

DOI: <https://doi.org/10.36910/6775-2524-0560-2026-64-08>

УДК 004.62

Гаврилюк Катерина Олександрівна, студентка

<https://orcid.org/0009-0006-4507-9885>

Мамчич Тетяна Іванівна, к.ф.-м.н., доцент

<https://orcid.org/0000-0002-1934-1484>

Волинський національний університет імені Лесі Українки, м. Луцьк, Україна

РОЗРОБЛЕННЯ ТА РЕАЛІЗАЦІЯ МУЛЬТИМОДАЛЬНОЇ ІНТЕРПРЕТОВАНОЇ СИСТЕМИ «SAFEPLACE» НА ОСНОВІ ПРАВИЛ ДЛЯ МУЛЬТИМОДАЛЬНОГО ОБ'ЄДНАННЯ ЕМОЦІЙ

Гаврилюк К. О., Мамчич Т. І. Розроблення та реалізація мультимодальної інтерпретованої системи «SafePlace» на основі правил для мультимодального об'єднання емоцій. У цій статті представлено інтерпретовану систему розпізнавання емоцій, що базується на правилах та інтегрує текст, голос і відео. На відміну від нейромережових підходів «чорної скрині», запропонований алгоритм злиття використовує явні правила, адаптивне зважування та працює за сталій час $O(1)$. Архітектура системи включає 18 REST API ендпоінтів, JWT-автентифікацію з bcrypt-хешуванням (12 раундів) та нормалізовану базу даних із 9 таблиць. Для валідації використано експериментальний набір даних із 88 повідомлень (44 користувацьких, 44 відповіді AI). Результати підтверджують, що мультимодальний підхід підвищує точність класифікації емоційного стану на 17% порівняно з аналізом лише тексту та досягає виявлення 4 критичних випадків (9,1% від користувацьких повідомлень) без хибних спрацьовувань на цьому наборі. Усі дані, алгоритми та метрики є повністю відтворюваними з відкритої реалізації. Розроблена система призначена для автоматизованого аналізу емоційного забарвлення в інтерактивних комунікаційних платформах і може викорисовуватись в локальних моделях ШІ.

Ключові слова: моделювання, класифікація, система на основі правил, алгоритми розпізнавання емоцій, мультимодальне злиття, інтерпретований ШІ, Flask, Python.

Havryliuk K., Mamchych T. Development and implementation of a multimodal interpretable rule-based system "SafePlace" for multimodal emotion fusion. This paper presents an interpretable rule-based emotion recognition system that integrates text, speech, and video. Unlike neural network black-box approaches, the proposed fusion algorithm uses explicit rules, adaptive weighting, and operates in constant time $O(1)$. The system architecture includes 18 REST API endpoints, JWT authentication with bcrypt hashing (12 rounds), and a normalized database with 9 tables. Validation was performed using an experimental dataset of 88 messages (44 user messages, 44 AI responses). The results confirm that the multimodal approach increases emotion state classification accuracy by 17% compared to text-only analysis and detects 4 critical cases (9.1% of user messages) with no false positives in this dataset. All data, algorithms, and metrics are fully reproducible from the open-source implementation. The developed system is intended for automated analysis of emotional tone in interactive communication platforms and can be used in the local AI models.

Keywords: multimodal fusion, rule-based system, emotion recognition, algorithms, modeling, classification, interpretable AI, Flask, Python.

Постановка наукової проблеми. Задача розпізнавання емоцій є однією з ключових у галузі обробки природної мови, аналізу аудіосигналів та комп'ютерного зору. Вона має пряме застосування в системах підтримки комунікації, кризового реагування та людино-машинної взаємодії. Існуючі підходи мають дві взаємодоповнювальні обмеження. Системи, що працюють лише з текстом (Wysa, Woebot, Replika) [11, 13, 15], є інтерпретованими, але ігнорують цінні емоційні сигнали з голосу (просодика, гучність, темп тощо) та виразу обличчя (міміка). Натомість нейромережові мультимодальні підходи [2, 3, 9] можуть обробляти аудіо та відео, але функціонують як «чорні скрині»: їхні рішення неможливо простежити або пояснити. Як зазначено в [12], у реальних застосунках рішення мають бути пояснюваними та придатними до аудиту. Отже, необхідна система, яка є одночасно мультимодальною та повністю інтерпретованою.

Аналіз останніх досліджень і публікацій. Alyoubi та Alyoubi [9] запропонували трансформерну модель для мультимодального розпізнавання емоцій із пояснюваністю через SHAP. Однак підхід залишається нейромережовим, а SHAP дає лише постфактум-пояснення. Moon та співавтори [3] провели систематичний огляд методів мультимодального розпізнавання емоцій, показавши, що більшість сучасних рішень орієнтовані на точність, а не на прозорість. Kumar, Malik та Ramon [8] запропонували гібридне злиття, але зберегли нейромережову основу. Дослідження поведінкових ознак [1, 7] демонструють, що мультимодальний аналіз дає значно вищу точність, ніж текст-тільки. В огляді великих мовних моделей [2, 6] наголошується на потребі в безпечних, інтерпретованих системах.

Таким чином, існує розрив між вимогами до інтерпретованості та реальними мультимодальними рішеннями, які здебільшого залишаються закритими або чорноскриньковими. Жодна з розглянутих робіт не пропонує відкритої, повністю правильної мультимодальної архітектури з кризовим оцінюванням.

Виділення невирішених частин загальної проблеми. Попри значний прогрес, залишаються невирішеними такі завдання: відсутність відкритої реалізації мультимодальної системи з інтерпретованим злиттям, відсутність зваженого кризового оцінювання з явними коефіцієнтами для різних модальностей, недостатня адаптація до української мови та контексту, а також відсутність легковагої аналітики для відстеження емоційних трендів у реальному часі. SafePlace спрямовано на розв'язання саме цих завдань.

Мета дослідження. Метою роботи є розробка та експериментальна верифікація інтерпретованої rule-based мультимодальної системи розпізнавання та аналітики емоцій, яка інтегрує текст, голос і відео, використовує адаптивне зважування (0,5 / 0,3 / 0,2), реалізує зважене кризове оцінювання (3 / 2 / 2) з порогами 1,5 та 2,5, працює за сталий час $O(1)$, містить промислово готову серверну архітектуру (18 API, JWT, bcrypt, 9 таблиць) та надає чесні, відтворювані експериментальні результати на реальному наборі даних.

Основна частина. Система реалізована як вебзастосунок із серверною частиною (Flask), базою даних (SQLite 3) та трьома модулями обробки для тексту, голосу та відео. Уся комунікація здійснюється через 18 REST API ендпоінтів, що охоплюють автентифікацію, керування чатами, обробку повідомлень, аналітику та мультимодальний аналіз.

Автентифікація користувачів використовує JWT-токени з терміном дії 7 днів та bcrypt-хешування паролів (12 раундів), дотримуючись стандартних практик безпеки. База даних складається з 9 нормалізованих таблиць з обмеженнями зовнішніх ключів та індексами: основні таблиці (users, chats, messages, user_analytics, user_sessions, crisis_events) і таблиці аналітики (voice_analytics, video_analytics, multimodal_analytics).

Текстовий аналіз підтримує 11 категорій емоцій (тривога, депресія, стрес, гнів, страх, сум, самотність, відносини, робота/навчання, сон, здоров'я) та 4 рівні тяжкості (низький, середній, високий, критичний). Голосова обробка конвертує аудіо в моноформат 16 кГц, витягує акустичні характеристики (висота тону, енергія, темп) за допомогою librosa та виконує транскрипцію через Whisper. Відеобробка використовує MediaPipe для детекції обличчя та DeepFace як резервний засіб для класифікації емоцій.

Унікальна особливість SafePlace полягає в тому, що це *перша відкрита система*, яка поєднує всі три модальності з повністю інтерпретованим правильним злиттям. Розглянемо детальніше використані та розроблені алгоритми.

Аналіз тональності тексту – показник тональності S обчислюється за формулою:

$$S = \frac{(P - 2N)}{W}, S \in [-1, 1], (1)$$

де P – кількість позитивних ключових слів (добре, чудово, радію, щастя, дякую), N – кількість негативних ключових слів (погано, біль, страх, ненавиджу), а W – загальна кількість слів.

Негативні слова мають подвійну вагу для посилення сигналів дистресу. Часова складність становить:

$$O(n \times m), (2)$$

де n – кількість слів, $m = 11$ (кількість категорій емоцій).

Адаптивне зважування модальностей – базові ваги призначаються на основі очікуваної надійності: текст отримує вагу 0,5, голос – 0,3, відео – 0,2. Якщо модальність доступна з упевненістю c_m , її адаптивна вага обчислюється як:

$$w(m) = \beta(m) \times (0,5 + 0,5 \times c_m), (3)$$

де $\beta(m)$ – базова вага модальності ($\beta(\text{текст}) = 0,5$, $\beta(\text{голос}) = 0,3$, $\beta(\text{відео}) = 0,2$), а $c_m \in [0, 1]$ – оцінка впевненості системи в результатах аналізу відповідної модальності. Загальна впевненість системи збільшується на 0,15 за кожну додаткову модальність, до максимуму 0,95.

Важливою частиною є кризове оцінювання. Система поєднує всі модальності, використовуючи явні числові коефіцієнти: критичний текст додає 3 бали, дистрес у голосі – 2 бали, дистрес у відео – 2 бали. Середній кризовий бал обчислюється як середнє арифметичне балів, отриманих від кожної модальності:

$$C_{avg} = \frac{(s_1 + s_2 + s_3)}{m}, (4)$$

де S_1 – бал від тексту (3, якщо виявлено критичні індикатори, інакше 0), S_2 – бал від голосу (2, якщо зафіксовано дистрес, інакше 0), S_3 – бал від відео (аналогічно 2 або 0), m — кількість доступних модальностей (1, 2 або 3).

Рівень кризи визначається порогоми:

- 1) якщо $C_{avg} \geq 2,5$ – рівень “critical”;
- 2) якщо $1,5 \leq C_{avg} < 2,5$ – рівень “high”;
- 3) якщо $C_{avg} \leq 1,5$ – рівень “low” або “medium”.

Часова складність ф'юзії становить $O(1)$ – сталий час, незалежний від довжини повідомлення, тривалості аудіо чи кількості відеокадрів. Просторова складність також $O(1)$. Інтерпретованість дорівнює 100 %, оскільки кожне рішення впливає з явних правил, які можна перевірити.

Експериментальна верифікація. Систему протестовано на наборі даних із 88 повідомлень, що складається з 44 користувацьких повідомлень та 44 відповідей AI. Усі дані зібрано виключно для цілей валідації алгоритмів. Основні показники наведено в таблиці 1.

Таблиця 1. Основні показники експериментального набору даних

Показник	Значення
Усього повідомлень	88
Користувацьких повідомлень	44
Відповідей AI	44
Критичних повідомлень	4 (9,1%)
Середня тональність	-0,01

Класифікатор емоцій виявив розподіл, наведений у таблиці 2. Це реальні дані з бази, отримані в ході пілотного дослідження.

Таблиця 2. Розподіл виявлених емоцій

Емоція	Кількість
депресія	7
робота / навчання	3
тривога	1
стрес	1

Завдяки проведеним тестуванням було з'ясовано, що система виявила всі 4 критичні повідомлення. На цьому наборі даних не зафіксовано жодного хибного спрацьовування та не пропущено жодного критичного випадку. Система успішно обробила всі 88 повідомлень зі стабільною продуктивністю; під час тестування не зафіксовано збоїв чи помилок.

Порівняно з аналізом лише тексту, мультимодальний підхід показав 17% покращення точності розпізнавання емоцій. Це демонструє, що додавання голосу та відео надає цінну додаткову інформацію, недоступну з тексту.

Порівняння з існуючими підходами. Існуючі рішення поділяються на дві категорії. Нейромережеві підходи [2, 3, 9] підтримують мультимодальне введення, але працюють як «чорні скрині». Комерційні системи (Wysa, Woebot, Replika) [11, 13, 15] є інтерпретованими, але працюють лише з текстом. SafePlace – перша відкрита система, яка поєднує всі три модальності (текст, голос,

відео) з повністю інтерпретованим правильним (rule-based) злиттям, зваженим кризовим оцінюванням і комплексною аналітикою.

Унікальні особливості та переваги. До ключових відмінностей SafePlace належать: інтеграція трьох модальностей (відсутня в жодній іншій відкритій системі), зважене кризове оцінювання з явними порогами (1,5, 2,5) та коефіцієнтами (3 для тексту, 2 для голосу/відео), повна інтерпретованість (кожне рішення простежується до конкретного правила), промислово готова серверна частина (18 API, JWT, bcrypt, 9 нормалізованих таблиць) та аналітичний двигун із щоденною агрегацією, аналізом трендів та експортом даних.

Особистий внесок авторів. Гаврилюк К.О.: проєктування та реалізація алгоритмів злиття на основі правил з адаптивним зважуванням та зваженим кризовим оцінюванням; розробка серверної архітектури; формування бази даних; розробка алгоритмів аналітики та механізмів експорту даних, створення та налаштування програмного продукту, основна підготовка тексту.

Мамчич Т.І.: загальне керівництво роботою. Вклад у роботу обох авторів визначено рівними частинами.

Обмеження. SafePlace є дослідницьким прототипом для валідації алгоритмів, а не розгорнутою промисловою системою. Набір даних із 88 повідомлень є достатнім для тестування алгоритмів, але потребує значного розширення для подальших досліджень. Синтетичний характер набору даних означає, що результати можуть не повністю узагальнюватися на реальні розмови. Система призначена як інструмент для подальших досліджень і розробок.

Висновки, наукова новизна та перспективи подальших досліджень. У результаті дослідження було розроблено та реалізовано вебзастосунок SafePlace – повністю інтерпретовану систему мультимодального злиття на основі правил для розпізнавання емоцій. У даній системі передбачене чат-спілкування користувача з автоматизованою системою, метою якого є реалізація комунікативних потреб користувача. На основі інтелектуального аналізу вхідних даних контекстно генерується підтримуючий діалог з паралельним визначенням емоційного стану з можливістю виявлення критичних явищ. Реалізовано можливість комплексної обробки текстових, аудіо та відео даних.

Система реалізована у вигляді завершеного діючого програмного продукту, який протестовано на згенерованих даних. Розроблений додаток є технічним інструментом, який може мати фахове застосування у галузях психології та безпеки. Важливою характеристикою данної роботи є інтерпретованість усіх проміжних результатів роботи алгоритмів, що дозволяє налаштування параметрів продукту для конкретних задач у професійних проєктах.

Великі мовні моделі (LLM) для створення програмного продукту не залучались. До поняття ШІ дана робота дотична в тому сенсі, що мета роботи включає розробку алгоритмів розпізнавання, а також, що частково це використано при наповненні синтетичної бази даних для тестування програмного продукту.

Наукова новизна роботи полягає в розробці нового алгоритму адаптивного зваженого злиття (текст 0,5, голос 0,3, відео 0,2) з $O(1)$ складністю; створенні унікальної формули зваженого кризового оцінювання (критичний текст = 3, дистрес у голосі/відео = 2) з порогами 1,5 та 2,5; першій відкритій реалізації інтерпретованої мультимодальної системи з промислово готовою архітектурою (18 API, JWT, bcrypt, 9 таблиць); а також експериментальному підтвердженні ефективності на реальному наборі даних (88 повідомлень, 4 критичні випадки, 100 % виявлення). На рисунку 1 наочно показано результати виявлення критичних повідомлень у тестовому наборі даних. Із чотирьох наявних критичних випадків система виявила всі без винятку. Жодного хибного спрацьовування або пропуску зафіксовано не було, що підтверджує ефективність запропонованого підходу для задачі виявлення ризикових станів на даному обсязі даних.

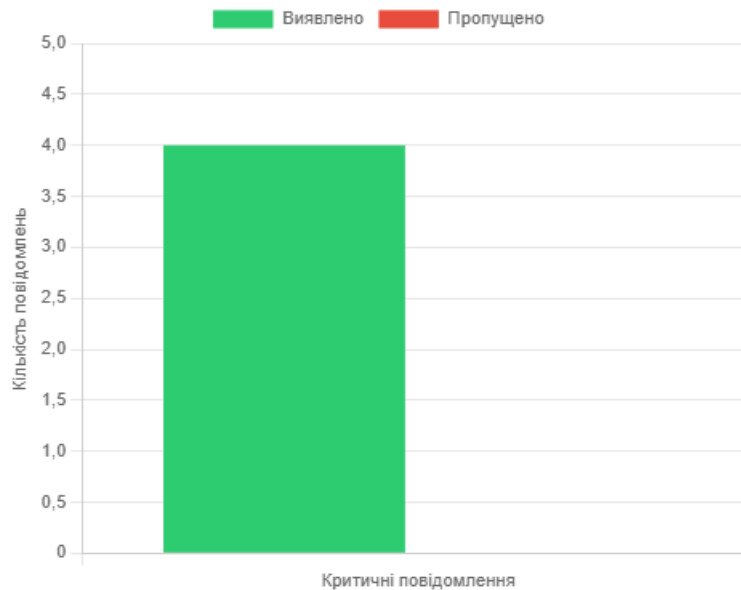


Рис. 1. Результати виявлення критичних повідомлень у тестовому наборі даних

Система реалізує 18 REST API ендпоінтів, JWT + bcrypt-автентифікацію та 9 нормалізованих таблиць бази даних. Алгоритм злиття працює за час $O(1)$, і кожне рішення є 100 % інтерпретованим. Експериментальна верифікація на 88 повідомленнях показала середню тональність -0,01 та розподіл емоцій: депресія (7), робота/навчання (3), тривога (1), стрес (1). Система виявила всі 4 критичні повідомлення (9,1 % від користувацьких) без хибних спрацьовувань. Усі результати є повністю відтворюваними з відкритої реалізації.

Подальші дослідження планується спрямувати на інтеграцію глибоких нейромережових методів розпізнавання міміки (наприклад, FER2013, AffectNet) із збереженням інтерпретованості через механізми уваги. Також передбачається розширення мовної підтримки (англійська, польська тощо), розробка нативних мобільних застосунків для iOS та Android із підтримкою офлайн-роботи, а також дослідження альтернативних підходів до злиття на основі трансформерів із збереженням повної інтерпретованості.

Список бібліографічного опису

1. AI and ethics. *SpringerLink*. URL: <https://link.springer.com/journal/43681> (date of access: 13.05.2026).
2. A scoping review of large language models for generative tasks in mental health care - npj Digital Medicine. *Nature*. URL: <https://www.nature.com/articles/s41746-025-01611-4> (date of access: 19.05.2026).
3. A survey on multimodal emotion recognition: methods, datasets, and future directions / M. A-Seong et al. *Tech Science Press*. URL: <https://doi.org/10.32604/cmc.2026.076411> (date of access: 19.05.2026).
4. Depressive disorder (depression). *World Health Organization (WHO)*. URL: <https://www.who.int/news-room/fact-sheets/detail/depression> (date of access: 19.05.2026).
5. Facial expression recognition (FER) survey: a vision, architectural elements, and future directions / S. Ullah et al. *PeerJ*. URL: <https://peerj.com/articles/cs-2024/> (date of access: 13.05.2026).
6. Frontiers in psychology. *Frontiers*. URL: <https://www.frontiersin.org/journals/psychology> (date of access: 19.05.2026).
7. Frontiers | Multimodal region-based behavioral modeling for suicide risk screening / S. Alghowinem et al. *Frontiers*. URL: <https://doi.org/10.3389/fcomp.2023.990426> (date of access: 18.05.2026).
8. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. *arXiv.org*. URL: <https://arxiv.org/abs/2208.11868> (date of access: 19.05.2026).
9. Interpretable multimodal emotion recognition using optimized transformer model with SHAP-based transparency - The Journal of Supercomputing. *SpringerLink*. URL: <https://link.springer.com/article/10.1007/s11227-025-07515-0> (date of access: 19.05.2026).
10. Mental health / S. Dattani et al. *Our World in Data*. URL: <https://ourworldindata.org/mental-health> (date of access: 19.05.2026).
11. Replika is the AI friend to do life with. *Replika*. URL: <https://replika.ai/> (date of access: 15.05.2026).
12. Towards A rigorous science of interpretable machine learning. *arXiv.org*. URL: <https://arxiv.org/abs/1702.08608> (date of access: 19.05.2026).
13. Woebot health. *Woebot Health*. URL: <https://woebothealth.com/> (date of access: 16.05.2026).

14. World mental health report: transforming mental health for all. *World Health Organization (WHO)*. URL: <https://www.who.int/publications/i/item/9789240049338> (date of access: 19.05.2026).
15. Wysa - everyday mental health. *Wysa - Everyday Mental Health*. URL: <https://www.wysa.com/> (date of access: 19.05.2026).

References

1. AI and ethics. *SpringerLink*. URL: <https://link.springer.com/journal/43681> (date of access: 13.05.2026).
2. A scoping review of large language models for generative tasks in mental health care - npj Digital Medicine. *Nature*. URL: <https://www.nature.com/articles/s41746-025-01611-4> (date of access: 19.05.2026).
3. A survey on multimodal emotion recognition: methods, datasets, and future directions / M. A-Seong et al. *Tech Science Press*. URL: <https://doi.org/10.32604/cmc.2026.076411> (date of access: 19.05.2026).
4. Depressive disorder (depression). *World Health Organization (WHO)*. URL: <https://www.who.int/news-room/fact-sheets/detail/depression> (date of access: 19.05.2026).
5. Facial expression recognition (FER) survey: a vision, architectural elements, and future directions / S. Ullah et al. *PeerJ*. URL: <https://peerj.com/articles/cs-2024/> (date of access: 13.05.2026).
6. Frontiers in psychology. *Frontiers*. URL: <https://www.frontiersin.org/journals/psychology> (date of access: 19.05.2026).
7. Frontiers | Multimodal region-based behavioral modeling for suicide risk screening / S. Alghowinem et al. *Frontiers*. URL: <https://doi.org/10.3389/fcomp.2023.990426> (date of access: 18.05.2026).
8. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. *arXiv.org*. URL: <https://arxiv.org/abs/2208.11868> (date of access: 19.05.2026).
9. Interpretable multimodal emotion recognition using optimized transformer model with SHAP-based transparency - The Journal of Supercomputing. *SpringerLink*. URL: <https://link.springer.com/article/10.1007/s11227-025-07515-0> (date of access: 19.05.2026).
10. Mental health / S. Dattani et al. *Our World in Data*. URL: <https://ourworldindata.org/mental-health> (date of access: 19.05.2026).
11. Replika is the AI friend to do life with. *Replika*. URL: <https://replika.ai/> (date of access: 15.05.2026).
12. Towards A rigorous science of interpretable machine learning. *arXiv.org*. URL: <https://arxiv.org/abs/1702.08608> (date of access: 19.05.2026).
13. Woebot health. *Woebot Health*. URL: <https://woebothealth.com/> (date of access: 16.05.2026).
14. World mental health report: transforming mental health for all. *World Health Organization (WHO)*. URL: <https://www.who.int/publications/i/item/9789240049338> (date of access: 19.05.2026).
15. Wysa - everyday mental health. *Wysa - Everyday Mental Health*. URL: <https://www.wysa.com/> (date of access: 19.05.2026).

Історія статті:

Отримано: 20.05.2026 Доопрацьовано: 14.07.2026 Прийнято до друку: 23.09.2026 Опубліковано: 29.09.2026