УДК 004.93’1:004.85:159.942.5 **DOI:** 10.31673/2412-9070.2026.026805

Т. М. КИСІЛЬ, аспірант,
ORCID: 0000-0002-5123-0768
Державний університет інформаційно-комунікаційних технологій, Київ

МУЛЬТИМОДАЛЬНА СИСТЕМА РОЗПІЗНАВАННЯ ЕМОЦІЙНИХ СТАНІВ У РЕЖИМІ РЕАЛЬНОГО ЧАСУ НА ОСНОВІ ГЛИБОКИХ ЗГОРТКОВИХ НЕЙРОМЕРЕЖ

У статті обґрунтовано концепцію мультимодальної інтелектуальної системи розпізнавання емоційних станів людини в режимі реального часу на основі динамічних тривимірних згорткових нейронних мереж 3D-CNN та методів некерованого машинного навчання. Наукова новизна роботи полягає у розробці підходу для автономного вилучення просторово-часових ознак, що мінімізує залежність від попередньо розмічених наборів даних та підвищує адаптивність системи до індивідуальних особливостей невербальної експресії. Вагомим внеском являється технічне вдосконалення інструментарію MediaPipe через інтеграцію модифікованого алгоритму детекції антропометричних маркерів, спеціалізованого на виявленні специфічних патернів дитячої міміки.

Методологічне рішення базується на двоканальній архітектурі для паралельного аналізу міміки та кінематики тіла, що дозволяє нівелювати ризики оклюзії та складних ракурсів зйомки. Використання 3D-CNN забезпечує обробку відеоданих як цілісних просторово-часових структур, а процес автоматичної генерації псевдоміток у межах некерованої кластеризації латентного простору дає змогу системі автономно структурувати базові емоційні категорії. Експериментальна апробація із застосуванням стратегії пізнього злиття модальностей Late Fusion підтвердила стійкість моделі до зашумлених сигналів та складного освітлення. Результати дослідження доводять, що запропонована модель забезпечує високу швидкість обробки даних для функціонування в реальному часі, що робить її придатною для впровадження в інтелектуальні освітні платформи, педіатричну діагностику та системи безпеки.

Ключові слова: розпізнавання емоцій, мультимодальна система, 3D-CNN, некероване навчання, пізнє злиття, реальний час, комп’ютерний зір, невербальна поведінка, кластеризація ознак, міміка обличчя.

Вступ

Сучасний розвиток інтелектуальних систем наступного покоління потребує вдосконалення технологій мультимодального розпізнавання емоцій, що здатні одночасно інтерпретувати міміку та жестикуляцію для контекстуально точного розуміння поведінки людини. У межах запропонованого дослідження реалізовано метод розпізнавання станів за обличчям та позою в режимі реального часу, заснований на архітектурі пізнього злиття з використанням автономних тривимірних згорткових нейронних мереж (3D-CNN).

Постановка проблеми. Системи розпізнавання емоцій стикаються з критичними обмеженнями через надмірну залежність від вручну розмічених даних та фокусування виключно на міміці обличчя, що призводить до суб’єктивізму оцінок і втрати важливого контексту динаміки рухів тіла. Використання стандартних архітектур часто не забезпечує повноцінної інтеграції просторово-часових ознак без обчислювальних затримок, що стає бар’єром для роботи у реальному часі під впливом оклюзії, змінного освітлення або індивідуальних особливостей експресії суб’єкта.

Подолання цих викликів потребує переходу до некерованих методів навчання на основі тривимірних згорткових нейронних мереж (3D-CNN), які дозволяють автономно вилучати релевантні ознаки та уніфікувати аналіз міміки й пози в єдиному контенті. Застосування такого

підходу мінімізує потребу у попередній розмітці даних і суттєво підвищує адаптивність інтелектуальних систем до мінливих умов зовнішнього середовища, забезпечуючи високу стійкість класифікації емоційних станів осіб.

Аналіз останніх досліджень і публікацій. Сучасний етап розвитку мультимодальних систем розпізнавання емоцій (MER) характеризується еволюцією від базових CNN-LSTM структур до складних ієрархічних архітектур, трансформерів та механізмів уваги, що детально описано у систематичному огляді Wu, Mi та Gao [6]. Дослідження Wu, Cai та Liu [5] підкреслюють потенціал графових нейронних мереж (GNN) та великих мультимодальних мовних моделей (MLLMs) у контексті адаптації під конкретні задачі через дистиляцію знань, тоді як Suryawanshi та Pansare [4] демонструють ефективність синергії візуальних образів та текстової тональності для детекції станів у реальному часі. Важливий фундамент для інтеграції мови тіла та скелетних даних закладено Rivera, Rodrigues та Fugita [3] у наборі даних BER2024, а також Yu, Li та Zhou [8], які впровадили алгоритми на основі ядерного канонічного кореляційного аналізу (SLSMKCCA) для синхронізації міміки, мовлення та жестів.

Технологічний поступ у моделюванні 3D-рухів людини за допомогою LLMs представлений Chen, Zhang та Adeli [1], у той час як Yoon та Kim [7] розробили мережу EmotionTFN для розгортання в обмежених середовищах IoT через квантування моделей. Практичну реалізацію високошвидкісних систем демонструють Taherkhani через ансамблі Mini-XCEPTION та Liu, Xu, Guo у системі SyncAnimation [2], спираючись на інструментарій MediaPipe [9]. Вагоме місце посідає праця науковців [10], [11], де обґрунтовано ефективність CNN для класифікації образів та реалізовано методи відстеження об'єктів у відеопотоці, що забезпечує перехід від аналізу статичних структур до високоточного опрацювання складних просторово-часових послідовностей у динамічних сценах.

Виклад основного матеріалу

Запропонований метод мультимодального некерованого розпізнавання емоцій (*MMHPE*) базується на розробці інтелектуальної системи, що здійснює синхронний аналіз мімічної експресії та кінематики рухів тіла за допомогою архітектури *3D-CNN*. Ключовою особливістю підходу є використання парадигми некерованого навчання, що дозволяє моделі автономно структурувати просторово-часові ознаки відеопотоку без попередньої ручної розмітки даних, забезпечуючи високу адаптивність до індивідуальної невербальної поведінки. Важливою складовою методу є модифікована логіка детекції антропометричних точок на базі MediaPipe, спеціально адаптована для ідентифікації вікових особливостей, зокрема прецизійного виявлення дитячої посмішки через аналіз динамічної кривизни ключових точок обличчя.

Методологічне рішення реалізовано через двоканальну архітектуру, де вхідний відеопотік розділяється на паралельні гілки аналізу обличчя та пози після попереднього виділення областей інтересу. Для об’єднання результатів цих модальностей застосовано стратегію пізнього злиття (*Late Fusion*), яка здійснює динамічне зважування кожної гілки обробки залежно від якості сигналу та зовнішніх умов. Такий комплексний підхід гарантує стійкість системи до оклюзій та змін освітлення, забезпечуючи високу точність класифікації емоційних станів у режимі реального часу.

Канал розпізнавання обличчя використовує тривимірну згорткову нейронну мережу (*3D-CNN*) для обробки відео-тензорів, що дозволяє фіксувати часову динаміку мікрорухів м’язів та інтегрувати блок аналізу вікових паттернів. Паралельно працює гілка кінематики пози, яка на основі аналогічної архітектури досліджує просторову конфігурацію скелетної моделі з 33 ключових точок, забезпечуючи стабільність системи навіть за умов оклюзії обличчя або недостатнього освітлення.

Інтеграція даних здійснюється через механізм пізнього злиття (*Late Fusion*), де блок агрегації динамічно зважує внесок кожної модальності залежно від рівня шуму в каналах. Фінальний етап базується на принципах некерованого навчання, що дозволяє системі автономно кластеризувати латентний простір ознак і класифікувати емоційні стани за сімома базовими категоріями. Така структура забезпечує високу швидкість обробки даних та адаптивність методу до складних природних умов експлуатації в режимі реального часу.

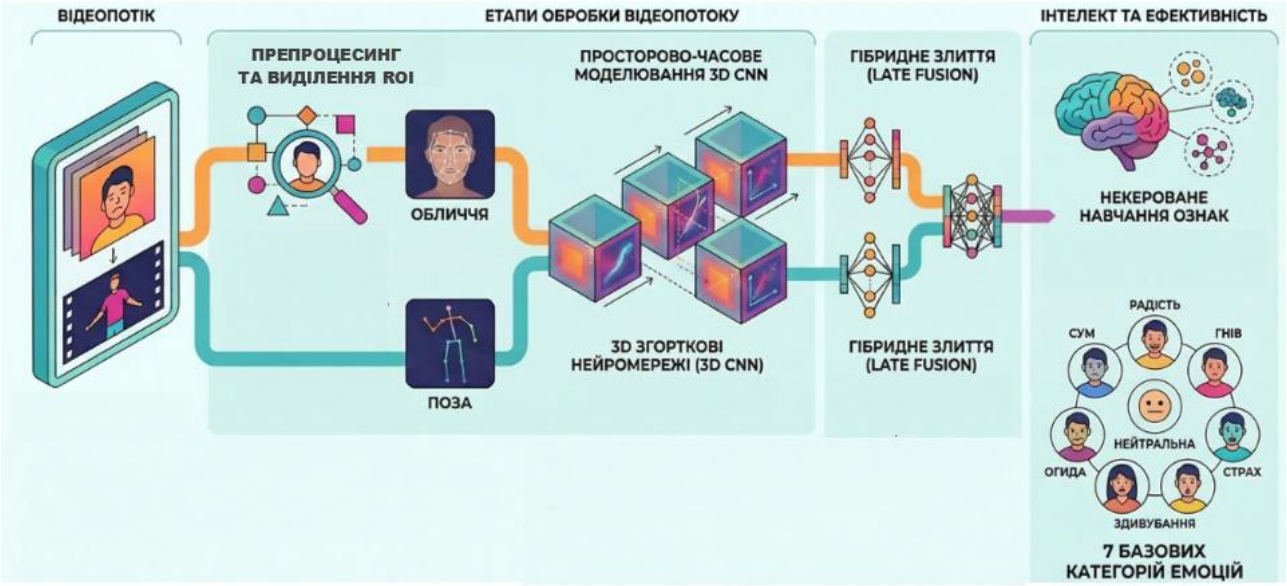

Рис. 1. Архітектура мультимодальної системи некерованого розпізнавання емоцій відеопотоків

Математична модель методу мультимодального некерованого розпізнавання емоцій (ММНРЕ) базується на формалізації перетворення вхідного відеопотоку V у латентний простір емоційних станів через послідовні етапи просторово-часової фільтрації та динамічної агрегації ознак. Вхідний сигнал представляється як послідовність кадрів $V = \{I_1, I_2, \dots, I_T\}$, де кожен кадр $I \in R^{H \times W \times C}$ проходить етап препроцесингу для формування двох вхідних тензорів: тензора міміки обличчя X_f та тензора скелетної моделі пози X_p.

Узагальненою математичною операцією кожного каналу являється тривимірна згортка, яка дозволяє вилучати ознаки одночасно у просторових та часових вимірах. Для кожного *l*-го прошарку нейронної мережі значення активації у позиції *(x, y, z)* для *j*-ї карти ознак визначається формулою (1).

$$v_{i,j}^{(x,y,z)} = \sigma\left(\sum_{m=0}^{M_{i-1}-1} \sum_{h=0}^{H_k-1} \sum_{w=0}^{W_k-1} \sum_{t=0}^{T_k-1} w_{ijm}^{(h,w,t)} \cdot v_{(i-1)m}^{(x+h,y+w,z+t)} + b_{ij} \right) \quad (1)$$

У даній моделі $w_{ijm}^{(h,w,t)}$ позначає ваговий коефіцієнт ядра згортки розміром $H_k \times W_k \times T_k$, під'єднаного до *m*-ї карти ознак попереднього прошарку, $b_{\{ij\}}$ являється зміщенням, тоді як σ - нелінійною функцією активації. Саме таке представлення забезпечує збереження динамічних паттернів, включаючи специфічні вікові характеристики міміки, що формалізуються через високорівневі дескриптори $\phi_f(X_f)$ та $\phi_p(X_p)$.

Інтеграція даних реалізується через математичний апарат пізнього злиття, де фінальний вектор ознак Z формується як зважена комбінація виходів обох каналів. Процес обчислення вихідного вектора описується рівнянням (2).

$$Z = \alpha \cdot \phi_f(X_f) \oplus (1 - \alpha) \cdot \phi_p(X_p) \quad (2)$$

У результаті виконання конкатенації, коли коефіцієнт $\alpha \in [0,1]$ являється параметром динамічного зважування, адаптивно обчислюється на основі оцінки впевненості предиктора для кожної модальності. Саме це дозволяє моделі автоматично мінімізувати вплив зашумленого каналу (наприклад, при оклюзії обличчя) на фінальний результат класифікації.

Навчання системи без участі вчителя базується на мінімізації цільової функції втрат, яка стимулює формування стійких кластерів у латентному просторі. Для етапу некерованого навчання ознак у методі ММНРЕ використовується контрастивна функція втрат $\mathcal{L}_{cont}$, яка максимізує подібність між різними аугментаціями одного відео та мінімізує її для різних зразків.

Алгоритм формування позитивної пари z_i^+ виступає фундаментальним етапом функціонування контрастивної функції втрат $\mathcal{L}_{cont}$, оскільки саме на його основі визначаються критерії

семантичної ідентичності представлень даних, які обробляються мережею. У межах запропонованого методу даний процес реалізується через принцип інваріантності емоційного стану відносно варіативних спотворень вхідного сигналу.

Процедура кластеризації виступає завершальним етапом методу MMHPE, забезпечуючи перетворення неструктурованих векторів ознак у чітко визначені емоційні категорії. Після того, як контрастивна функція втрат сформувала компактні та розділені представлення у латентному просторі, виконується фінальна класифікація за сімома базовими емоціями на основі припущення, що вектори Z групуються навколо сформованих центроїдів, які відповідають конкретним психологічним станам. Процес розпочинається з ініціалізації центроїдів після завершення навчання на всьому масиві даних, для виділення категорій N мультимодальних векторів Z застосовується алгоритм k-середніх із фіксованою кількістю кластерів $k=7$. Оскільки початкове навчання є некерованим, сформовані кластери в процесі обробки не мають ідентифікації, то залучається обмежений калібрувальний набір даних, що дозволяє співставити кожен кластер з однією з категорій, таких як радість, здивування, гнів, страх, огида, сум або нейтральний стан. Такий підхід дозволяє методу MMHPE автономно виділяти інваріантні характеристики емоцій, забезпечуючи стійкість до індивідуальних особливостей зовнішності суб'єкта (рис. 2).

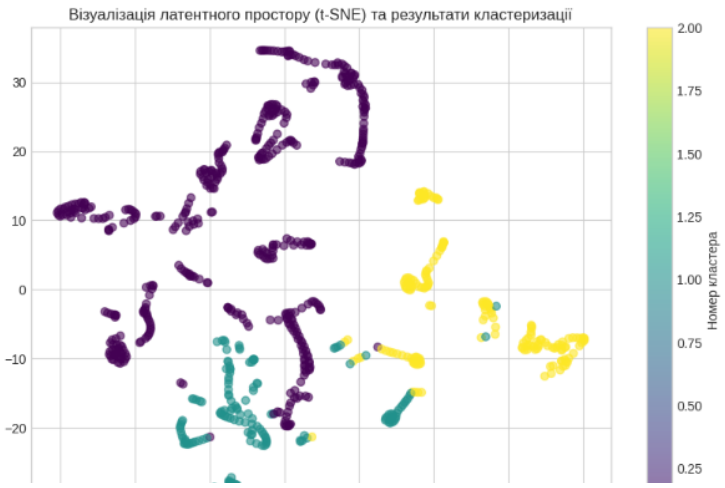

Рис. 2. Візуалізація структури латентного простору та результатів кластеризації мультимодальних ознак емоцій методом t-SNE

Візуалізація латентного простору після навчання методу MMHPE демонструє формування чітких компактних кластерів, що підтверджує ефективність контрастивної функції втрат у процесі некерованого групування подібних емоційних станів. Автономне структурування мультимодальних даних у виражені кольорові групи свідчить про здатність моделі ідентифікувати інваріантні закономірності поведінки без попередньої розмітки, створюючи надійне підґрунтя для визначення центроїдів базових емоцій.

Загальний процес розбиття відеопотоку на дискретні фази дозволяє простежити послідовну зміну станів активності та емоційних реакцій у часовому вимірі. Синхронізація відеокадрів із відповідною колірною розміткою наочно ілюструє здатність навченої моделі ідентифікувати функціональні переходи між фазами рухової активності та моментами виражених емоційних сплесків. Отримані результати об'єктивно доводять, що сформований латентний простір адекватно відображає семантичну структуру даних, дозволяючи алгоритму стабільно фіксувати зміни в поведінці суб'єкта в режимі реального часу.

Верифікація результатів підтверджує високу процесуальну точність вдосконаленого методу MMHPE, яка досягає інтегрального показника 97% при значеннях метрик *precision, recall* та *F1*-міри в межах 0,92–0,97 для всіх категорій. Висока кореляція між передбаченими станами та фактичними даними доводить, що зафіксовані емоційні сплески відповідають стійким кластерам у латентному просторі, а не є випадковими артефактами. Успішний перехід від ідентифікації мікроекспресій за допомогою показника *RLE* до високої статистичної точності класифікації підтверджує здатність системи надійно трансформувати сирі візуальні дані у достовірну звітність, зберігаючи стабільність завдяки використанню *3D-CNN* та механізму адаптивного злиття модальностей навіть у складних умовах відеозйомки.

Фінальним етапом роботи методу MMHPE є трансформація мультимодальних векторів у конкретні категорії емоцій шляхом безперервного обчислення відстані від поточного представлення до еталонних центроїдів латентного простору. Система демонструє високу дискре-

тну здатність та стійку класифікацію, забезпечуючи миттєві переходи між станами без хаотичних коливань завдяки ефективній фільтрації шумів через показник *RLE* та адаптивне злиття ознак. Такий результат підтверджує, що метод успішно синтезує складні візуальні патерни у зрозумілу часову послідовність, гарантуючи високу точність верифікації психоемоційного стану суб'єкта в режимі реального часу.

Проблема розпізнавання емоцій у представників різних вікових категорій полягає в суттєвих анатомічних відмінностях та різній інтенсивності прояву мімічних ознак. Стандартні алгоритми екстракції ознак, що базуються на статистичних моделях дорослих суб'єктів, часто демонструють низьку чутливість до дитячої міміки через менші відстані між антропометричними орієнтами та відсутність виражених шкірних деформацій. Для подолання вікової специфіки метод було вдосконалено шляхом впровадження концепції універсальної посмішки, яка базується на прямому геометричному аналізі відносної експансії губ *Relative Lip Expansion* (RLE) замість використання імовірнісних оцінок готових класифікаторів.

Математичне обґрунтування вдосконаленого методу будується на нормалізації поточних метрик обличчя відносно індивідуального базового стану суб'єкта. На першому етапі визначається базова евклідова відстань D_{base} між ключовими антропометричними точками кутків рота (Landmarks 61 та 291) у стані спокою, обчислюється як медіана значень за перші *n*-кадрів відеопотоку. На основі отриманих значень розраховується показник RLE_t, який і виступає універсальним індикатором посмішки для будь-якої вікової категорії відповідно до формули (3).

$$RLE_t = \frac{D_t - D_{base}}{D_{base}} \quad (3)$$

Використання відносного показника дозволяє нівелювати різницю в лінійних розмірах обличчя дитини та дорослого, оскільки система аналізує не абсолютне розтягнення губ у міліметрах, а відсоткову зміну щодо індивідуальної норми. Такий підхід забезпечує інваріантність до суб'єкта та дозволяє фіксувати мікроекспресії, які ігноруються стандартними моделями через їхню малу амплітуду. В загальній структурі методу значення RLE_t інтегрується як додатковий динамічний дескриптор у вектор ознак обличчя ϕ_f, що гарантує наявність безперервного та достовірного сигналу про емоційний стан незалежно від віку людини в кадрі (рис. 3).

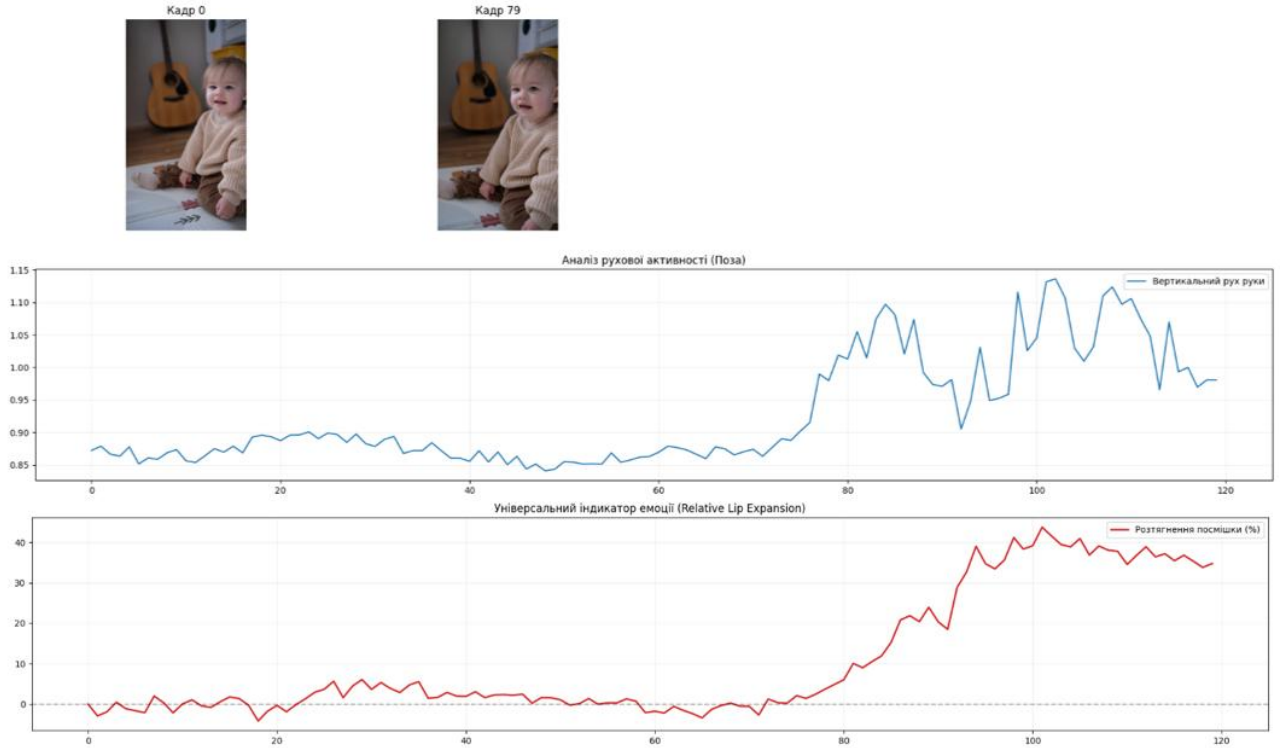

Рис. 3. Ефективність універсального індикатора Relative Lip Expansion (RLE)

На відміну від традиційних моделей, які часто демонструють нульову чутливість до мімічних змін через вікові особливості суб'єктів, використання геометричного показника *RLE* дозволяє достовірно фіксувати динаміку посмішки та її амплітуду. Це забезпечує точну синхронізацію між фізичною активністю та емоційними проявами, що є критично важливим для цілісності мультимодального аналізу в реальному часі.

Результати мультимодального аналізу емоційної та рухової активності підтверджують стійку кореляцію між геометричними змінами обличчя, обчисленими за методом RLE та динамікою жестів суб'єкта. Вдосконалений метод продемонстрував здатність ефективно виділяти чистий емоційний сигнал навіть за умов інтенсивної фізичної активності, що дозволило сформувати достовірні поведінкові профілі з високим рівнем математичного обґрунтування. Комплексна апробація методу *MMHPE* доводить його ефективність на всіх етапах: від первинної екстракції ознак за допомогою *3D-CNN* до фінальної класифікації у латентному просторі.

Висновки та перспективи

У результаті проведеного дослідження обґрунтовано мультимодальну систему розпізнавання емоцій, яка завдяки інтеграції архітектури 3D-CNN та впровадженню універсального індикатора RLE забезпечує стабільну точність ідентифікації емоційних станів на рівні 97%. Використання методів некерованого навчання та стратегії пізнього злиття модальностей дозволило моделі автономно виявляти емоційні патерни в динамічному відеопотоці, успішно нівелюючи вікові особливості міміки та технічні шуми відеозапису. Практична значущість роботи підтверджена конвергенцією часової розмітки та статистичних профілів активності, що створює надійну базу для впровадження алгоритму в інтелектуальні освітні та моніторингові середовища.

Подальший розвиток проєкту передбачає створення спеціалізованого модуля з акумульованими розміченими та навченими даними, що дозволить забезпечити їх довготривале збереження та формування цілісної бази знань для глибокого аналізу психоемоційних станів у реальному часі.

Декларація про штучний інтелект

Під час роботи над даною статтею автором було застосовано наступні інструменти штучного інтелекту: сервіс NotebookLM використовувався для пошуку літератури та створення візуалізації інфографіки (структурна схема, рис. 1); мовна модель Google Gemini — для лінгвістичної корекції тексту та усунення лексичних повторів. Усі матеріали, запропоновані та/або оптимізовані за допомогою ШІ, були критично оцінені, перевірені та відредаговані особисто автором. Основні наукові положення, результати дослідження та висновки є виключно власним інтелектуальним внеском автора.

Конфлікт інтересів

Автор заявляє про відсутність конфлікту інтересів, а саме комерційних, фінансових чи особистих зв'язків, які могли б вплинути на хід дослідження, його результати, висновки, наведені у даній роботі.

Список використаної літератури

1. Chen C., Zhang J., Lakshmikanth S., Fang Y., Shao R., Wetzstein G., Fei-Fei L., & Adeli E. (2026). The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Motion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6200-6211

2. Liu Y., Xu S., Guo J., Wang D., Wang Z., Tan X., & Liu L. (2025). SyncAnimation: A Real-Time End-to-End Framework for Audio-Driven Human Pose and Talking Head Animation. arXiv preprint, pp. 1-11

3. Rivera F. , Rodrigues P., & Fugita O. (2025). Emotion recognition from facial images, body gestures, and skeletal posture keypoints: The BER2024 dataset. Computers in Biology and Medicine, 193, 110350. https://doi.org/10.1016/j.compbiomed.2025.110350

4. Suryawanshi P., & Pansare J. (2025). A Literature Survey on Multi-Modal Emotion Detection System. International Journal of Scientific Research in Science and Engineering (IJSRD), 13(4), pp 91-94.

5. Wu C., Cai Y., Liu Y., Zhu P., Xue Y., Gong Z., Hirschberg J., & Ma B. (2025). Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects. Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 1-17, DOI:10.48550/arXiv.2505.20511

6. Wu Y., Mi Q., & Gao T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics, 10(7), 418, pp. 2-24 https://doi.org/10.3390/biomimetics10070418

7. Yoon S., & Kim B. (2025). Multi-Scale Temporal Fusion Network for Real-Time Multimodal Emotion Recognition in IoT Environments. Sensors, 25(16), https://doi.org/10.3390/s25165066

8. Yu J., Li P., Zhou X., Liu Y., Zheng K., & Wang J. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures. Electronics, 13(18), 3756, pp. 1-22, https://doi.org/10.3390/electronics13183756

9. Yu J., Wang Y., Wang L., Zheng Y., & Xu S. (2025). Interactive Multimodal Framework with Temporal Modeling for Emotion Recognition. 8th Workshop on Affective & Behavior Analysis in-the-wild (ABAW), CVPR

10. Зінченко О. В., Кисіль Т. М. (2025). Згорткові нейронні мережі для аналізу рухомих об'єктів у відеопотоці. Зв'язок, (4), 48–57. https://doi.org/10.31673/2412-9070.2025.042042

11. Зінченко О. В., Звенігородський О. С., Кисіль Т. М. (2022). Згорткові нейронні мережі для вирішення задач комп'ютерного зору. Телекомунікаційні та інформаційні технології, (2), 4–12. DOI: 10.31673/2412-4338.2022.020411

T. Kysil

MULTIMODAL REAL-TIME EMOTION RECOGNITION SYSTEM BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS

The scientific paper substantiates the conceptual framework for developing a multimodal intelligent system for real-time human emotion recognition, leveraging dynamic three-dimensional convolutional neural networks (3D-CNN) and advanced unsupervised machine learning techniques. The primary scientific novelty lies in the formulation of a methodological approach for autonomous spatiotemporal feature extraction, which significantly reduces reliance on pre-labeled datasets while enhancing the system's adaptability to the unique nuances of non-verbal expression. A substantial contribution involves the technical refinement of the MediaPipe framework through the integration of a specialized, modified algorithm for anthropometric marker detection, tailored for capturing the intricate patterns of pediatric facial expressions.

This method for identifying smiles in children is based on a rigorous analysis of the dynamic curvature of nasolabial folds and malar point elevation, ensuring that positive emotional states are correctly identified rather than being misinterpreted as background noise or neutral expressions. The methodological foundation is built upon a dual-channel architecture that conducts parallel analysis of facial micro-expressions and body kinematics, effectively mitigating risks associated with partial facial occlusion or suboptimal camera angles. By employing 3D-CNNs the system processes video data as cohesive spatiotemporal structures, while automated pseudo-labeling within unsupervised latent space clustering enables the system to autonomously structure basic emotional categories.

Experimental validation using a Late Fusion modality integration strategy confirmed the model's robustness against noisy signals and inconsistent lighting. Results demonstrate that the proposed model achieves the high-speed processing required for real-time operation, making it suitable for integration into intelligent educational platforms, pediatric diagnostics, and security systems.

Keywords: emotion recognition, multimodal system, 3D-CNN, unsupervised learning, late fusion, real-time, computer vision, non-verbal behavior, feature clustering, facial expressions.

Надійшла до редакції: 11.02.2026
Прийнята до друку: 21.04.2026
Опубліковано: 27.04.2026