC = \frac{QC_{in}}{QC_{out}} \tag{3}$$

Нижчі значення QC свідчать про кращу щільність і відділення кластерів.

Додатково застосовано внутрішні валідаційні метрики з пакету clusterCrit (R), які широко використовуються для оцінювання кластеризації у високовимірних просторах.

Таким чином, у роботі застосовано поєднання власного критерію QC та класичних індексів (CH, SH, PBM), що дало змогу комплексно оцінити якість кластеризації в умовах високої розмірності та шумності даних експресії генів (табл. 1).

Таблиця 1

Внутрішні індекси якості кластеризації

Індекс	**Опис**	**Оптимум**
Calinski–Harabasz	Співвідношення міжкластерної та внутрішньокластерної дисперсії	Максимізація
Silhouette	Середня «силуетна» ширина; показує, наскільки добре об’єкти відповідають своєму кластеру	Максимізація
PBM	Композитний критерій, що враховує відстань між кластерами, компактність та їх кількість	Максимізація

Для оцінки узгодженості та діагностичної цінності результатів кластеризації було реалізовано класифікаційний конвеєр на основі ансамблевого навчання зі стекінгом. Концепція методу полягає у побудові незалежних класифікаторів для кожного кластера та подальшій інтеграції їхніх прогнозів за допомогою метамоделі. Така архітектура підвищує робастність, знижує ризик перенавчання та забезпечує регуляризацію на рівні метаагрегації. У ролі базових класифікаторів застосовано алгоритм Random Forest (RF), оскільки він демонструє високу ефективність на даних великої розмірності, стійкий до шуму та має вбудовані механізми відбору ознак. Агрегація результатів окремих моделей здійснюється за допомогою логістичної регресії (LR) як метакласифікатора. Простота й інтерпретованість LR роблять її оптимальним вибором для інтеграції й водночас мінімізують ризик перенавчання на стекінговому рівні.

Для оптимізації гіперпараметрів Random Forest використано байєсівську оптимізацію з 5-кратною крос-валідацією, що забезпечило ефективний пошук у просторі параметрів за обмежених обчислювальних ресурсів. Щоб уникнути інформаційного витоку та переоцінки результатів, застосовано дворівневу схему валідації: спершу отримувалися out-of-fold прогнози від базових моделей, які формували ознакову матрицю для навчання метамоделі, після

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

чого фінальна оцінка здійснювалася на незалежній тестовій вибірці. У результаті алгоритм валідації кластеризації на основі SOTA за допомогою ансамблевої класифікації та стекінгу виглядає так:

1. Завантажити кластери експресії генів $X_1, X_2, \ldots, X_6$ та вектор цільових міток y .
2. Розділити вибірку на тренувальну та тестову підмножини.
3. Для кожного кластера $i = 1..6$:

a.Виконати байєсівську оптимізацію гіперпараметрів Random Forest X_i^{train} із 5-кратною крос-валідацією.

b. Отримати мітки класів y_i^{train} для X_i^{train} .

c.Навчити базовий класифікатор RF_i на тренувальній підмножині X_i^{train} .

d. Отримати прогнози для тестової підмножини та обчислити метрики (Accuracy, F1, Weighted F1).

4. Побудувати stacking-матрицю $X_{train}^{stack} = \left[y_1^{train}, \ldots, y_6^{train} \right]$ з out-of-fold прогнозів базових класифікаторів.
5. Закодувати мітки класів у числовий формат.
6. Навчити логістичну регресію як метакласифікатор на stacking-матриці X_{train}^{stack} .
7. Сформувати тестову stacking-матрицю та згенерувати фінальні прогнози y_i^{stack} .
8. Оцінити точність, повноту, F1-міру та побудувати матрицю неточностей у вигляді теплокарти.

Отримані результати засвідчили, що застосування алгоритму k-medoids призводить до зниження якості кластеризації як на рівні окремих кластерів, так і на рівні метамоделі. Формування консенсусних матриць на основі k-medoids виявилося обчислювально затратним і слабко масштабованим для великомасштабних транскриптомних даних. Це пояснюється обмеженою здатністю методу відображати складні структури у високовимірних просторах та жорстким механізмом призначення зразків.

Натомість алгоритм SOTA продемонстрував значно вищу ефективність завдяки врахуванню топологічних та щільнісних характеристик простору ознак. Саме тому в подальшому аналізі він був вибраний як основний метод.

Для оцінювання класифікаційної здатності підходу було використано матрицю експресії, яка охоплювала 6 310 зразків і понад 18 тис генів у 14 діагностичних класах (13 типів пухлин і контрольна група)[14]. На рис. 2 подано загальну структуру класифікаційної задачі, що ілюструє складність розподілу вибірки між різними онкологічними категоріями.

Було встановлено, що використання алгоритму k-medoids супроводжується погіршенням якості кластеризації. Алгоритм виявився обчислювально затратним і обмеженим у здатності відображати складні топології високорозмірних даних. Як наслідок, формуються слабо відокремлені кластери з низькою інтерпретованістю.

Натомість алгоритм SOTA продемонстрував значно кращі результати. Його здатність ураховувати топологічні та щільнісні властивості простору ознак забезпечила формування стабільніших структур. На рис. 3 показано розподіл генів у різних конфігураціях SOTA та k-medoids, що ілюструє перевагу першого підходу.

Для аналізу було використано комбінований критерій QC , а також відомі індекси якості – Calinski–Harabasz (CH), Silhouette (SH) та PBM. Аналогічний підхід до оцінювання ефективності кластеризації було реалізовано у роботі [15].

Результати показали, що:

– класичний SOTA (без консенсусу) вже формує збалансовані двокластерні структури;

– поєднання агломеративної та спектральної консенсусної кластеризації додатково покращує метрики якості;

– алгоритм k-medoids продемонстрував найгірші результати з низькими значеннями CH, SH та PBM.

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

Таким чином, найбільш стабільні конфігурації – це SOTA з агломеративним або спектральним консенсусом за $K = 2$.

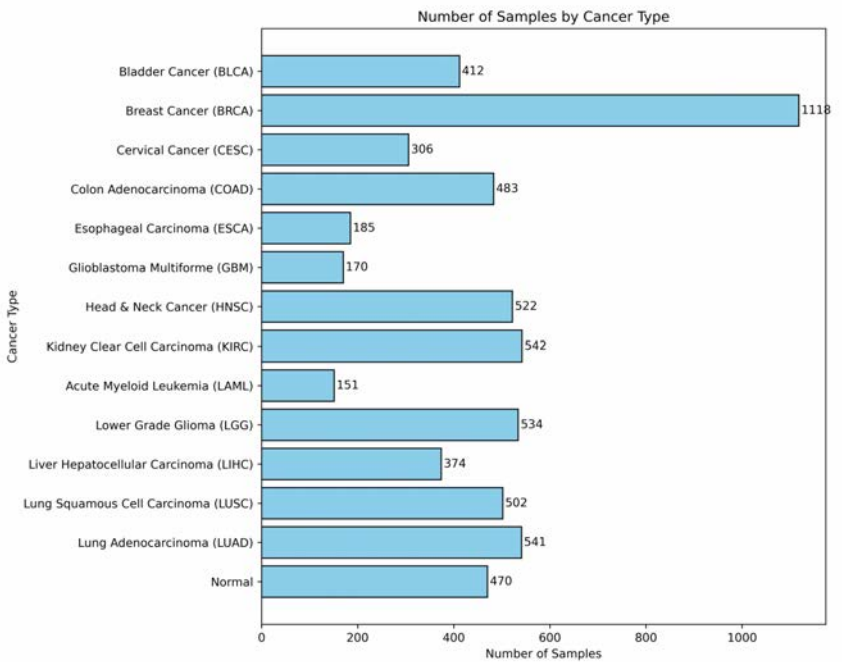

Рис. 2. Класифікація біологічних зразків за профілями експресії генів у межах 14 діагностичних категорій (13 типів пухлин та контрольна група)

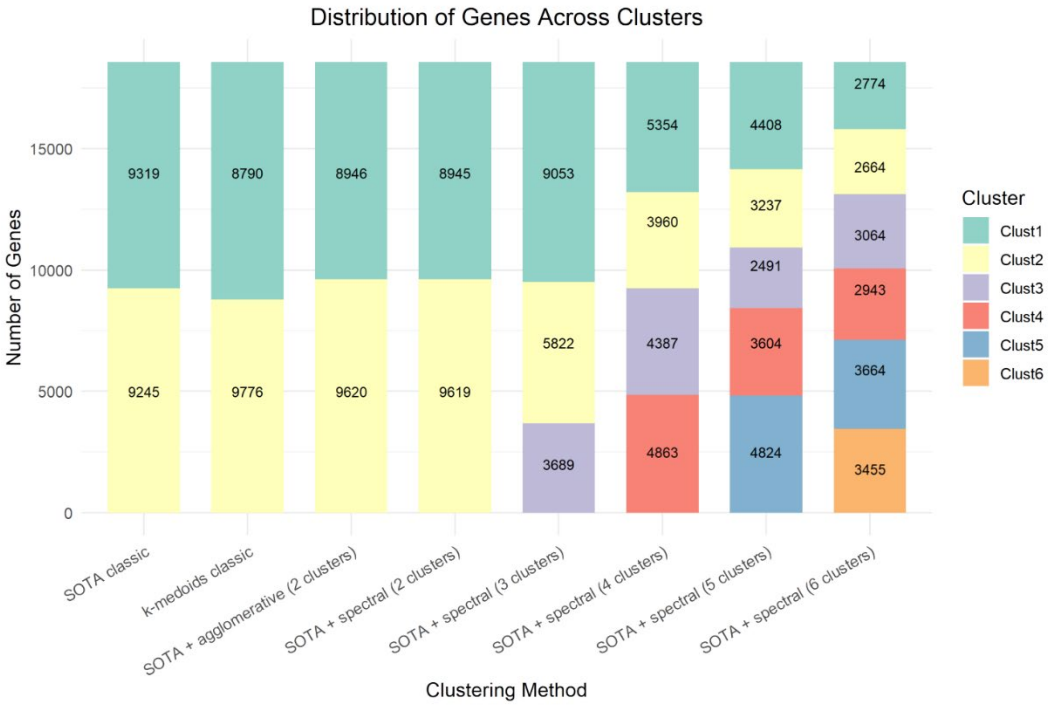

Рис. 3. Розподіл кількості генів між кластерами для різних конфігурацій алгоритмів (SOTA та k-medoids, у т. ч. з агломеративним і спектральним консенсусом)

На рис. 4 відображено залежність значень показників QC, CH, SH та PBM від використаного алгоритму.

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

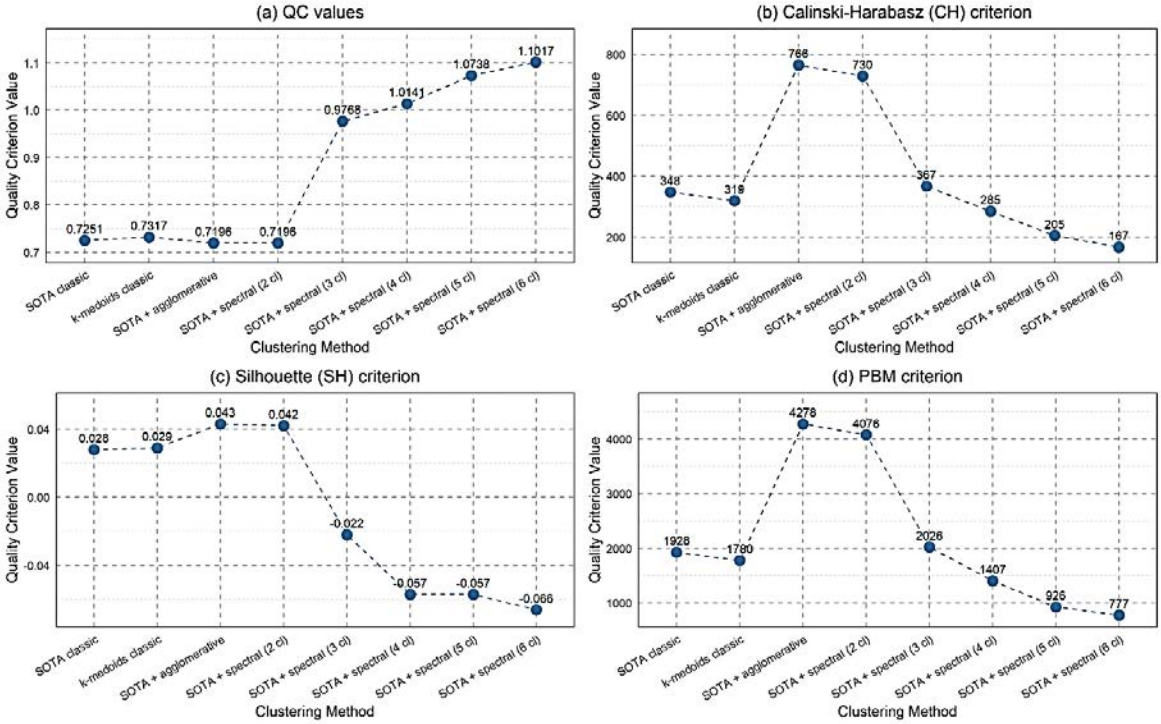

Рис. 4. Залежність метрик QC, CH, SH та PBM від алгоритму та конфігурації SOTA

Результати моделювання підтверджують ключову роль вибору алгоритму кластеризації у формуванні якісних та інтерпретованих структур профілів експресії генів. Алгоритм k-medoids, незважаючи на простоту та зручність реалізації, продемонстрував обмежену ефективність у випадку високовимірних транскриптомних даних. Його використання супроводжувалося формуванням слабо відокремлених кластерів та низькими значеннями внутрішніх індексів якості (CH, SH, PBM), що узгоджується з відомими обмеженнями цього методу для складних біологічних просторів ознак. Окрім того, обчислювальна складність k-medoids суттєво зростала зі збільшенням розмірності даних, що обмежує його масштабованість у практичних застосуваннях.

На відміну від цього алгоритм SOTA продемонстрував значно кращі результати, зокрема завдяки здатності враховувати топологічні та щільнісні характеристики простору даних. У поєднанні з консенсусними стратегіями (агломеративною та спектральною) він формував більш компактні та відокремлені кластери, що підтверджувалося підвищеними значеннями метрик CH, SH та PBM. Особливої уваги заслуговує спектральний консенсус, який показав найвищу стабільність результатів у різних конфігураціях та забезпечив максимальні значення інтегрального критерію QC. Це свідчить про здатність методу ефективно зменшувати вплив шуму та локальних аномалій у даних.

Додатково слід відзначити, що використання кластерів як нових ознак для класифікаційної моделі на основі Random Forest забезпечило суттєве підвищення точності діагностики. Стекінгова архітектура, яка поєднала результати кількох базових моделей через логістичну регресію, гарантувала високу узгодженість прогнозів та знизила ризик перенавчання. У конфігураціях із трьома-п'ятьма кластерами досягнуто найкращих показників (Accuracy та F1 = 1.0), що підтверджує високу діагностичну значущість виділених груп генів та потенціал методу для практичного використання в онкології.

Таким чином, проведене порівняння чітко показало перевагу SOTA з консенсусними стратегіями над класичними методами, що робить його перспективним інструментом для побудови

https://doi.org/10.32782/mathematical-modelling/2025-8-2-33 ISSN 2618-0332 (print), 2618-0340 (online)

масштабованих і стійких діагностичних моделей. Отримані результати створюють підґрунтя для інтеграції біологічної інтерпретації кластерів через збагачення онтологій (GO, KEGG) та подальшої побудови функціональних мереж, що стане наступним етапом дослідження.

Висновки

У статті представлено комплексний ансамблевий конвеєр кластеризації та класифікації даних експресії генів, спрямований на вдосконалення процесів діагностики стану складних систем. Запропонована методологія ґрунтується на інтеграції алгоритму самоорганізованих дерев (Self-Organizing Tree Algorithm, SOTA) із консенсусними підходами агломеративної та спектральної кластеризації, що дало змогу подолати обмеження класичних методів, особливо у випадках високорозмірних і зашумлених даних.

Проведене моделювання засвідчило, що SOTA у поєднанні зі спектральним консенсусом формує найбільш стабільні та інформативні кластери, що підтверджується внутрішніми критеріями якості (QC, Calinski–Harabasz, Silhouette, PBM) та зовнішньою оцінкою через класифікацію. Отримані результати показали перевагу цього підходу над альтернативними методами та забезпечили досягнення ідеальних значень Accuracy і F1-міри (= 1.0) у конфігураціях із трьома-п'ятьма кластерами. Це свідчить про високу діагностичну значущість сформованих кластерів та їх придатність для практичного використання.

Детальний аналіз результатів підтвердив, що k-medoids має обмежену ефективність для транскриптомних даних, тоді як гібридні стратегії на основі SOTA демонструють кращу масштабованість, стійкість до шуму та інтерпретованість. Важливим є також поєднання кластеризації з ансамблевою класифікацією на базі Random Forest та стекінгової архітектури, що забезпечило узгодженість прогнозів і зниження ризику перенавчання.

У цій роботі основний акцент зроблено на методології кластеризації та класифікації. Подальші дослідження будуть зосереджені на повній біологічній інтерпретації отриманих результатів, зокрема на аналізі збагаченості шляхів та побудові мережевих моделей із використанням баз KEGG та Gene Ontology. Це дасть змогу визначити функціональні механізми, пов'язані з виявленими кластерами, та підсилити їх релевантність.

Таким чином, запропонований підхід є масштабованим і стандартизованим інструментом для аналізу транскриптомних даних, що підвищує точність діагностики, покращує інтерпретованість результатів і створює підґрунтя для розроблення моделей діагностики складних систем.

Список використаної літератури

1. Ryan C., O'Driscoll A., Coughlan J., Luo J. Cancer diagnosis and prognosis through gene expression. *Briefings in Bioinformatics.* 2023. Vol. 24. Art. no. bbac527. DOI: https://doi.org/10.1093/bib/bbac527
2. Golalipour K., Akbari E., Hamidi S., Lee M., Enayatifar R. From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence.* 2021. Vol. 104. Art. no. 104388. DOI: https://doi.org/10.1016/j.engappai.2021.104388
3. Babichev S., Yasinska-Damri L., Liakh I. A hybrid model of cancer diseases diagnosis based on gene expression data with joint use of data mining methods and machine learning techniques. *Applied Sciences.* 2023. Vol. 13. Art. no. 6022. DOI: https://doi.org/10.3390/app13106022
4. Galluzzo Y. A comprehensive review of the data and knowledge graph approaches in bioinformatics. *Computer Science and Information Systems.* 2024. Vol. 21. P. 1055–1075. DOI: https://doi.org/10.2298/CSIS230530027G
5. Shen J., Guo X., Bai H., Luo J. CAEM-GBDT: a cancer subtype identifying method using multi-omics data and convolutional autoencoder network. *Frontiers in Bioinformatics.* 2024. Vol. 15. Art. no. 1403826. DOI: https://doi.org/10.3389/fbinf.2024.1403826

6. Khalsan M., Machado L., Al-Shamery E., Liu R. A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access.* 2022. Vol. 10. P. 27522–27534. DOI: https://doi.org/10.1109/ACCESS.2022.3146312
7. Xianyu H., Zhenglin W., Qing W. Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: A multi-omics consensus ensemble clustering. *Biomedicine & Pharmacotherapy.* 2021. Vol. 144. Art. no. 112222. DOI: https://doi.org/10.1016/j.biopha.2021.112222
8. Figueroa-Martínez J., Saz-Navarro D. M., López-Fernández A., Rivera J. Computational ensemble gene co-expression networks for breast and prostate cancer biomarker identification. *Informatics*. 2024. Vol. 11. Art. no. 14. DOI: https://doi.org/10.3390/informatics11020014
9. Mubeen S., Hoyt C., Gemünd A., & Smith K. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Frontiers in Genetics.* 2019. № 22. Art. no. 1203. DOI: https://doi.org/10.3389/fgene.2019.01203
10. Jianxia L., Liu R., Mingyang Z., Yangyang L. Ensemble-based multi-objective clustering algorithms for gene expression data sets. *IEEE Congress on Evolutionary Computation (CEC)* : Donostia–San Sebastián, Spain, 5–8 June. 2017. P. 333–340.
11. Panwong P., Boongoen T., Iam-On N., Mullaney J. Exploiting consensus clustering for light curve data analysis. *IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)* : Yunlin, Taiwan, October 3–6. 2019. P. 498–501.
12. Dopazo J., Carazo J. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution.* 1997. Vol. 44. P. 226–233. DOI: https://doi.org/10.1007/PL00006139
13. Heidari J., Daneshpour N., Zangeneh A. A novel k-means and k-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers. *Pattern Recognition.* 2024. Vol. 155. Art. no. 110639. DOI: https://doi.org/10.1016/j.patcog.2024.110639
14. Babichev S., Yarema O., Savchenko A. Evaluating proximity metrics for gene expression data: A hybrid model integrating data mining and machine learning techniques for disease diagnosis systems. *Biomedical Signal Processing and Control.* 2025. Vol. 110. Art. no. 108115. DOI: https://doi.org/10.1016/j.bspc.2025.108115
15. Babichev S., Yarema O., Liakh I., Shumylo N. A gene ontology-based pipeline for selecting significant gene subsets in biomedical applications. *Applied Sciences.* 2025. Vol. 15. Art. no. 4471. DOI: https://doi.org/10.3390/app15084471

References

1. Ryan, C., O'Driscoll, A., Coughlan, J., & Luo, J. (2023). Cancer diagnosis and prognosis through gene expression. *Briefings in Bioinformatics*, 24, 112–123. https://doi.org/10.1093/bib/bbac527 [in English].
2. Golalipour, K., Akbari, E., Hamidi, S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, 104, 104388. https://doi.org/10.1016/j.engappai.2021.104388 [in English].
3. Babichev, S., Yasinska-Damri, L., & Liakh, I. (2023). A hybrid model of cancer diseases diagnosis based on gene expression data with joint use of data mining methods and machine learning techniques. *Applied Sciences*, 13, 6022. https://doi.org/10.3390/app13106022 [in English].
4. Galluzzo, Y. (2024). A comprehensive review of the data and knowledge graph approaches in bioinformatics. *Computer Science and Information Systems*, 21, 1055–1075. https://doi.org/10.2298/CSIS230530027G [in English].
5. Shen, J., Guo, X., Bai, H., & Luo, J. (2024). CAEM-GBDT: a cancer subtype identifying method using multi-omics data and convolutional autoencoder network. *Frontiers in Bioinformatics*, 15, 1403826. https://doi.org/10.3389/fbinf.2024.1403826 [in English].

6. Khalsan, M., Machado, L., Al-Shamery, E., & Liu, R. (2022). A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access*, 10, 27522–27534. https://doi.org/10.1109/ACCESS.2022.3146312 [in English].
7. Xianyu, H., Zhenglin, W., & Qing, W. (2021). Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: A multi-omics consensus ensemble clustering. *Biomedicine & Pharmacotherapy*, 144, 112222. https://doi.org/10.1016/j.biopha.2021.112222 [in English].
8. Figueroa-Martínez, J., Saz-Navarro, D.M., López-Fernández, A., & Rivera, J. (2024). Computational ensemble gene co-expression networks for breast and prostate cancer biomarker identification. *Informatics*, 11, 14. https://doi.org/10.3390/informatics11020014 [in English]
9. Mubeen, S., Hoyt, C., Gemünd, A., & Smith, K. (2019). The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Frontiers in Genetics*, 22, 1203. https://doi.org/10.3389/fgene.2019.01203 [in English].
10. Jianxia, L., Liu, R., Mingyang, Z., & Yangyang, L. (2017). Ensemble-based multi-objective clustering algorithms for gene expression data sets. *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2017)*, Donostia–San Sebastián, Spain [in English].
11. Panwong, P., Boongoen, T., Iam-On, N., & Mullaney, J. (2019). Exploiting consensus clustering for light curve data analysis. *IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, Yunlin, Taiwan [in English].
12. Dopazo, J., & Carazo, J. (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution*, 44, 226–233. https://doi.org/10.1007/PL00006139 [in English].
13. Heidari, J., Daneshpour, N., & Zangeneh, A. (2024). A novel k-means and k-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers. *Pattern Recognition*, 155, 110639. https://doi.org/10.1016/j.patcog.2024.110639 [in English].
14. Babichev, S., Yarema, O., & Savchenko, A. (2025). Evaluating proximity metrics for gene expression data: A hybrid model integrating data mining and machine learning techniques for disease diagnosis systems. *Biomedical Signal Processing and Control*, 110, 108115. https://doi.org/10.1016/j.bspc.2025.108115 [in English].
15. Babichev, S., Yarema, O., Liakh, I., & Shumylo, N. (2025). A gene ontology-based pipeline for selecting significant gene subsets in biomedical applications. *Applied Sciences*, 15, 4471. https://doi.org/10.3390/app15084471 [in English].

Ярема Олег Романович – к.т.н., доцент, доцент кафедри цифрової економіки та бізнес-аналітики Львівського національного університету імені Івана Франка. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

Бабічев Сергій Анатолійович – д.т.н., професор, професор кафедри інформатики університету Яна Євангеліста Пуркінє в Усті на Лабі, Чехія; професор кафедри фізики Херсонського державного університету. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

Yarema Oleh Romanovych – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Digital Economy and Business Analytics of the Ivan Franko National University of Lviv. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

Babichev Sergii Anatoliiovych – Doctor of Technical Sciences, Professor, Professor at the Department of Informatics of the Jan Evangelista Purkyně University in Ústí nad Labem, Czech Republic; Professor at the Department of Physics at Kherson State University. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

Дата надходження статті: 01.09.2025
Дата прийняття статті: 28.11.2025
Опубліковано: 30.12.2025