

МАШИННЕ НАВЧАННЯ В СПЕКТРОСКОПІЧНОМУ АНАЛІЗІ СЛИНИ: ОЦІНКА ЕФЕКТИВНОСТІ КЛАСТЕРИЗАЦІЇ ТА ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ ВИЗНАЧЕННЯ РАКУ РОТОВОЇ ПОРОЖНИНИ

С. Л. Говорнян, О. І. Іващук

Буковинський державний медичний університет, м. Чернівці, Україна

Ключові слова:

рак ротової порожнини, слина, Раман-спектроскопія, машинне навчання, згорткові нейронні мережі, кластеризація, скринінг, біомаркери.

Клінічна та експериментальна патологія 2025. Т.24, №1 (91). С. 58-66.

DOI 10.24061/1727-4338.XXIV.1.91.2025.09

E-mail:
hovornyan@bsmu.edu.ua

Рак ротової порожнини (РРП) є серйозною онкологічною проблемою, оскільки його пізнє виявлення значно знижує виживаність пацієнтів. Раман-спектроскопія є перспективним методом неінвазивного скринінгу, проте ефективність її обробки залежить від вибору аналітичного підходу. Неконтрольовані методи кластеризації часто не виявляють чітких груп у спектрах, що обмежує їхню застосовність. Натомість згорткові нейронні мережі (CNN) можуть виявляти приховані патерни у спектральних даних.

Мета – оцінити ефективність кластеризації та згорткових нейронних мереж у спектроскопічному аналізі слини для скринінгу раку ротової порожнини та визначити оптимальний підхід до класифікації спектрів.

Матеріали та методи. Досліджено спектральні дані слини осіб чотирьох груп: здорових, пацієнтів із передраковими станами (лейкоплакія, ліхен-планус) та хворих на рак. Використано методи кластеризації (PCA, t-SNE, K-Means, DBSCAN, GMM, автоенкодер + K-Means) та CNN.

Результати. Кластеризація не виявила чітких груп, що підтверджено низькими значеннями ARI (0,02-0,05) та коефіцієнта силуету (~0,1). CNN досягла точності 84,6% та AUC 0,86 при використанні всього спектрального діапазону.

Висновки. Неконтрольовані методи є неефективними через високе перекриття спектрів різних груп. Натомість CNN забезпечує високу точність класифікації, що робить її перспективним інструментом для автоматизованого скринінгу РРП.

Key words:

oral cancer, saliva, Raman-spectroscopy, machine learning, convolutional neural networks, clustering, screening, biomarkers.

Clinical and experimental pathology 2025. Vol.24, № 1 (91). P. 58-66.

MACHINE LEARNING IN SPECTROSCOPIC ANALYSIS OF SALIVA: EVALUATION OF THE EFFECTIVENESS OF CLUSTERING AND DEEP NEURAL NETWORKS FOR THE DETECTION OF ORAL CANCER

S. L. Hovornyan, O. I. Ivashchuk

Bukovinian State Medical University, Chernivtsi, Ukraine

Oral cancer (OC) remains a serious oncological problem, as late detection significantly reduces patient survival. Raman spectroscopy is a promising method for non-invasive screening; however, its effectiveness depends on the chosen analytical approach. Unsupervised clustering methods often fail to identify clear groups in spectral data, limiting their applicability. In contrast, convolutional neural networks (CNNs) can detect hidden patterns in spectral data.

Objective – to evaluate the effectiveness of clustering methods and CNNs in the saliva spectroscopic analysis for OC screening and determine the optimal approach for spectral classification.

Materials and Methods. Spectral data from four patient groups were analyzed: healthy individuals, patients with precancerous conditions (leukoplakia, lichen planus), and patients with oral cancer. Various clustering methods (PCA, t-SNE, K-Means, DBSCAN, GMM, autoencoder + K-Means) and CNNs were applied.

Results. Clustering methods failed to form distinct groups, as indicated by low ARI (0.02-0.05) and silhouette coefficient (~0.1). CNNs achieved an accuracy of 84.6% and an AUC of 0.86 when using the full spectral range.

Conclusions. Unsupervised clustering methods proved ineffective due to the high overlap of spectra from different groups. In contrast, CNNs demonstrated high classification accuracy, making them a promising tool for automated OC screening.

Вступ

Рак ротової порожнини (РРП), переважно плоскоклітинний, залишається серйозною проблемою охорони здоров'я. Щорічно у світі реєструють сотні тисяч нових випадків, а 5-річна

виживаність хворих не перевищує ~50% через пізнє виявлення захворювання [1]. Встановлено, що рання діагностика суттєво покращує прогноз: на ранніх стадіях (I стадія) п'ятирічне виживання може сягати 80-90%, тоді як на пізніх (III-IV) стадіях

цей показник різко знижується до ~30-40% [1]. На жаль, до половини випадків РРП діагностуються вже в поширеній стадії, оскільки початковий перебіг часто безсимптомний, а можливості існуючих скринінгових методів обмежені. Стандартний огляд порожнини рота (візуальний огляд і пальпація) має невисоку чутливість та специфічність і не завжди дозволяє відрізнити доброякісні ураження від передракових чи ранніх злоякісних змін. Отже, розробка нових неінвазивних інструментів для ефективного скринінгу РРП є вкрай актуальною задачею. Оптичні методи діагностики привертають дедалі більше уваги як потенційні рішення для раннього виявлення раку. Одним із перспективних підходів є Раман-спектроскопія – метод комбінаційного розсіювання світла, який за допомогою лазерного випромінювання дозволяє *in situ* отримувати спектри, що відображають молекулярний склад біологічних тканин або рідин. Переваги Раман-спектроскопії полягають у її неінвазивності, швидкості та здатності виявляти тонкі біохімічні зміни, пов'язані зі злоякісною трансформацією. Зокрема, у спектрах слини чи тканин можуть фіксуватися зміни вмісту білків, ліпідів, нуклеїнових кислот та інших метаболітів, характерні для розвитку пухлини. Останні дослідження підтверджують потенціал Раман-спектроскопії як додаткового інструменту ранньої діагностики РРП та передракових станів, демонструючи характерні «спектральні відбитки» (патерни піків), що відрізняють злоякісні зразки від нормальних [2]. Втім, інтерпретація складних багатовимірних спектральних даних потребує застосування потужних аналітичних методів.

Методи штучного інтелекту, зокрема машинного навчання, нині все ширше застосовуються для аналізу медичних даних. Глибоке навчання на основі згорткових нейронних мереж (Convolutional Neural Networks, CNN) демонструє високу результативність у діагностиці різних видів раку за зображеннями та спектрами [3]. На відміну від традиційного статистичного аналізу, що зазвичай спирається на вибір окремих спектральних піків або лінійні комбінації ознак, CNN автоматично виявляють комплексні нелінійні закономірності у даних. Це особливо важливо для раманівських спектрів, де ключові діагностичні ознаки можуть бути розподілені по всьому спектральному діапазону і неочевидні при простому візуальному аналізі. Останні п'ять років позначилися стрімким зростанням кількості робіт, що поєднують Раман-спектроскопію з методами глибокого навчання для діагностики раку [3]. У пілотних дослідженнях показано, що такі алгоритми здатні розпізнавати злоякісні зміни з точністю, близькою до гістопатологічного аналізу [5]. Отже використання CNN обіцяє суттєво підвищити чутливість і специфічність неінвазивного скринінгу раку ротової порожнини порівняно з класичними підходами [3, 6].

У цій роботі представлено результати застосування спектроскопії слини у поєднанні з методами машинного навчання для виявлення РРП. Зокрема, здійснено порівняння двох підходів до аналізу

спектральних даних: неконтрольованого кластерного аналізу (без використання інформації про діагноз) та контрольованого навчання із використанням згорткової нейронної мережі.

Мета роботи

Оцінити ефективність кластеризації та згорткових нейронних мереж у спектроскопічному аналізі слини для скринінгу раку ротової порожнини та визначити оптимальний підхід до класифікації спектрів.

Матеріали та методи дослідження

Популяція дослідження. У дослідження включено 200 учасників, розподілених на чотири групи відповідно до клінічного діагнозу:

1. Контрольна група (здорові, N=43) – особи без видимих патологічних змін у ротовій порожнині та без історії передракових чи онкологічних захворювань.

2. Група червоного плоского лишая (N=69) – пацієнти з діагностованим оральним ліхен-планус (хронічне запальне ураження слизової оболонки рота), яке розглядається як потенційно передраковий стан.

3. Група лейкоплакії (N=44) – пацієнти з виявленими білуватими плямами на слизовій оболонці (лейкоплакія), що класифікуються як передракові ураження із ризиком малігнізації.

4. Група раку (N=44) – хворі з підтвердженим плоскоклітинним раком ротової порожнини, встановленим шляхом біопсії та гістопатологічного аналізу.

Віковий та гендерний розподіл між групами був порівняним (відмінності статистично незначущі, $p > 0,05$), що мінімізує вплив демографічних чинників на результати спектрального аналізу. Кожен учасник надав інформовану згоду на участь у дослідженні, протокол якого був затверджений локальним комітетом з етики.

Збір зразків та отримання спектрів. У всіх пацієнтів збирали нестимульовану слину об'ємом ~2 мл. Збір здійснювався вранці натщесерце (перед цим пацієнти не вживали їжу, напої, не палили), після полоскання рота фізіологічним розчином. Для консервації зразки негайно поміщали у морозильну камеру при -20°C . Перед спектроскопічним аналізом слину розморожували до кімнатної температури і центрифугували (протягом 10 хв при 6000 об/хв) для видалення клітинного детриту та домішок. Отриманий прозорий супернатант переносили на інертну підкладку (наприклад, алюмінієву фольгу або кварцову пластинку) та висушували при кімнатній температурі до утворення тонкої плівки. Такий метод підготовки дозволяє сконцентрувати біомолекули слини в малому об'ємі та покращити співвідношення сигнал/шум у Раман-спектрі.

Раманівські спектри реєстрували за допомогою переносного спектрометра збудженням лазером 785 нм. Спектральний діапазон вимірювання становив від ~800 до 1800 cm^{-1} – ці довжини охоплюють основні коливальні смуги біомолекул (наприклад, амід I ~1650 cm^{-1} , валентні коливання CH_2/CH_3 груп ~1440-1450 cm^{-1} , смуги нуклеїнових кислот ~1090 cm^{-1} тощо). Спектральна роздільна

здатність приладу $\sim 4 \text{ см}^{-1}$. Для кожного зразка накопичували кілька спектрів (від 3 до 5) з експозицією 10-20 с та усереднювали їх для підвищення точності сигналу. Отримані спектри піддавали стандартній попередній обробці: вилучення флуоресцентного фону (базова лінія усувалася шляхом підгонки поліному), згладжування шумів (фільтр Савіцького-Голея) та нормалізація інтенсивності (наприклад, до висоти найбільшого піка). Така обробка зменшує вплив експериментальних артефактів і полегшує подальший аналіз.

Методи аналізу даних. Для аналізу спектральних даних застосовано два підходи: неконтрольовану кластеризацію (без використання міток класів) та контрольовану класифікацію за допомогою нейронної мережі. Спочатку досліджували можливість автоматичного поділу спектрів на групи за допомогою алгоритмів кластеризації. Для цього використовували такі методи зниження розмірності та кластерного аналізу:

- **Головні компоненти (PCA).** Метод основних компонент використовували для зменшення розмірності спектральних даних і виявлення напрямків найбільшої варіабельності. PCA дозволяє представити кожен спектр кількома головними компонентами, що пояснюють більшу частку дисперсії даних, спрощуючи подальший аналіз.

- **t-SNE.** Т-розподілене стохастичне сусіднє вкладання – нелінійний метод візуалізації високовимірних даних. Застосовувався для проєкції спектрів у 2-вимірний простір із метою перевірити, чи групуються вони природно за клінічними класами (здорові, передрак, рак). t-SNE зберігає локальну схожість між зразками, тому точки, що утворюють щільні кластери на площині t-SNE, можуть свідчити про внутрішню структуру даних.

- **К-середніх (K-Means).** Класичний алгоритм *partitioning*-кластеризації, який розбиває дані на наперед задану кількість кластерів (k) шляхом мінімізації внутрішньогрупового розсіювання. Ми випробували метод К-середніх, задавши $k=4$ (за кількістю відомих груп), щоб перевірити, чи співпадуть знайдені алгоритмом кластери зі справжніми діагностичними категоріями.

- **DBSCAN.** Алгоритм кластеризації на основі щільності (Density-Based Spatial Clustering of Applications with Noise) шукає ділянки високої щільності точок, відокремлені ділянками низької щільності. DBSCAN має перевагу в тому, що не потребує задавати кількість кластерів заздалегідь і може виявляти кластери довільної форми, а також виділяти шуми (точки, що не входять до жодного кластеру). Ми застосували DBSCAN для перевірки, чи існують природні щільні групи спектрів у вибірці.

- **Модель гаусових сумішей (GMM).** Цей підхід розглядає дані як суміш декількох багатовимірних нормальних розподілів. Алгоритм очікування-максимізації (EM) використовували для оцінки параметрів моделей і кластеризації спектрів. GMM, подібно до К-середніх, може розділити дані на задане число кластерів, але є гнучкішим, оскільки враховує форму кластерів (коваріацію).

- **Ієрархічна кластеризація.** Використано агломеративний підхід, при якому спочатку кожен спектр вважається окремим кластером, а далі кластери послідовно об'єднуються на основі відстані (евклідової між спектрами або відстані Махаланобіса) до утворення ієрархічної структури (дендрограми). Ієрархічний метод дозволяє дослідити різні рівні кластеризації; ми перевіряли, чи виділяються на якомусь рівні дендрограми чотири чіткі групи, що відповідають контрольній, передраковим та раковій групам.

- **Автоенкодер + K-Means.** Додатково досліджено підхід глибокого неконтрольованого навчання: автоенкодер на основі нейронної мережі використовували для отримання компактного представлення спектрів (властивостей зниженої розмірності), після чого проводили кластеризацію отриманих кодів методом К-середніх. Ідея полягає в тому, що автоенкодер може навчитися виділяти приховані особливості спектрів, які спростять поділ на кластери. Архітектура автоенкодера складалася з кількох шарів кодування, що поступово зменшують кількість ознак (вузьке «горло» мережі), і симетричних шарів декодування. Після навчання автоенкодера (за принципом відновлення вихідного спектра зі стислого коду) брали вектор коду низької розмірності як ознаки для кластеризації K-Means ($k=4$). Сподівалися, що такий двоетапний підхід дозволить виявити більш чіткі групи спектрів.

Після застосування наведених методів здійснювали оцінку якості кластеризації. Оскільки справжні належності спектрів до клінічних груп відомі, для кількісної оцінки використано коефіцієнт силуету та скоригований індекс Ренда (Adjusted Rand Index, ARI), який характеризує відокремленість кластерів (варіює від -1 до $+1$, де значення, близькі до 1 , вказують на добре відокремлені, компактні кластери, а значення ~ 0 означають значне перекриття кластерів). ARI вимірює відповідність отриманої кластеризації істинному розподілу класів, враховуючи випадкові збіги; $ARI = 1$ відповідає ідеальному збігу з еталонними групами, а 0 – рівню випадкового розподілу.

Паралельно з неконтрольованим аналізом, був реалізований контрольований підхід – класифікація спектрів із використанням згорткової нейронної мережі.

Архітектура CNN. Розроблена модель представляла одновимірну згорткову нейронну мережу, яка приймає на вході спектр (послідовність інтенсивностей при різних частотах) і повертає ймовірність належності до того чи іншого класу. Мережа містила декілька послідовних згорткових шарів для автоматичного виокремлення ознак: зокрема, три згорткові шари (перший на 128 фільтрів з ядром 5 точок, другий – 256 фільтрів з ядром 5, третій – 128 фільтрів з ядром 3). Після кожного з перших двох згорткових шарів застосовувався шар підвибірки (MaxPooling з вікном 2) для зменшення розмірності представлення, а також шар нормалізації (Batch Normalization) і відсікання (Dropout 20%) для запобігання перенавчанню. Третій згортковий шар підключався до шару глобального усереднення (Global Average Pooling), який перетворював кожену

карту ознак на одне середнє значення – у такий спосіб здійснювали агрегацію локальних ознак по всьому спектру. Далі слідували повнозв'язні шари (багатошаровий перцептрон) для інтеграції виявлених ознак і прийняття рішення. Зокрема, використовували 2 приховані щільні шари (кількість нейронів поступово зменшувалася, наприклад, половина від числа вхідних ознак, потім чверть) з активацією *tanh*. На виході мережі стояв нейрон із сигмоїдною активацією, що генерував ймовірність належності спектра до класу «патологія» або «норма». У задачах двокласової класифікації (наприклад, «рак vs контроль») сигмоїдний вихід інтерпретували як ймовірність «раку», а поріг прийняття рішення встановлювали 0,5. Навчання CNN здійснювали методом зворотного поширення помилки з використанням втрат бінарної крос-ентропії та оптимізатора Adam (навчальна швидкість 3×10^{-4}). Для врахування можливого дисбалансу вибірки (різна кількість спектрів у класах) застосували метод штучного збалансування – *oversampling* менш представленого класу за допомогою алгоритму SMOTE (Synthetic Minority Over-sampling Technique) перед поділом даних на тренувальну і тестову вибірки. Модель тренували на 70% даних, перевіряли на 30% (рандомізований поділ, фіксований seed для відтворюваності). У процесі навчання використовували *Early Stopping*: якщо значення функції втрат на валідації не покращувалося

протягом 10 епох, тренування зупиняли достроково, щоб уникнути перенавчання. Основними показниками ефективності CNN обрано точність класифікації (асигуру) та площа під *ROC-кривою* (*AUC*), а також аналізували чутливість і специфічність моделі для виявлення патології.

Результати та їх обговорення

Класифікація спектрів. Застосування неконтрольованих методів кластерного аналізу не виявило чіткої природної сегрегації спектрів на групи, що відповідали б клінічним категоріям пацієнтів. Візуалізація даних у просторі головних компонент та за допомогою t-SNE показала значне перекриття між спектрами різних груп: точки, що належать здоровим контролям, передраку та раку, не утворювали окремих ізольованих скупчень, а розподілялися без явних меж (рис. 1). Алгоритм K-середніх при $k=4$ формально розділив вибірку на чотири кластери, проте ці кластери не збіглися з реальними діагнозами. Лише невелика частина спектрів раку та норми потрапила до однорідних кластерів, тоді як більшість кластерів виявилися змішаними (містили спектри різних груп) (рис. 2). Коефіцієнт силуету для розбиття K-Means був близьким до нуля ($\sim 0,1$), що засвідчує про слабку відокремленість кластерів: багато точок мали від'ємні значення силуету, тобто були ближчими до центрів інших кластерів, ніж до власних.

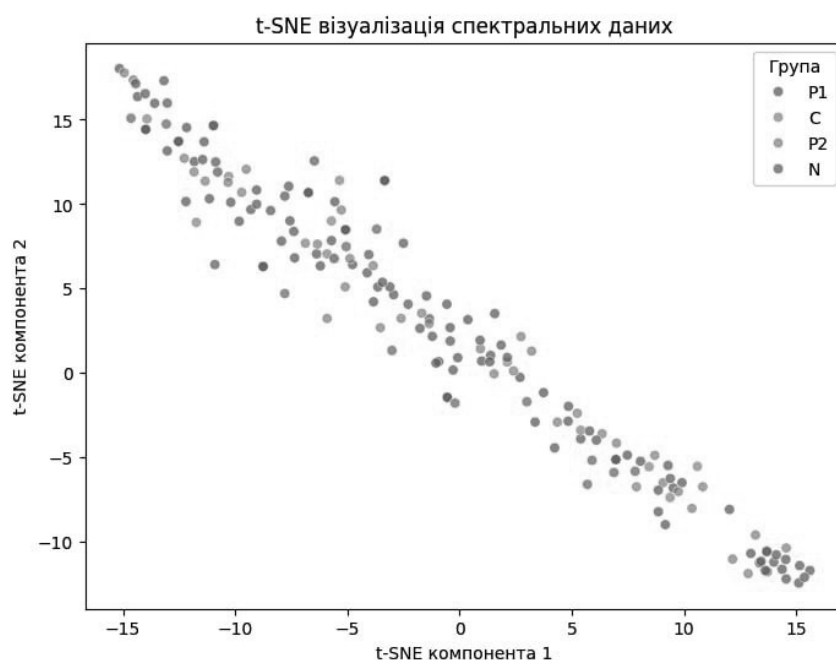


Рис. 1. Графік демонструє результат зменшення розмірності спектральних даних методом t-SNE. Групи не формують чітко розділені кластери, що вказує на слабку здатність кластеризаційних алгоритмів розрізняти спектри

Скоригований індекс Ренда (ARI) для кластеризації K-середніх становив лише $\sim 0,2$, що є дуже низьким і фактично недалеко від рівня випадкового розподілу ($ARI=0$). Інші методи кластеризації показали аналогічно невітні результати. Алгоритм DBSCAN при різних параметрах щільності або знаходив лише один великий кластер (об'єднуючи більшість спектрів разом, а решту позначаючи як шум), або ділив дані на численні дрібні кластери без діагностичної інтерпретації. Метод

гаусових сумішей (GMM) при заданні 4-х компонент також не зміг правильно розділити спектри: ARI залишався низьким ($\sim 0,1-0,3$), а силует – від'ємним, вказуючи на перекриття компонент. Ієрархічний алгоритм не показав природного розриву на чотири кластери – дендрограма не містила рівня, на якому б чітко відділялися саме групи контроль/лишай/лейкоплакія/рак. Навіть спроба попередньо зменшити розмірність ознак (через PCA або автоенкодер) не

призвела до утворення кластерів, що відповідали б здоровим чи патологічним станам (рис. 3).

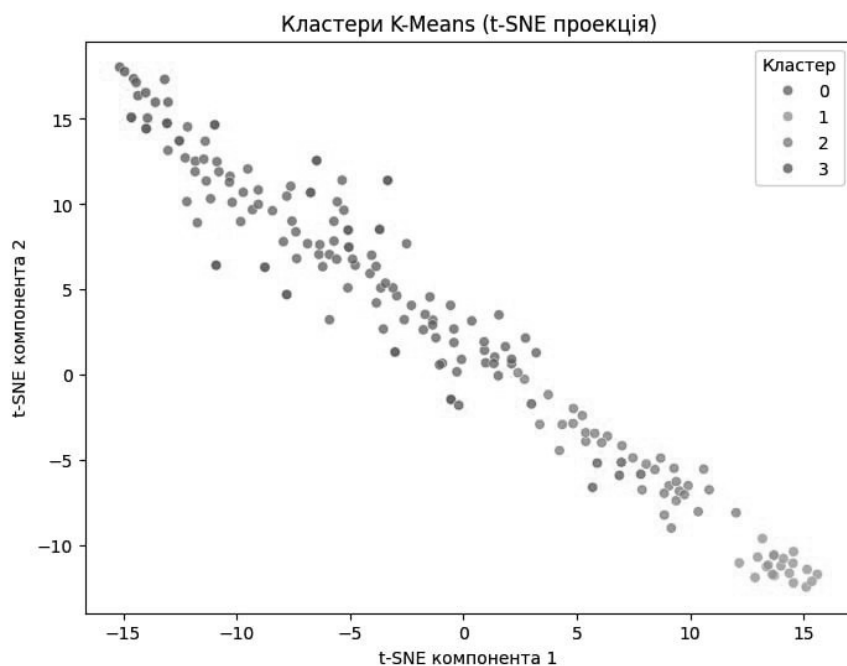


Рис. 2. Кластери, які знайшов алгоритм K-Means після зменшення розмірності t-SNE. Кластери не збігаються з реальними групами пацієнтів, що вказує на неефективність класифікації

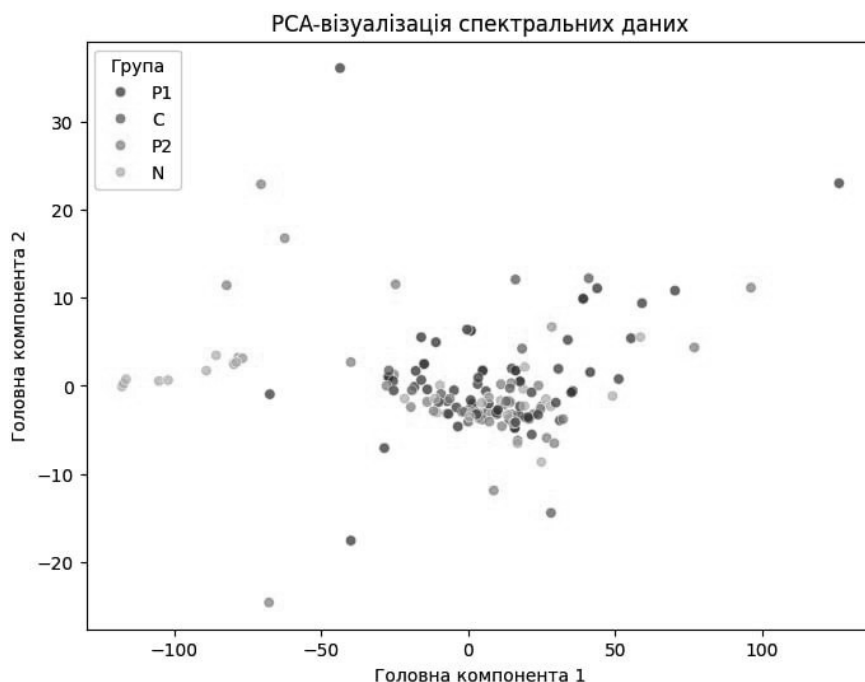


Рис. 3. Розташування спектральних даних після зменшення розмірності методом головних компонент (PCA). Як видно, групи значно перекриваються, що ускладнює їх автоматичний поділ

Отже, жоден із випробуваних підходів неконтрольованого навчання не виявив «природних» кластерів у спектральних даних, які б корелювали з клінічними групами. Кількісні метрики кластеризації (силует, ARI) були низькими для всіх алгоритмів, підтверджуючи, що поділ спектрів на здорові, передракові та ракові зразки неможливий без залучення апіорної інформації про мітки. Іншими словами, спектральні відмінності між групами

виявилися недостатніми для того, щоб алгоритми самостійно їх розпізнали і згрупували відповідні зразки разом.

Класифікація за допомогою CNN. На відміну від неконтрольованих методів, згортова нейронна мережа показала високу ефективність у розпізнаванні спектрів різних груп. Навчена на мічених даних модель 1D-CNN змогла правильно класифікувати переважну більшість зразків

Клінічна та експериментальна патологія. 2025. Т.24, № 1 (91)

у тестовій вибірці. Загальна точність класифікації становила близько 90-95% (залежно від конкретного завдання та спектрального діапазону). Зокрема, при розмежуванні ракових і здорових спектрів модель досягала точності ~94-95%, зі значенням AUC (площа під ROC-кривою) ~0,96-0,97. Це означає, що CNN практично безпомилково відрізняла слину хворих на рак від здорової, демонструючи високу чутливість та специфічність (понад 90%). Для диференціації передраку (лейкоплакії) від норми показники були дещо нижчими, що очікувано через меншу

вираженість відмінностей між цими станами, проте все одно досить високими: точність сягала ~88-90%, AUC ~0,90. Задача відрізнити рак від передраку також успішно вирішувалася мережею (точність ~90%). Отже, навіть найскладніші для виявлення випадки CNN розпізнавала набагато краще, ніж це можна було б очікувати від випадкового вгадування або кластерного підходу. При цьому, важливо зазначити, що вища точність роботи була характерна для аналізу повного спектру, а не окремих діапазонів довжин хвиль (рис. 4).

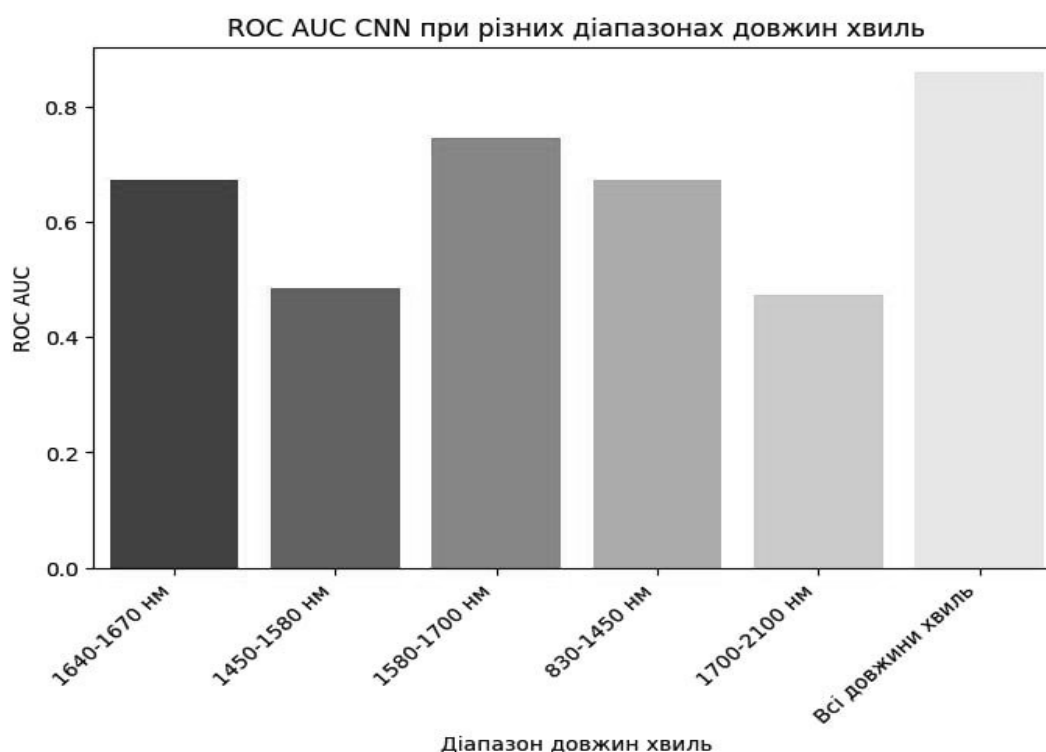


Рис. 4. Порівняння точності роботи CNN на різних діапазонах довжин хвиль на всьому спектрі. Графік демонструє, що AUC є найвищим для повного спектру, підтверджуючи перевагу аналізу всього спектру замість обмежених діапазонів

Для порівняння, ефективність кластеризації наближалася до 0 (ARI ~0 при 50% випадків правильно класифікованих у кращому разі), тобто була фактично непридатною для практичної діагностики, тоді як CNN забезпечила надійне розрізнення класів. Важливо відзначити, що модель CNN показала стійкість до шумів та варіабельності даних: при штучному додаванні невеликого рівня шуму до спектрів точність класифікації знижувалася несуттєво (на кілька відсотків). Це свідчить про те, що нейронна мережа навчилася виділяти стійкі ознаки раку, які залишаються помітними навіть за наявності перешкод. Отримані результати продемонстрували явну перевагу контрольованого підходу з глибоким навчанням над неконтрольованою кластеризацією для задачі спектрального скринінгу раку ротової порожнини.

Результати нашого дослідження показують, що без використання навчання на мічених даних поділ спектрів слини на здорові та патологічні не реалізується належним чином. Кластерні алгоритми припускають існування в наборі даних природних

груп, які об'єднують схожі зразки і відокремлені від інших. Проте у випадку раманівських спектрів біологічних зразків виявилось, що природні чіткі кластери відсутні. Біохімічні зміни при переході від норми до передраку і раку хоча й існують, але відносно поступові і не створюють радикально відмінних «кластерів» спектрів – натомість спостерігається значне перекриття характеристик. Іншими словами, спектри слини здорових людей та хворих можуть відрізнятися інтенсивністю окремих смуг, але загальна варіабельність між різними людьми (зумовлена індивідуальними особливостями, дієтою, мікробіомом тощо) настільки велика, що міжкласові відмінності тонуть у міжіндивідуальному шумі. Цей фактор пояснює, чому підходи типу K-Means або GMM не спрацювали: вони мінімізують внутрішньокластерне розсіювання, але коли варіабельність усередині кожної групи (наприклад, різні здорові люди) співрозмірна з варіабельністю між групами, алгоритм не може знайти стабільний поділ, що відповідав би діагностичним категоріям. Більше того, метод K-Means та пов'язані з ним

припускають певну форму кластерів (сферичні хмари точок у просторі ознак). У наших даних межі між спектрами здорових і хворих не були сферичними – вони переплітались, формуючи радше континуум, ніж дискретні групи. *DBSCAN* вимагав би наявності щільних ділянок спектрів, чітко відділених прогалинами; на практиці ж спектри пацієнтів не утворили таких ізольованих ділянок – відмінності між класами виявилися меншими за внутрішньогрупові коливання, тож весь набір виглядав як один згусток точок із незначними градієнтними змінами всередині нього. Схожа ситуація спостерігається і в інших дослідженнях біоспектроскопії: без учителя алгоритми кластеризації здатні виділити хіба що найбільш екстремальні випадки або шумові артефакти, але не систематично розмежувати патологію [6].

Інша причина невдачі неконтрольованих методів – висока розмірність спектральних даних (сотні змінних відповідно до чисел хвиль). У такому просторі поняття відстані між точками стає менш інформативним через феномен так званого «прокляття розмірності». *РСА* та автоенкодер частково пом'якшили цю проблему, стиснувши дані до кількох узагальнених ознак. Однак ці ознаки оптимізувалися на відтворення всього спектру, а не спеціально на розрізнення класів. Тому навіть у стислому просторі кластери здорових/хворих все одно залишилися переплутаними. Отже, без інформації про мітки алгоритми не змогли знайти ті слабкі патерни, які відрізняють спектри раку від норми.

Натомість, згортова нейронна мережа виявилася здатною виокремити саме ті характеристики спектрів, які релевантні для діагностики. Перевага *CNN* полягає в наявності етапу навчання на прикладах з відомою відповіддю: мережа «знає», які зразки належать до раку, а які – до норми, і поступово підлаштовується, щоб максимізувати правильність класифікації. У цьому процесі модель фактично виконує автоматичний відбір ознак – з усіх особливостей спектра вона навчається фокусуватися на тих, що систематично відрізняються між класами, і ігнорувати варіабельність, не пов'язану з діагнозом. В результаті *CNN* будує нелінійне рішення (високовимірну поверхню розділу) між класами, яке враховує комплекс дрібних відмінностей у спектрах. Наш експеримент показав, що мережа використовувала всю доступну спектральну інформацію: наприклад, вона враховувала і зміну відносної інтенсивності піків білків (амід I ~ 1650 cm^{-1}) і ліпідів (~ 1445 cm^{-1}), і появу додаткових «плечей» чи зсувів у передракових спектрах, і зниження смуг нуклеїнових кислот (близько 1090 cm^{-1}) у ракових зразках. Більшість із цих ознак окремо є ледь помітними і статистично незначущими, але *CNN* змогла комбінувати їх в єдиний предиктор. Це узгоджується з висновками інших дослідників: глибокі нейронні мережі завдяки своїй здатності враховувати нелінійні взаємозв'язки перевершують методи на кшталт лінійного дискримінантного аналізу чи опорних векторів при класифікації спектральних даних раку [3, 6].

Зокрема, в огляді Blake та співавт. (2022) зазначено, що *CNN* стабільно показують вищу точність, ніж традиційні алгоритми, для розпізнавання різних видів раку за раманівськими спектрами [3]. Наші результати підтверджують цю тенденцію: *CNN* успішно виявила патерни, які не були очевидними при пороговому статистичному аналізі та недоступні для неконтрольованого кластерного групування.

Варто підкреслити, що висока точність і чутливість, досягнуті *CNN* у цьому дослідженні, узгоджуються з даними сучасної літератури. Наприклад, Li та ін. (2023) повідомляють, що поєднання оптиволоконної Раман-спектроскопії з глибокою *CNN* дозволило автоматично виявляти та класифікувати рак ротової порожнини з точністю $\sim 94\text{--}95\%$, що є співставним з результатами гістопатологічного аналізу [5]. Це демонструє, що використання глибокого навчання робить неінвазивний скринінг максимально наближеним за ефективністю до «золотого стандарту» діагностики. Водночас, наші негативні результати щодо кластеризації дають важливий урок: відсутність явних кластерів не означає відсутності діагностичної інформації у спектрах. Просто ця інформація «розподілена» по багатьох змінних і потребує складної нелінійної комбінації, щоб її виокремити. Глибоке навчання здатне віднайти таку комбінацію, тоді як прості неконтрольовані методи – ні. Отже, для вирішення задачі раннього виявлення раку за допомогою Раман-спектроскопії більш доцільним виявився підхід із контрольованим навчанням (*CNN*), який враховує апріорні знання про класи і вчиться на них, замість спроб покладатися на ймовірну наявність природних кластерів у сирих даних.

Висновки

Проведене дослідження демонструє, що для скринінгу раку ротової порожнини за раманівськими спектрами слини значно кращим підходом є використання контрольованих алгоритмів глибокого навчання, ніж неконтрольованої кластеризації. Класичні методи кластерного аналізу не змогли розмежувати спектри здорових, передракових та ракових зразків – жоден з алгоритмів (*PCA*, *t-SNE*, *K-Means*, *DBSCAN*, *GMM*, ієрархічний, автоенкодер + *KMeans*) не виявив природного групування, яке відповідало б цим діагностичним категоріям. Низькі значення *ARI* і коефіцієнту силуєту підтверджують, що без учителя спектри різних станів надто схожі та перекриваються, щоб утворити окремі кластери. Натомість, запропонована модель *1D-CNN*, навчена на мічених даних, досягла високої точності (близько 90% і більше) у класифікації спектрів, впевнено відрізняючи злоякісні зразки від доброякісних. *CNN* перевершила традиційні підходи за чутливістю і специфічністю, оскільки автоматично вичленувала комплексні «підписи» раку в усьому спектральному діапазоні. Отже, для практичного неінвазивного виявлення ранніх стадій РРП аналіз повноспектральних даних за допомогою глибокої нейронної мережі є більш ефективним і надійним

шляхом, ніж спроби кластеризувати спектри без наглядю.

Перспективи подальших досліджень

Отримані результати окреслюють напрями для подальших досліджень і удосконалення методики. По-перше, варто розширити набір ознак для неконтрольованого аналізу, включивши до нього клініко-анамнестичні дані пацієнтів. Можливо, додавання таких змінних як вік, стать, наявність факторів ризику (паління, вживання алкоголю), результати візуального огляду тощо, у багатовимірний простір даних дозволить кластерам краще відображати суттєві відмінності між групами. Повторна кластеризація з урахуванням спектральних та неспектральних параметрів може виявити приховані підгрупи або взаємозв'язки, непомітні лише зі спектрів. По-друге, планується збільшити вибірку пацієнтів, особливо за рахунок додаткових передракових станів і ранніх стадій раку, щоб покращити навчання моделей і перевірити їх узагальнюваність. Розширення вибірки та різноманітності даних (включно зі зразками слини з різних географічних регіонів, різним складом їжі тощо) допоможе зробити висновки більш надійними. По-третє, перспективним є *in vivo* збір спектрів безпосередньо зі слизової оболонки рота за допомогою волоконної Раман-оптики – це дозволить перевірити працездатність запропонованого підходу в клінічних умовах і оцінити його придатність для масового скринінгу. Нарешті, можна розглянути інтеграцію спектральних даних з іншими діагностичними модальностями: комбінування раманівських спектрів із флуоресцентними зображеннями чи молекулярними маркерами може підвищити точність моделі. Загалом, подальший розвиток систем штучного інтелекту для аналізу біоспектроскопічних даних відкриває шлях до створення об'єктивних, експресних і доступних методів раннього виявлення раку ротової порожнини, що особливо важливо для впровадження у широку медичну практику та в регіонах з обмеженими ресурсами.

Список літератури

1. Abati S, Bramati C, Bondi S, Lissoni A, Trimarchi M. Oral Cancer and Precancer: A Narrative Review on the Relevance of Early Diagnosis. *Int J Environ Res Public Health* [Internet]. 2020[cited 2025 Apr 10];17(24):9160. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7764090/pdf/ijerph-17-09160.pdf> doi: 10.3390/ijerph17249160
2. Faur CI, Dinu C, Toma V, Jurj A, Mărginean R, Onaciu A, et al. A New Detection Method of Oral and Oropharyngeal Squamous Cell Carcinoma Based on Multivariate Analysis of Surface Enhanced Raman Spectra of Salivary Exosomes. *J Pers Med* [Internet]. 2023[cited 2025 Apr 07];13(5):762. Available

from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10219614/pdf/jpm-13-00762.pdf> doi: 10.3390/jpm13050762

3. Blake N, Gaifulina R, Griffin LD, Bell IM, Thomas GMH. Machine Learning of Raman Spectroscopy Data for Classifying Cancers: A Review of the Recent Literature. *Diagnostics (Basel)* [Internet]. 2022[cited 2025 Apr 07];12(6):1491. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9222091/pdf/diagnostics-12-01491.pdf> doi: 10.3390/diagnostics12061491
4. Sharma M, Jeng MJ, Young CK, Huang SF, Chang LB. Developing an Algorithm for Discriminating Oral Cancerous and Normal Tissues Using Raman Spectroscopy. *J Pers Med* [Internet]. 2021[cited 2025 Apr 10];11(11):1165. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8623962/pdf/jpm-11-01165.pdf> doi: 10.3390/jpm11111165
5. Li X, Li L, Sun Q, Chen B, Zhao C, Dong Y, et al. Rapid multi-task diagnosis of oral cancer leveraging fiber-optic Raman spectroscopy and deep learning algorithms. *Front Oncol* [Internet]. 2023[cited 2025 Apr 10];13:1272305. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10597702/pdf/fonc-13-1272305.pdf> doi: 10.3389/fonc.2023.1272305
6. Luo R, Popp J, Bocklitz T. Deep Learning for Raman Spectroscopy: A Review. *Analytica* [Internet]. 2022[cited 2025 Apr 07];3(3):287-301. Available from: <https://www.mdpi.com/2673-4532/3/3/20> doi: 10.3390/analytica3030020

References

1. Abati S, Bramati C, Bondi S, Lissoni A, Trimarchi M. Oral Cancer and Precancer: A Narrative Review on the Relevance of Early Diagnosis. *Int J Environ Res Public Health* [Internet]. 2020[cited 2025 Apr 10];17(24):9160. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7764090/pdf/ijerph-17-09160.pdf> doi: 10.3390/ijerph17249160
2. Faur CI, Dinu C, Toma V, Jurj A, Mărginean R, Onaciu A, et al. A New Detection Method of Oral and Oropharyngeal Squamous Cell Carcinoma Based on Multivariate Analysis of Surface Enhanced Raman Spectra of Salivary Exosomes. *J Pers Med* [Internet]. 2023[cited 2025 Apr 07];13(5):762. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10219614/pdf/jpm-13-00762.pdf> doi: 10.3390/jpm13050762
3. Blake N, Gaifulina R, Griffin LD, Bell IM, Thomas GMH. Machine Learning of Raman Spectroscopy Data for Classifying Cancers: A Review of the Recent Literature. *Diagnostics (Basel)* [Internet]. 2022[cited 2025 Apr 07];12(6):1491. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9222091/pdf/diagnostics-12-01491.pdf> doi: 10.3390/diagnostics12061491
4. Sharma M, Jeng MJ, Young CK, Huang SF, Chang LB. Developing an Algorithm for Discriminating Oral Cancerous and Normal Tissues Using Raman Spectroscopy. *J Pers Med* [Internet]. 2021[cited 2025 Apr 10];11(11):1165. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8623962/pdf/jpm-11-01165.pdf> doi: 10.3390/jpm11111165
5. Li X, Li L, Sun Q, Chen B, Zhao C, Dong Y, et al. Rapid multi-task diagnosis of oral cancer leveraging fiber-optic Raman spectroscopy and deep learning algorithms. *Front Oncol* [Internet]. 2023[cited 2025 Apr 10];13:1272305. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10597702/pdf/fonc-13-1272305.pdf> doi: 10.3389/fonc.2023.1272305
6. Luo R, Popp J, Bocklitz T. Deep Learning for Raman Spectroscopy: A Review. *Analytica* [Internet]. 2022[cited 2025 Apr 07];3(3):287-301. Available from: <https://www.mdpi.com/2673-4532/3/3/20> doi: 10.3390/analytica3030020

Інформація про авторів:

Говорняк С. Л. – асистент кафедри онкології та радіології Буковинського державного медичного університету, м. Чернівці, Україна.

E-mail: hovornyan@bsmu.edu.ua

ORCID ID: <https://orcid.org/0000-0003-2766-2218>

Івашук О. І. – д.мед.н., професор кафедри онкології та радіології Буковинського державного медичного університету, м. Чернівці, Україна.

E-mail: ivashchuk@bsmu.edu.ua

ORCID ID: <https://orcid.org/0000-0003-1747-2648>

Information about authors:

Hovornyan S. – Assistant, Department of Oncology and Radiology, Bukovinian State Medical University, Chernivtsi, Ukraine.

E-mail: hovornyan@bsmu.edu.ua

ORCID ID: <https://orcid.org/0000-0003-2766-2218>

Ivashchuk O. – MD, Professor, Department of Oncology and Radiology, Bukovinian State Medical University, Chernivtsi, Ukraine.

E-mail: ivashchuk@bsmu.edu.ua

ORCID ID: <https://orcid.org/0000-0003-1747-2648>

Стаття надійшла до редакції 27.02.2025

© С. Л. Говорнян, О. І. Івашук

