Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСИФІКАЦІЇ ЕМОЦІЙ В МЕДІАДАНИХ ДЛЯ АНАЛІЗУ ЕМОЦІЙНОГО СТАНУ КОРИСТУВАЧІВ
(тема)

Виконав:
здобувач 2 року навчання,
групи ІНФм-24-1
Цісаренко О. І.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Науковий керівник доц. Руденко Д. О.
(посада, прізвище, ініціали)

Допускається до захисту

Завідувач кафедри інформатики _____ Кобилін О. А.
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту

Кафедра Інформатики

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Цісаренку Олександру Ігоровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів класифікації емоцій в медіаданих для аналізу емоційного стану користувачів
затверджена наказом університету від 14 листопада 2025 року № 1045Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 18 грудня 2025 р.
3. Вихідні дані до роботи технології обробки текстових та мультимедійних даних, методи машинного навчання для класифікації емоцій у тексті, зображеннях та аудіо, літературні джерела щодо застосування алгоритмів ML у sentiment analysis та emotion recognition, інструменти для попередньої обробки текстових даних (токенізація, стемінг, лемматизація), програмні засоби для реалізації прогнозних моделей Python, opencv_python, pillow, numpy, tensorflow, keras, scikit-learn, методи оцінки якості класифікації та метрики точності, набори даних для класифікації емоцій з платформ Kaggle, Hugging Face Datasets, SemEval, допоміжні діаграми, графіки та статистичні матеріали, результати навчання та тестування моделей, синтетично згенеровані дані для моделювання різних емоційних станів користувачів.

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз методів роботи з медіаданими та машинного навчання для класифікації емоцій у тексті, зображеннях та аудіо.
2. Дослідження лексичних, синтаксичних, семантичних та контекстуальних ознак у медіаданих та визначення ключових факторів, що впливають на точність класифікації емоцій.
3. Формування вибірки та проведення попередньої обробки даних.
4. Розробка прототипу програмного забезпечення для інтерактивного аналізу емоційного користувачів на основі метаданих.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) актуальність класифікації емоцій у медіаданих, об'єкт і мета дослідження, постановка задачі; блок-схема обробки медіаданих; діаграми попередньої обробки та нормалізації текстових та мультимедійних даних; ілюстрації формування вибірки й побудови ознак; приклад синтетичних емоційних даних; схема архітектури прототипу застосунку; інтерфейс головної сторінки та графік розподілу емоцій; сторінка статистики з аналітичними графіками; матриці плутанини та візуалізація розподілу емоційних станів; порівняльні графіки ефективності моделей; підсумкові діаграми точності та перспективи розвитку системи.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	29.09.2025	
2	Аналіз завдання, підбір літератури	30.09.25-07.10.25	
3	Аналіз літератури з досліджуваної проблеми	08.10.25-14.10.25	
4	Особливості методів класифікації емоцій у метаданих	15.10.25-27.10.25	
5	Дослідження методів класифікації емоцій у метаданих	15.10.25-27.10.25	
6	Програмна реалізація	28.10.25-05.11.25	
7	Обґрунтування отриманих результатів	06.11.25-11.11.25	
8	Оформлення пояснювальної записки	12.11.25-24.11.25	
9	Перевірка на нормоконтроль	14.12.25	
10	Перевірка на плагіат	15.12.25	
11	Рецензування	16.12.25	
12	Підготовка презентації та доповіді	20.12.25	
13	Занесення роботи в електронний архів	22.12.25	
14	Попередній захист кваліфікаційної роботи	22.12.25	

Дата видачі завдання 29 вересня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Руденко Д. О
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи: 80 с., 1 табл., 15 рис., 32 джерела.

КЛАСИФІКАЦІЯ ЕМОЦІЙ, КОМП'ЮТЕРНИЙ ЗІР, РОЗПІЗНАВАННЯ ЄМОЦІЙ, МАШИНЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ ЗОБРАЖЕНЬ, ЕМОЦІЙНИЙ СТАН, ЄМОЦІЇ, ГЛИБОКЕ НАВЧАННЯ, ГЛИБОКІ ЗГОРТКОВІ МЕРЕЖІ, ВІЗУАЛІЗАЦІЯ ДАНИХ, CNN, LSTM, МЕДІАДАНІ, DEAP, FER-2013, PYTHON, TENSORFLOW, NUMPY, KERAS.

Об'єктом дослідження є медіадані, що містять ознаки емоційних проявів (зображення облич, аудіозаписи голосу, текстові повідомлення).

Предметом дослідження є методи автоматичної класифікації емоцій для аналізу емоційного стану користувачів за такими даними.

Метою дослідження є порівняння сучасних моделей розпізнавання емоцій шляхом розробки й апробації програмного прототипу, який аналізує медіадані користувачів та класифікує емоційний стан із заданою точністю.

Використано методи аналізу медіаданих з використанням глибоких згорткових нейронних мереж (CNN), рекурентних нейронних мереж (LSTM), CNN-LSTM, CNN-Transformer, а також класичні моделі глибокого навчання, навчання з переносом та донавчання на відкритих датасетах (FER-2013, DEAP).

Взаємозв'язок з іншими роботами проявляється у соціальних мережах, системах рекомендацій, дистанційне навчання, телемедичні сервіси тощо.

Рекомендації щодо використання результатів роботи є є програмний комплекс придатний для класифікації емоцій у соціальних платформах, чат- ботах для психологічної підтримки, дистанційних навчальних системах, телемедичних і діагностичних додатках. Отримані алгоритми можна використовувати для поліпшення якості розпізнавання емоцій у сервісах для моніторингу емоційного здоров'я, підвищення ефективності комунікації між людиною та цифровими системами.

ABSTRACT

Explanatory note to the qualification work: 80 pages, 6 tables, 14 figures, 32 sources.

EMOTION CLASSIFICATION, COMPUTER VISION, EMOTION RECOGNITION, MACHINE LEARNING, IMAGE CLASSIFICATION, EMOTIONAL STATE, EMOTIONS, DEEP LEARNING, DEEP CONVOLUTIONAL NETWORKS, CNN, LSTM, MEDIA DATA, DEAP, FER-2013, PYTHON, TENSORFLOW, NUMPY, KERAS.

The object of the study is media data containing signs of emotional manifestations (images of faces, audio recordings of voices, text messages).

The subject of the study is methods of automatic emotion classification for analysing the emotional state of users based on such data.

The aim of the study is to compare modern emotion recognition models by developing and testing a software prototype that analyses user media data and classifies emotional states with a given accuracy.

Methods of media data analysis using deep convolutional neural networks (CNN), recurrent neural networks (LSTM), CNN-LSTM, CNN-Transformer, as well as classical deep learning models, transfer learning and fine-tuning on open datasets (FER-2013, DEAP) are used.

Interconnection with other works is evident in social networks, recommendation systems, distance learning, telemedicine services, etc.

Recommendations for using the results of the work are a software complex suitable for classifying emotions in social platforms, chatbots for psychological support, distance learning systems, telemedicine and diagnostic applications. The algorithms obtained can be used to improve the quality of emotion recognition in services for monitoring emotional health and to increase the effectiveness of communication between humans and digital systems.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ШІ – штучний інтелект

ML – Machine Learning (машинне навчання)

API – Application Programming Interface (програмний інтерфейс застосунку)

BERT – Bidirectional Encoder Representations from Transformers

CNN – Convolutional Neural Network (згорткова нейронна мережа)

Emotion Recognition (розпізнавання емоцій) – процес автоматичного виявлення та класифікації емоційних станів у тексті, зображеннях або аудіо

Lemmatization (лематизація) – приведення слів до їхньої нормальної (словникової) форми з урахуванням морфологічного аналізу

LSTM – Long Short-Term Memory (довга короткочасна пам’ять)

NLP – Natural Language Processing (обробка природної мови)

RNN – Recurrent Neural Network (рекурентна нейронна мережа)

Stemming (стемінг) – приведення слів до їхньої кореневої форми шляхом відсікання афіксів

Sentiment Analysis (аналіз тональності) – метод обробки природної мови для визначення емоційного забарвлення тексту (позитивне, негативне, нейтральне)

Tokenization (токенізація) – розбиття тексту на окремі лексичні одиниці (токени) для подальшої обробки

ВСТУП

У сфері комп’ютерної графіки існують три основні напрями: створення візуалізацій, обробка растрових і векторних зображень, а також розпізнавання образів. Візуалізація передбачає побудову зображень на основі заданих моделей, що може включати графіки, схеми чи імітацію тривимірних світів, наприклад у відеоіграх або архітектурному проєктуванні. Завдання розпізнавання образів полягає в отриманні семантичного опису об’єктів, зображених на зображеннях. Це може бути як виділення конкретних елементів зображення, так і класифікація зображення загалом. По суті, розпізнавання образів є оберненим процесом до візуалізації. Провідні області застосування розпізнавання включають системи розпізнавання тексту та створення 3D-моделей людини на основі фотознімків.

Основне завдання розпізнавання образів полягає в одержанні семантичного опису зображених об’єктів. Мета розпізнавання може бути різна: як виділення окремих елементів на зображенні, так і класифікація зображення в цілому. У якомусь сенсі завдання розпізнавання є зворотним стосовно завдання візуалізації. Областями застосування можуть бути системи розпізнавання текстів, створення тривимірних моделей людини по фотографіях.

Актуальність роботи полягає у стрімкому зростанні значущості автоматичного розпізнавання емоцій у медіаданих, що є ключовим чинником розвитку сучасних технологій людино-комп’ютерної взаємодії, систем психологічного моніторингу, дистанційного навчання, телемедицини та соціальних платформ. Розвиток глибокого навчання та доступність великих мультимодальних датасетів відкривають нові можливості для точного аналізу емоційного стану користувачів на основі різних типів медіа – зображень, аудіо та тексту. Актуальність також зумовлена зростанням впливу онлайн-комунікацій на повсякденне життя та необхідністю підвищення якості персоналізованих сервісів, які враховують емоційний контекст. Потреба у підвищенні ефективності та точності систем розпізнавання емоцій особливо важлива в умовах поширення дистанційної роботи, онлайн-навчання та телемедичних

консультацій, де емоційний стан користувача критично впливає на прийняття рішень та якість взаємодії. Ця проблема є актуальною як з наукової, так і з практичної точки зору, оскільки вплив емоцій на поведінку людини потребує системного вивчення й інтеграції в штучний інтелект.

Огляд сучасного стану проблеми показує, що останніми роками значний прогрес досягнуто завдяки використанню глибоких нейронних мереж (CNN, LSTM, трансформерів) та attention-механізмів для багатомодальної класифікації емоційних станів. Існують численні публічно доступні датасети, які включають різні види медіаданих, зокрема FER-2013 (зображення емоцій облич), DEAP (фізіологічні сигнали), EmoSet тощо. Водночас виклики пов’язані зі складністю інтеграції мультимодальних даних, високими обчислювальними витратами та потребою розробки моделей із високою інтерпретованістю. Значну увагу приділяють питанням узагальнення моделей на нові типи даних та їх адаптації до різних контекстів користувачів.

Наукова задача полягає у створенні та оптимізації гібридних моделей глибокого навчання, які ефективно поєднують різні типи нейронних мереж і механізмів уваги для всебічного розпізнавання емоцій у медіаданих. Це потребує розробки методів попередньої обробки, фільтрації та векторизації мультимодальних даних, вибору оптимальних архітектур моделей, а також проведення емпіричного порівняння їхньої точності та стійкості на різних датасетах. Одним із ключових завдань є забезпечення балансу між точністю моделі, її швидкодією та здатністю до інтеграції у реальні інформаційні системи.

1 АНАЛІЗ ОСНОВНИХ МЕТОДІВ КЛАСИФІКАЦІЇ ЕМОЦІЙ В МЕДІАДАНИХ

1.1 Метод класифікації на основі згорткових нейронних мереж (CNN)

Класифікація емоцій у медіаданих є важливою задачею сучасного машинного навчання, що знаходить застосування в аналізі поведінки користувачів, системах рекомендацій, медичній діагностиці та багатьох інших галузях. Одним із найпоширеніших підходів до розв'язання цієї задачі є використання згорткових нейронних мереж (Convolutional Neural Networks, CNN), які демонструють високу ефективність при обробці візуальних та аудіо даних.

Перевагами CNN для класифікації емоцій є автоматичне виявлення ознак без необхідності ручного проєктування, інваріантність до невеликих зсувів та деформацій, а також можливість навчання наскрізних моделей. Однак CNN вимагають великих обсягів даних для навчання та значних обчислювальних ресурсів.

Згорткові нейронні мережі базуються на операції згортки, яка дозволяє виявляти локальні ознаки в вхідних даних. Для класифікації емоцій на зображеннях обличчя або в аудіозаписах CNN використовує ієрархічну структуру шарів, де кожен наступний шар виявляє все більш абстрактні ознаки. Архітектура типової CNN для класифікації емоцій складається з декількох згорткових шарів, шарів пулінгу та повнозв'язних шарів.

Математично операція згортки для двовимірного сигналу (зображення) описується наступним чином:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) \cdot K(i-m, j-n), \quad (1.1)$$

де I – вхідне зображення;

K – ядро згортки (фільтр);

$S(i,j)$ – значення карти ознак у позиції (i,j);

m,n – індекси, що визначають розмір ядра згортки.

Після кожного згорткового шару зазвичай застосовується функція активації, найчастіше ReLU (Rectified Linear Unit), яка вводить нелінійність у модель:

$$f(x) = max(0,x), \tag{1.2}$$

де x – вхідне значення нейрона;

$f(x)$ – вихідне значення після застосування функції активації.

Шар пулінгу використовується для зменшення просторової розмірності карт ознак, що знижує обчислювальну складність та ризик перенавчання. Найпоширенішим є макс-пулінг, який обирає максимальне значення з кожного регіону:

$$P(i,j) = \max_{(m,n)\in R_{ij}} S\,(m,n), \tag{1.3}$$

де $P(i,j)$ – значення після операції пулінгу;

R_{ij} – регіон пулінгу навколо позиції (i,j);

$S(m,n)$ – значення карти ознак у позиції (m,n).

Для класифікації емоцій на зображеннях облич часто використовується датасет FER-2013 (Facial Expression Recognition), який містить 35887 зображень розміром 48×48 пікселів, розподілених на сім категорій емоцій: гнів, огида, страх, радість, сум, здивування та нейтральний стан [1]. Архітектура CNN для цієї задачі може включати 3-5 згорткових блоків, кожен з яких складається зі згорткового шару, функції активації та шару пулінгу.

Вихідний шар мережі використовує функцію softmax для отримання ймовірностей належності до кожного класу емоцій:

$$P\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad (1.4)$$

де z_j – логіт для класу j;

K – загальна кількість класів емоцій;

$\sigma(z)_j$ – ймовірність належності до класу j;

Навчання мережі відбувається шляхом мінімізації функції втрат, зазвичай категоріальної крос-енентропії:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\widehat{y_{ij}}), \quad (1.5)$$

де N – кількість зразків у навчальній вибірці;

y_{ij} – істинна мітка класу (1, якщо зразок i належить класу j, інакше 0);

$\widehat{y_{ij}}$ – передбачена ймовірність належності зразка i до класу j.

Оптимізація параметрів мережі здійснюється за допомогою алгоритму зворотного поширення помилки та градієнтного спуску. Сучасні модифікації, такі як Adam (Adaptive Moment Estimation), показують кращу збіжність:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\widehat{v_t}}+\epsilon}\widehat{m_t}, \quad (1.6)$$

де θ_t – параметри моделі на кроці t;

α – швидкість навчання;

$\hat{m}_t$ – оцінка першого моменту градієнта з корекцією зміщення;

$\hat{v}_t$ – оцінка другого моменту градієнта з корекцією зміщення;

ϵ – мала константа для числової стабільності (зазвичай 10^{-8}).

Важливим аспектом навчання CNN для класифікації емоцій є попередня обробка даних. Зображення облич повинні бути нормалізовані за розміром, зазвичай до квадратного формату 48×48, 64×64 або 224×224 пікселів, залежно від обраної архітектури. Нормалізація значень пікселів до діапазону [0, 1] або

[- 1, 1] прискорює збіжність навчання та покращує стабільність моделі.

Математично нормалізація пікселів виконується за формулою:

$$x_{norm} = \frac{x-\mu}{\sigma}, \tag{1.7}$$

де x – вихідне значення пікселя;

x_{norm} – нормалізоване значення пікселя;

μ – середнє значення по всьому датасету (зазвичай 127.5 для зображень у діапазоні [0, 255]);

σ – стандартне відхилення (зазвичай 127.5 для масштабування до [-1, 1]).

Аугментація даних є критично важливою технікою для покращення узагальнюючої здатності моделі та запобігання перенавчанню. Для зображень облич застосовуються такі трансформації, як горизонтальне відображення, невеликі повороти (до ±15°), зсуви, масштабування та зміна яскравості.

Ймовірність застосування аугментації під час навчання можна описати як:

$$I_{aug} = T_{\theta}(I_{orig}), \tag{1.8}$$

де I_{orig} – оригінальне зображення;

I_{aug} – аугментоване зображення;

T_{θ} – випадкова трансформація з параметрами θ (кут повороту, зсув, масштаб тощо).

Архітектура CNN для класифікації емоцій зазвичай складається з кількох згорткових блоків. Кожен блок включає згортковий шар, функцію активації, нормалізацію батчу та шар пулінгу. Нормалізація батчу стабілізує навчання та дозволяє використовувати вищу швидкість навчання [14]:

$$\widehat{x_i} = \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\epsilon}}, \tag{1.9}$$

$$y_i = \gamma \widehat{x_i} + \beta, \quad (1.10)$$

де x_i – вхідне значення для і-го елемента в батчі;

μ_B – середнє значення по батчу;

σ_B^2 – дисперсія по батчу;

ϵ – мала константа для числової стабільності;

$\widehat{x_i}$ – нормалізоване значення;

γ, β – навчальні параметри масштабування та зсуву;

y_i – вихідне значення після нормалізації батчу.

Для оцінки якості класифікації емоцій використовуються різні метрики. Окрім загальної точності, важливими є точність (precision), повнота (recall) та F1-міра для кожного класу емоцій:

$$Precision = \frac{TP}{TP+FP}, \quad (1.11)$$

$$Recall = \frac{TP}{TP+FN}, \quad (1.12)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}, \quad (1.13)$$

де TP – кількість істинно позитивних передбачень;

FP – кількість хибно позитивних передбачень;

FN – кількість хибно негативних передбачень.

Експериментальні дослідження показують, що CNN досягають точності 60-70% на датасеті FER-2013, що є складним через низьку роздільну здатність зображень та значну варіативність умов зйомки. На більш якісних датасетах, таких як CK+ (Extended Cohn-Kanade) або AffectNet, точність може досягати 85- 95% [20].

Згорткові нейронні мережі є фундаментальним та високоефективним методом для класифікації емоцій на зображеннях облич, що забезпечує

автоматичне виявлення ієрархічних візуальних ознак без необхідності ручного проєктування дескрипторів. Аналіз методу демонструє його здатність виявляти складні патерни емоційних виразів через послідовне застосування операцій згортки, активації та пулінгу, що дозволяє будувати багаторівневі репрезентації від простих країв та текстур до складних семантичних ознак емоцій.

Ключовою перевагою CNN є здатність до автоматичного навчання ієрархії ознак. Нижні згорткові шари виявляють базові візуальні елементи, такі як краї, кути та текстури, які є загальними для різних типів зображень. Середні шари комбінують ці базові елементи у більш складні структури, такі як частини обличчя (очі, брови, рот), а верхні шари формують високорівневі репрезентації, специфічні для кожної емоції. Така ієрархічна структура дозволяє моделі ефективно узагальнювати знання та розпізнавати емоції навіть при значній варіативності умов зйомки, освітлення та індивідуальних особливостей облич.

Таким чином, згорткові нейронні мережі є потужним інструментом для класифікації емоцій на зображеннях облич, що забезпечує автоматичне виявлення ієрархічних візуальних ознак та високу точність розпізнавання при достатньому обсязі навчальних даних [23, 24].

1.2 Метод трансферного навчання з попередньо навченими моделями

Трансферне навчання (transfer learning) є потужним підходом у машинному навчанні, що дозволяє використовувати знання, отримані при розв’язанні однієї задачі, для покращення продуктивності на іншій пов’язаній задачі. У контексті класифікації емоцій цей метод передбачає використання моделей, що попередньо навчені на великих датасетах зображень, таких як ImageNet, з подальшим дообученням на специфічних даних емоцій.

Основна ідея трансферного навчання полягає в тому, що нижні шари глибоких нейронних мереж виявляють загальні ознаки (краї, текстури, кольори), які є корисними для широкого спектру задач комп’ютерного зору. Верхні шари виявляють більш специфічні ознаки, характерні для конкретної задачі. Таким

чином, можна використати попередньо навчені ваги нижніх шарів та дообучити лише верхні шари або додати нові класифікаційні шари.

Математично процес трансферного навчання можна описати як оптимізацію функції втрат на цільовому домені D_T з використанням знань з вихідного домену D_S:

$$\theta^* = \arg\min_{\theta} L_{D_T}\left(f_\theta(X_T), Y_T\right) + \lambda R(\theta, \theta_S), \tag{1.14}$$

де θ – параметри моделі для цільової задачі;

θ_S – параметри попередньо навченої моделі на вихідному домені;

X_T, Y_T – дані та мітки цільового домену;

f_θ – функція моделі з параметрами θ;

L_{D_T} – функція втрат на цільовому домені;

$R(\theta, \theta_S)$ – регуляризаційний член, що забезпечує близькість до вихідних параметрів;

λ – коефіцієнт регуляризації.

Існує кілька стратегій трансферного навчання:

– заморожування, коли ваги всіх шарів, крім останніх класифікаційних, залишаються незмінними:

$$\theta = [\theta_{frozen}, \theta_{trainable}], \tag{1.15}$$

де θ_{frozen} – заморожені параметри попередньо навчених шарів (не оновлюються під час навчання);

$\theta_{trainable}$ – параметри нових або верхніх шарів, що піддаються навчанню.

– тонке налаштування (або донавчання), коли всі або частина шарів донавчаються з малою швидкістю навчання:

$$\theta_{l+1}^{(t)} = \theta_l^{(t)} - \alpha_l \nabla_{\theta_l} L, \tag{1.16}$$

де $\theta^{(t)}$ – параметри шару l;

α_l – швидкість навчання для шару l (зазвичай $\alpha_l < \alpha_{l+1}$ для нижчих шарів);

∇_{θ_l} L – градієнт функції втрат відносно параметрів шару l.

Для класифікації емоцій часто використовуються такі попередньо навчені архітектури, як VGG-16, ResNet-50, InceptionV3, MobileNet та EfficientNet [13, 14]. Ці моделі навчені на датасеті ImageNet, що містить понад 14 мільйонів зображень з 1000 класів.

Архітектура ResNet (Residual Network) використовує залишкові з'єднання (residual connections), що дозволяє навчати дуже глибокі мережі без проблеми зникаючого градієнта:

$$y = F(x, \{W_l\}) + x, \tag{1.17}$$

де x – вхідний тензор блоку;

y – вихідний тензор блоку;

$F(x, \{W_l\})$ – залишкове відображення, що навчається;

$\{W_l\}$ – набір вагів шарів у блоці.

Дослідження показують, що трансферне навчання особливо ефективне при обмежених обсягах даних. Для класифікації емоцій на датасетах розміром 5000 - 10000 зображень трансферне навчання може забезпечити покращення точності на 10-20% порівняно з навчанням з нуля [2].

Трансферне навчання з попередньо навченими моделями є одним із найбільш ефективних та практичних підходів до класифікації емоцій у медіаданих, особливо в умовах обмежених обчислювальних ресурсів та невеликих обсягів специфічних навчальних даних. Аналіз методу показує його численні переваги порівняно з навчанням моделей з нуля, що робить його оптимальним вибором для широкого спектру практичних застосувань.

Важливим аспектом є гнучкість методу, що проявляється у можливості вибору різних стратегій адаптації. Стратегія заморожування параметрів нижніх шарів дозволяє значно скоротити час навчання та обчислювальні витрати,

зберігаючи при цьому загальні візуальні ознаки, виявлені на великих датасетах. Стратегія тонкого налаштування, навпаки, забезпечує більш глибоку адаптацію моделі до специфіки емоційних виразів, що може бути критично важливим для досягнення максимальної точності. Вибір оптимальної стратегії залежить від подібності між вихідним доменом (ImageNet) та цільовим доменом (емоційні вирази), що може бути кількісно оцінено через косинусну подібність векторів ознак.

Далі, є необхідність правильного вибору гіперпараметрів для тонкого налаштування. Швидкість навчання для різних шарів повинна бути ретельно підібрана: нижні шари, що містять загальні візуальні ознаки, повинні навчатися з меншою швидкістю, ніж верхні шари, специфічні для задачі класифікації емоцій. Неправильний вибір швидкості навчання може призвести до руйнування корисних репрезентацій або, навпаки, до недостатньої адаптації моделі до нової задачі [3].

Перспективи розвитку методу пов’язані з використанням більш спеціалізованих попередньо навчених моделей. Замість ImageNet можна використовувати моделі, попередньо навчені на великих датасетах облич, таких як VGGFace2 або MS-Celeb-1M, що містять мільйони зображень облич різних людей. Такий підхід, відомий як domain-specific transfer learning, може забезпечити ще кращі результати, оскільки вихідний домен буде ближчим до цільової задачі класифікації емоцій [25, 26]. Іншим перспективним напрямком є використання мультизадачного навчання у поєднанні з трансферним навчанням. Модель може одночасно навчатися розпізнавати емоції, визначати вік, стать та інші атрибути обличчя, що дозволяє виявляти більш узагальнені та інформативні ознаки. Спільне навчання на кількох пов’язаних задачах часто покращує якість розв’язання кожної окремої задачі завдяки ефекту регуляризації та кращому використанню обмежених даних.

Таким чином, трансферне навчання з попередньо навченими моделями є високоефективним, практичним та гнучким методом для класифікації емоцій у медіаданих. Метод забезпечує оптимальний баланс між точністю, швидкістю

навчання та обчислювальною ефективністю, що робить його одним із найбільш популярних підходів у сучасних системах аналізу емоційного стану користувачів. Правильний вибір архітектури, стратегії адаптації та гіперпараметрів дозволяє досягати точності класифікації на рівні 80-90% на стандартних датасетах емоційних виразів, що є достатнім для більшості практичних застосувань у системах дистанційної освіти, телемедицини, маркетингових досліджень та аналізу соціальних медіа.

1.3 Метод класифікації на основі рекурентних нейронних мереж для аудіоданих.

Класифікація емоцій в аудіоданих є важливою складовою аналізу емоційного стану користувачів, оскільки голос містить багату інформацію про емоційний стан людини через просодичні характеристики, такі як тон, інтонація, темп мовлення та енергія сигналу. Рекурентні нейронні мережі (Recurrent Neural Networks, RNN) та їх модифікації, зокрема LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory) та GRU (Gated Recurrent Unit), є ефективними інструментами для обробки послідовних даних, таких як аудіосигнали [4].

1.3.1 Специфіка аудіоданих та просодичні ознаки емоцій

Аудіосигнал, на відміну від статичних зображень, є інформацією, розташованою в часовій послідовності. Кожна секунда аудіозапису репрезентує часовий ряд амплітудних значень, які містять мільйони точок дискретизації. Просодія – це набір фізичних параметрів мовлення, які суттєво впливають на передачу емоцій. Цей набір складається з:

– основна частота (pitch, F0) – висота голосу, яка корелює з емоційним

збудженням; гнів характеризується вищим pitch, сум – нижчим;

– інтенсивність – енергія звукового сигналу, що відображає силу мовлення; радість звичайно супроводжується вищою енергією, ніж сум;

– темп мовлення – швидкість вимовляння слів; прискорений темп часто пов'язаний з тривогою, сповільнений – з сумом чи задумливістю;

– тривалість та паузи – характер стислості або розтягнутості звуків; тривалі паузи можуть сигналізувати про невпевненість;

– форманти – резонансні частоти голосового тракту, які варіюють залежно від емоційного напруження;

– периодичність та шум – невеликі коливання частоти та амплітуди, що зростають при стресі.

Традиційні методи обробки звука спиралися на ручне виділення цих ознак. Однак такі підходи мали обмеження: вони залежали від експертної думки щодо того, які ознаки найбільш релевантні, і часто не могли виявити складні взаємодії між ознаками. Рекурентні нейронні мережі вирішили цю проблему завдяки здатності автоматично виділяти ознаки з часових послідовностей.

1.3.2 Архітектура та принцип роботи RNN

Рекурентні нейронні мережі (Recurrent Neural Networks, RNN) розроблені спеціально для обробки послідовних даних. У контексті аналізу емоцій в аудіоданих це є критично важливим, оскільки емоційне забарвлення мовлення є динамічним процесом, де поточний стан залежить від попередніх фонем, слів чи інтонаційних патернів.

Ключовою особливістю архітектури RNN є наявність зворотних зв'язків (loops), які дозволяють інформації зберігатися в мережі. Формально це реалізується через наявність прихованого стану, який діє як пам'ять мережі.

RNN обробляє послідовність входів $X = (x_1, x_2, \dots, x_T)$ крок за кроком. На кожному часовому кроці t мережа отримує вхідний вектор x_t та власний

прихований стан з попереднього кроку h_{t-1}.

Рекурентну мережу можна представити у вигляді "розгорнутого" (unfolded) графа. Це означає, що для послідовності довжиною T ми фактично створюємо ланцюг із T ідентичних блоків нейронної мережі, кожен з яких передає повідомлення наступному, як представлено на рисунку 1.1.

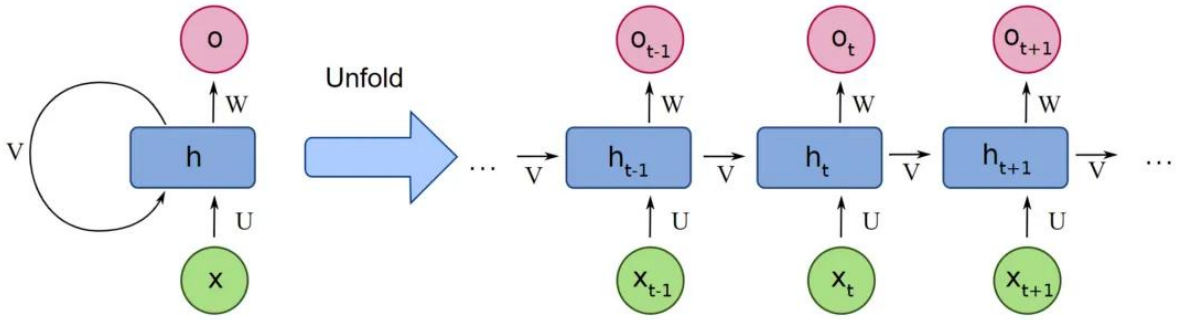

Рисунок 1.1 – Ланцюг ідентичних блоків нейронної мережі (англ.)

Тепер розглянемо архітектуру цієї системи. Вона складається з трьох основних типів шарів:

Компонент 1. Вхідний шар (input layer) – шар, приймає вхідний вектор ознак (наприклад, фрейм спектрограми) на кожному часовому кроці.

Компонент 2. Приховані шари (hidden layers) – це ключовий елемент архітектури (рис. 3, на рисунку позначено зеленим та блакитним кольорами). Вигнуті стрілки, що виходять з нейронів прихованого шару і повертаються до них же (або до наступного стану цього ж шару), візуалізують рекурентний зв'язок. Це означає, що вихід нейрона в момент часу t подається йому ж на вхід в момент часу $t + 1$. Така структура дозволяє мережі формувати прихований стан (hidden state), який акумулює історію всієї послідовності. На рисунку 3 показано глибоку RNN (Deep RNN) з двома послідовними прихованими шарами, що дозволяє виділяти більш абстрактні ознаки.

Компонент 3. Вихідний шар (output layer) – шар, що генерує результат обробки (наприклад, ймовірність класу емоції) на основі станів останнього прихованого шару.

Структурна схема глибокої рекурентної нейронної мережі представлена на рисунку 1.2.

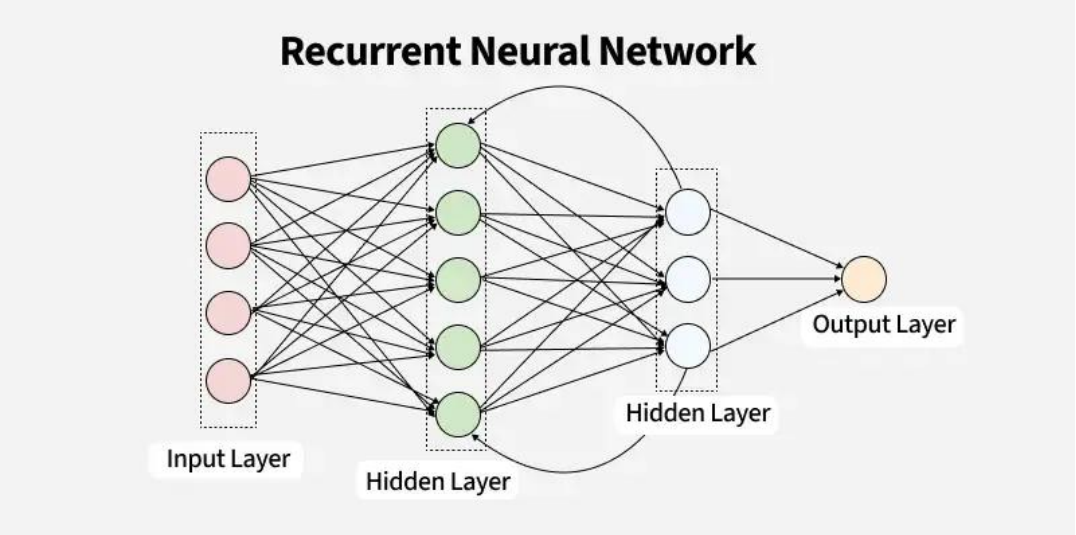

Рисунок 1.2 – Структурна схема глибокої рекурентної нейронної мережі (Deep RNN) (англ.)

Математично базова RNN описується рівняннями:

– оновлення прихованого стану: прихований стан h_t на момент часу t обчислюється як функція від поточного входу та попереднього стану:

$$h_t = tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad (1.18)$$

де $tanh$ – це нелінійна функція активації, що скорочено означає гіперболічний тангенс, яка передає значення з будь-якого діапазону в діапазон від -1 до +1;

$x_t \in R^d$ – вхідний вектор ознак на кроці t (наприклад, фрейм спектрограми);

$h_{t-1} \in R^h$ – прихований стан на попередньому кроці;

$W_{xh} \in R^{h\times d}$ – матриця вагових коефіцієнтів для вхідних даних;

$W_{hh} \in R^{h\times h}$ – матриця вагових коефіцієнтів для прихованого стану;

$b_h \in R^h$ – вектор зміщення.

– обчислення виходу: після оновлення прихованого стану мережа може згенерувати вихід y_t (якщо це необхідно на даному кроці, наприклад, у задачах

sequence-to-sequence, Seq2Seq) або використати останній стан h_T для класифікації всієї послідовності:

$$y_t = W_{hy}h_t + b_y, \tag{1.19}$$

де x_t – вхідний вектор на часовому кроці t;

h_t – прихований стан на кроці t;

W_{hy} – матриця ваг вихідного шару;

b_y – зміщення вихідного шару.

Важливою характеристикою RNN є принцип спільного використання параметрів. На відміну від звичайних глибоких мереж, де кожен шар має свої унікальні ваги, в розгорнутій RNN матриці W_{xh}, W_{hh} та W_{hy} є однаковими для всіх часових кроків t. Це дозволяє значно зменшити кількість параметрів, що навчаються, і дозволяє мережі обробляти послідовності довільної довжини [5].

Навчання RNN відбувається за алгоритмом зворотного поширення помилки в часі (Backpropagation Through Time, BPTT). Цей метод є узагальненням класичного алгоритму Backpropagation для розгорнутих графів. Помилка L (Loss function) обчислюється на кожному кроці або в кінці послідовності, після чого градієнти поширюються у зворотному напрямку від $t = T$ до $t = 1$. Повний градієнт для вагової матриці W_{hh} є сумою градієнтів на кожному часовому кроці. Математично повний градієнт визначається як сума часткових похідних помилки на кожному часовому кроці:

$$\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_{hh}}, \tag{1.20}$$

Критичною проблемою базових рекурентних нейронних мереж є явище зникаючих/вибухових градієнтів. Під час зворотного поширення помилки в часі градієнти можуть експоненціально спадати, що унеможливлює навчання довготривалих залежностей, або експоненціально зростати, спричиняючи нестабільність процесу навчання та появу значень NaN. У результаті, мережа

втрачає здатність ефективно запам’ятовувати інформацію, розділену на багато кроків послідовності.

1.3.3 Архітектура сучасних RNN-систем для аналізу емоцій в аудіоданих

Класифікація емоцій в аудіо (Audio Emotion Recognition, AER) є однією з найскладніших задач обробки мови. Вона вимагає від моделі не тільки розпізнавання лінгвістичного вмісту, а й інтерпретації просодичних та паралінгвістичних характеристик, які проявляються протягом усього висловлювання. Сучасні системи AER відійшли від використання «чистих» RNN/LSTM і покладаються на складні гібридні архітектури.

Фундаментом такої системи є двонаправлена LSTM (BiLSTM) система [6]. Дана архітектура є критично важливою для максимізації розуміння емоційного контексту, оскільки, на відміну від стандартної LSTM, яка враховує лише попередній контекст, BiLSTM дозволяє моделі «бачити» всю послідовність цілком, нівелюючи проблему втрати контексту в довгих аудіозаписах (рис. 1.3).

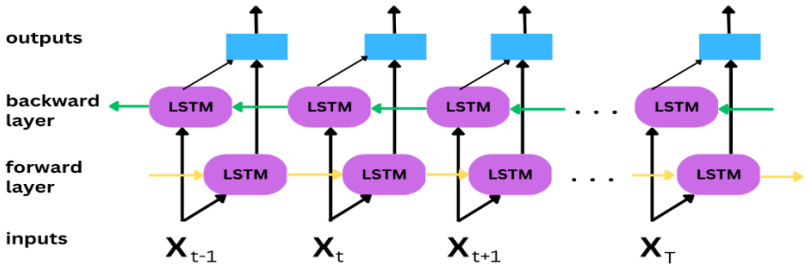

Рисунок 1.3 – Архітектура BiLSTM системи (англ.)

Архітектурно BiLSTM складається з двох незалежних шарів LSTM:

– forward layer – шар, що обробляє вхідну послідовність ознак x у прямому хронологічному порядку (від t = 1 до t = T);

– backward layer – шар, що обробляє ту ж послідовність у зворотному

порядку (від t = T до t = 1).

Остаточний прихований стан h_t для BiLSTM на часовому кроці t формується шляхом конкатенації виходів обох шарів. Це дозволяє отримати вектор, що містить інформацію як про передумови виникнення певного акустичного патерну, так і про його подальший розвиток:

$$h\_t = ht \oplus ht, \tag{1.21}$$

де $\oplus$ позначає операцію поелементної конкатенації;

$\vec{h}_t$ – вихідний вектор прямого шару;

$\overleftarrow{h}_t$ – вихідний вектор зворотного шару.

Така структура дозволяє мережі ефективно моделювати складні емоційні переходи, наприклад, іронію, яка стає зрозумілою лише в кінці фрази, або різку зміну тону.

Використання «чистої» BiLSTM архітектури для аналізу "сирих" спектрограм часто є неефективним через високу розмірність вхідних даних. Тому сучасним стандартом в AER (Audio Emotion Recognition) є використання гібридних моделей типу CRNN (Convolutional Recurrent Neural Network). Найбільш ефективні сучасні моделі поєднують згорткові шари для виділення локальних спектральних ознак з рекурентними шарами для обробки часових залежностей. У такій архітектурі вхідний потік даних спочатку проходить через згорткові шари CNN, які виступають у ролі екстракторів локальних ознак. Вони зменшують спектральну розмірність і виділяють ключові патерни (наприклад, форманти або гармоніки), формуючи компактні карти ознак, які потім подаються на вхід рекурентним шарам BiLSTM для моделювання часової динаміки. що складається з декількох етапів.

Етап 1. Спектрограма аудіо (представлена як 2D матриця) обробляється CNN для виділення спектральних ознак (формант, форманти, шумові характеристики).

Етап 2. Послідовність признак, виділених CNN, подається на вхід BiLSTM

для захоплення часових закономірностей.

Етап 3. Механізм уваги дозволяє моделі зосередитися на найбільш інформативних часових фрагментах.

Етап 4. Вихідний шар: повносвязна мережа з softmax активацією видає ймовірності для кожного класу емоції.

Успіх такої RNN залежить від якості представлення вхідних даних. Типовий пайплайн передобробки включає:

– зчитування аудіофайлу та вибір частоти дискретизації (44,1 кГц, 48 кГц);

– нормалізація амплітуди для забезпечення узгодженості між записами;

– видалення тиші та шумів через noise gate та спектральне віднімання;

– перетворення аудіо на спектрограму за допомогою короткочасного перетворення Фур'є (Short-Time Fourier Transform, STFT) або мел- спектрограми (mel-spectpogram), яка краще відповідає сприйманню людським вухом;

– розділення на перекриваючи вікна розміром 25-40 мс з кроком 10 мс для уловлювання динамічних змін просодичних параметрів;

– виділення просодичних ознак: основна частота, енергія, темп мовлення за допомогою спеціалізованих бібліотек (librosa, pyannote);

– аугментація даних: часовий зсув, зміна темпу, додавання фонового шуму для збільшення різноманітності тренувального набору.

Рекурентні нейронні мережі є критично важливими інструментами для аналізу емоцій в аудіоданих. Їх здатність обробляти часові послідовності та «запам'ятовувати» емоційні сигнали робить їх незамінними для розпізнавання просодичних характеристик мовлення. Однак для досягнення найкращих результатів необхідно комбінувати їх з іншими архітектурами (CNN, Attention, Transformer) та застосовувати мультимодальні підходи.

1.4 Мультимодальний підхід до класифікації емоцій

У реальній міжособистісній комунікації емоції виражаються не лише через один канал передачі інформації. Людина сприймає емоційний стан співрозмовника комплексно: через інтонацію голосу (аудіо), міміку та жести (відео), а також через семантичний зміст сказаного (текст).

Традиційні унімодальні системи (наприклад, тільки Audio Emotion Recognition) стикаються з проблемою емоційної неоднозначності. Наприклад, саркастичне «Яка чудова погода», сказане сумним голосом під дощем, буде класифіковано текстовою моделлю як «Радість» (через слово «чудова»), тоді як аудіо-модель визначить «Сум» або «Сарказм».

Мультимодальний підхід (Multimodal Emotion Recognition, MER) вирішує цю проблему шляхом інтеграції даних з різних джерел, досягаючи ефекту синергії, де точність об’єднаної системи перевищує точність кожної окремої модальності [29, 30].

Архітектура мультимодальної системи. Типова архітектура MER для медіаданих базується на обробці трьох основних потоків:

– акустична модальність для вилучення просодичних ознак (MFCC, хроматограми) за допомогою CRNN або трансформерів (наприклад, Wav2Vec);

– візуальна модальність для аналізу виразу обличчя (Facial Expression Recognition) за допомогою глибоких згорткових мереж, таких як ResNet або EfficientNet, що обробляють кадри відеопотоку;

– лексична модальність для аналізу транскрипції мовлення за допомогою NLP-моделей (наприклад, BERT або RoBERTa) для розуміння семантичного контексту.

Основною проблемою мультимодального навчання є ефективне об’єднання інформації з різних модальностей, які можуть мати різну природу, розмірність та часову структуру. Існує три основні стратегії об’єднання: раннє злиття, пізнє злиття та гібридне злиття.

Стратегія «раннього злиття». У цьому підході ознаки з різних

модальностей об'єднуються в єдиний вектор до подачі на вхід класифікатора. Нехай $x_a \in R^{d_a}$ – вектор ознак аудіо, а $x_v \in R^{d_v}$ – вектор візуальних ознак. Об'єднаний вектор z формується шляхом конкатенації:

$$z = [x_a \oplus x_v], \tag{1.22}$$

Цей вектор подається на вхід нейронної мережі для навчання.

Перевага полягає в тому, що є можливість моделі вивчати кореляції між ознаками низького рівня на ранніх етапах. Недоліком є проблема «прокляття розмірності» (висока розмірність вектора z) та складність синхронізації потоків, що мають різну частоту дискретизації (наприклад, аудіо 44.1 кГц та відео 30 fps).

Стратегія «пізнього злиття». Кожна модальність обробляється окремою незалежною моделлю, яка видає власний прогноз (вектор ймовірностей). Фінальне рішення приймається шляхом агрегації цих прогнозів. Якщо P_a, P_v, P_t – вектори ймовірностей емоцій для аудіо, відео та тексту відповідно, то фінальний прогноз P_{final} розраховується як зважена сума [7]:

$$P_{final} = \text{argmax}(w_a P_a + w_v P_v + w_t P_t), \tag{1.23}$$

де w_a, w_v, w_t – вагові коефіцієнти довіри до кожної модальності при умові $\sum w_i = 1$.

Перевагою є стійкість до відсутності однієї з модальностей та простота реалізації. Як недолік можна вказати втрату інформації про взаємозв'язок між модальностями (наприклад, рух губ і звук не корелюються моделлю).

Стратегія «гібридного злиття» є сучасним підходом (State-of-the-Art), де злиття відбувається на проміжних шарах глибокої нейронної мережі, часто з використанням механізмів Cross-Modal Attention. Це дозволяє одній модальності «фокусуватися» на важливих частинах іншої (наприклад, модель звертає більше уваги на вираз обличчя, коли інтонація голосу є нейтральною, але слова – емоційними). Ваги модальностей можуть бути фіксованими (рівними або

визначеними експертами) або навчальними параметрами моделі. Адаптивні ваги можуть обчислюватися динамічно для кожного зразка за допомогою механізму уваги.

Мультимодальний підхід забезпечує значно вищу точність класифікації (F1-score) та надійність системи порівняно з унімодальними методами. У рамках даного дослідження розуміння принципів мультимодальності є необхідним для побудови масштабованої системи аналізу емоційного стану користувачів [8].

1.5 Постановка задачі дослідження

Класифікація емоцій у медіаданих є актуальним завданням, що має широке практичне застосування в системах аналізу емоційного стану користувачів, автоматизованій психологічній діагностиці, інтерактивних системах та соціальних медіа. Прийнято рішення щодо розроблення програмного застосунку класифікації емоцій у медіаданих трьома методами, а саме із застосуванням CNN для зображень облич, рекурентних нейронних мереж з механізмом уваги (LSTM+Attention) для аудіоданих та мультимодального підходу з трансферним навчанням для комбінованих аудіовізуальних даних.

Предметом дослідження є методи машинного навчання та глибокого навчання для автоматичної класифікації емоцій у візуальних та аудіальних медіаданих.

Метою дослідження є порівняння ефективності різних методів класифікації емоцій у медіаданих шляхом розробки програмного застосунку, що класифікує базові емоції людини на зображеннях облич та в аудіозаписах мовлення.

Завдання дослідження передбачають:

– провести аналіз літературних джерел щодо апробації методів класифікації емоцій у медіаданих, включаючи огляд сучасних датасетів, метрик оцінювання та архітектур нейронних мереж;

– виконати порівняльний аналіз переваг та недоліків кожного методу,

визначити їх обчислювальну складність та вимоги до даних;

– сформувати покроковий алгоритм для кожного із вибраних методів класифікації емоцій, включаючи етапи попередньої обробки даних, виділення ознак, навчання моделі та класифікації;

– візуалізувати покроковий алгоритм кожного із вибраних методів блок-схемою з детальним описом кожного етапу обробки;

– підібрати та підготувати датасети для навчання та тестування моделей, включаючи FER-2013 для зображень облич та RAVDESS або TESS для аудіоданих;

– розробити програмний застосунок на мові Python з використанням бібліотек TensorFlow/Keras або PyTorch, що надасть змогу класифікувати емоції у медіаданих кожним із вибраних методів;

– реалізувати модулі попередньої обробки даних, включаючи детекцію облич на зображеннях, виділення MFCC-ознак з аудіо та нормалізацію даних;

– навчити та оптимізувати гіперпараметри кожної з моделей, використовуючи методи крос-валідації та регуляризації для запобігання перенавчанню;

– провести експериментальне дослідження ефективності одного з методів на тестових даних, обчислити метрики точності, повноти, точності та F1-міри для кожного класу емоцій;

– розробити інтерфейс користувача для програмного застосунку, що дозволить завантажувати медіадані та отримувати результати класифікації емоцій у зручному форматі;

– підготувати документацію до програмного застосунку, включаючи інструкції з встановлення, налаштування та використання.

Об’єктом дослідження є медіадані, що містять емоційну інформацію, зокрема фрагменти тексту, зображення облич людей та аудіозаписи мовлення з емоційним забарвленням.

2 ТЕОРЕТИЧНІ ОСНОВИ ВИКОРИСТАННЯ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

2.1 Фундаментальні концепції та архітектура згорткових нейронних мереж

У сучасній системі обробки медіаданих CNN відіграють ключову роль, оскільки забезпечують автоматичне та ефективне виділення ознак із зображень, відео та аудіосигналів. На відміну від традиційних методів обробки даних, де ознаки потрібно було конструювати вручну, CNN дозволяють моделі самостійно формувати набір релевантних характеристик, адаптований до складності вхідного сигналу [9]. Це робить їх фундаментальним інструментом для задач класифікації емоцій, де важливо вловлювати мінімальні зміни міміки, інтонації чи динаміки рухів.

Поява CNN стала можливою завдяки розвитку обчислювальних технологій та збільшенню обсягів даних, необхідних для навчання глибоких моделей. Сьогодні щосекунди генеруються мільйони зображень і відео, значна частина яких відображає емоційні стани користувачів – наприклад, у соціальних мережах, стримінгових сервісах чи під час онлайн-комунікацій. Обробити такі обсяги традиційними алгоритмами було б неможливо, тоді як CNN демонструють високу ефективність у виявленні складних нелінійних патернів.

Основною ідеєю CNN є використання згортки (convolution) – математичної операції, яка дозволяє аналізувати локальні області зображення або іншого сигналу [10, 11]. Згорткові ядра (фільтри) проходять через весь простір даних, виявляючи такі структурні елементи, як краї, кути, текстури, рухи або спектральні зміни. Завдяки цьому мережа формує ієрархію ознак: від низькорівневих (лінії, контури) до високорівневих (емоційні патерни виразу обличчя, артикуляція, динаміка погляду). На рисунку 2.1 представлено загальну схему формування ознак у CNN.

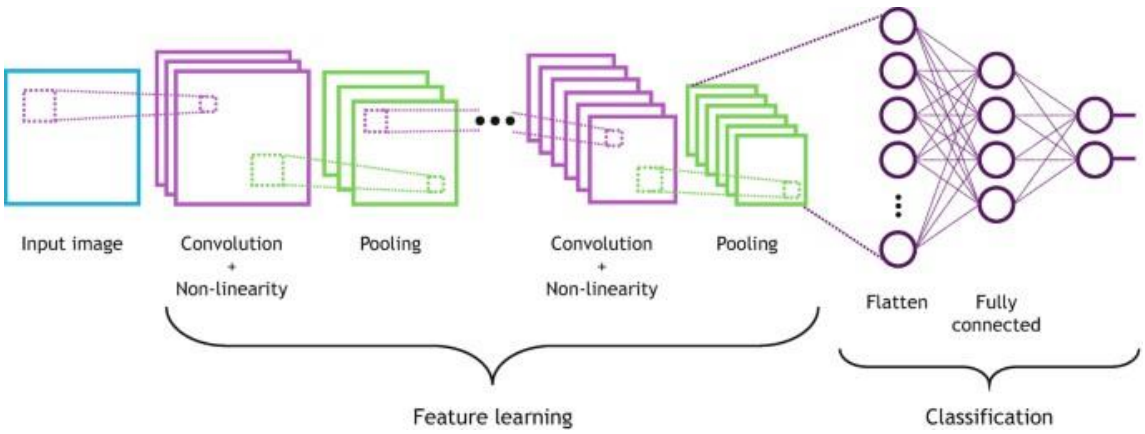

Рисунок 2.1 – Загальна архітектура згорткової нейронної мережі (англ.)

Архітектура згорткової нейронної мережі складається з двох основних етапів: навчання ознак та класифікації. Процес починається з вхідного зображення, яке надходить у мережу для обробки та аналізу. Вхідне зображення є початковою точкою обчислювального пайплайну, де кожен піксель або група пікселів стає джерелом інформації для подальшої обробки. Існує два етапи навчання мереж.

Етап навчання ознак (Feature Learning). На етапі навчання ознак вхідне зображення проходить через послідовність згорткових та пулінг-шарів, які витягують релевантні характеристики. Згортка з нелінійністю застосовує фільтри (ядра) до вхідного зображення, виявляючи локальні патерни, такі як краї, текстури або просторові залежності. Фільтри переміщуються по всьому зображенню, створюючи карти ознак, які підсилюють важливі деталі та пригнічують шум. Нелінійна функція активації, зазвичай ReLU [12], додає нелінійність до моделі, дозволяючи їй навчатися складніших залежностей між даними. Після згортки слідує шар пулінгу, який зменшує розмірність отриманих карт ознак, зберігаючи найважливішу інформацію. Пулінг виконує операції усереднення або вибору максимуму над невеликими областями, що дозволяє зменшити кількість параметрів та зробити модель більш стійкою до невеликих зсувів у вхідних даних. Цей процес також знижує обчислювальну складність та сприяє виявленню більш абстрактних ознак на наступних рівнях. Архітектура

передбачає повторення блоків згортки та пулінгу кілька разів, що дозволяє мережі навчатися ієрархії ознак від простих до складних. Кожен наступний шар виявляє все більш абстрактні та складні патерни, починаючи від базових країв та текстур до складних об'єктів та їх взаємозв'язків. Глибина мережі визначає її здатність розпізнавати складні структури у зображеннях.

Етап класифікації (Classification). Після завершення етапу навчання ознак отримані багатовимірні карти ознак проходять через операцію випрямлення (Flatten), яка перетворює двовимірні або тривимірні структури в одновимірний вектор. Це необхідно для підготовки даних до подальшої обробки повнозв'язаними шарами, які працюють з лінійними послідовностями значень. Випрямлення зберігає всю інформацію, накопичену під час етапу навчання ознак, але представляє її у форматі, зручному для фінальної класифікації. Випрямлений вектор ознак надходить у повнозв'язані шари, які виконують фінальну класифікацію або регресію. Кожен нейрон у повнозв'язаному шарі з'єднаний з усіма нейронами попереднього шару, що дозволяє моделі інтегрувати всі виявлені ознаки та приймати рішення на основі їх комбінації. Приховані повнозв'язані шари дозволяють моделі навчатися складним нелінійним залежностям між ознаками та цільовими класами. Вихідний шар мережі генерує фінальні передбачення, які можуть представляти ймовірності належності до різних класів, координати об'єктів або інші цільові значення залежно від конкретного завдання. Кількість нейронів у вихідному шарі зазвичай відповідає кількості класів для класифікації або кількості вихідних параметрів для регресійних задач [12-14].

Архітектура CNN дозволяє ефективно обробляти зображення та інші просторові дані, автоматично виявляючи релевантні ознаки та приймаючи інтелектуальні рішення на їх основі. Однією з ключових характеристик CNN є вагове спільне використання. На відміну від традиційних нейронних мереж, де кожен піксель має окремий набір ваг, в CNN один фільтр аналізує весь простір даних. Це значно зменшує кількість параметрів, дозволяє ефективно працювати з великими медіаданими та знижує ризик перенавчання. Особливо

важливо це у задачах емоційного аналізу, де моделі часто працюють з різноманітними джерелами – від відео високої роздільності до аудіоспектрограм.

Ще одним важливим аспектом є інваріантність до зсувів. Завдяки pooling-операціям CNN здатні розпізнавати емоційні патерни незалежно від точного положення об’єкта в кадрі або незначних змін у позі чи міміці. Це особливо актуально під час аналізу виразів облич у реальних умовах, де користувачі рухаються, змінюють освітлення або частково закривають обличчя.

Поширення CNN стало можливим завдяки появі графічних процесорів (GPU) та спеціалізованих бібліотек на зразок TensorFlow і PyTorch [15], які підтримують оптимізовані операції згортки. Це зробило моделі доступними для широкого кола завдань, включно з класифікацією емоцій в реальному часі. Сучасні CNN здатні обробляти десятки кадрів за секунду та адаптуватися до нових умов через transfer learning, використовуючи попередньо навчені моделі на великих датасетах.

Проте, попри високу ефективність, CNN мають і низку викликів. Зокрема, аналіз емоційних медіаданих вимагає врахування контексту– послідовності рухів, темпу зміни емоцій або аудіальних нюансів. У таких випадках класичні CNN доповнюють рекурентними мережами, трансформерами або механізмами уваги. Крім того, важливою проблемою залишається інтерпретованість: глибокі моделі часто діють як «чорна скринька», що ускладнює пояснення, які саме ознаки вплинули на класифікацію емоції. Серед актуальних напрямів досліджень – методи пояснюваного штучного інтелекту, теплові карти активацій та візуалізація фільтрів.

Таким чином, згорткові нейронні мережі надають потужний інструментарій для аналізу візуальних та аудіальних медіаданих, дозволяючи з високою точністю визначати емоційний стан користувачів. Їхня сила полягає у здатності працювати з великими обсягами інформації, автоматично формувати ознаки та адаптуватися до складних реальних умов, що робить CNN невід’ємною складовою сучасних систем емоційної аналітики [15].

2.2 Джерела та типи медіаданих для навчання та тестування систем

Розвиток аналізу емоційного стану залежить від різноманітності та якості медіаданих, що генеруються в цифровому середовищі. Сьогодні цифрова комунікація перестала бути лише обміном інформацією – вона стала багатовимірною системою, де майже кожне повідомлення, зображення чи аудіозапис можна виміряти, проаналізувати та перетворити на дані про емоційний стан користувача. Завдяки цьому з’явилася можливість не просто читати тексти чи переглядати контент, а глибше розуміти емоційну логіку комунікації, аналізувати закономірності вираження емоцій та навіть передбачати емоційні реакції на різні події чи контент. Медіадані поступово перетворилися на стратегічний ресурс: вони допомагають психологам та соціологам розуміти суспільні настрої, маркетологам – будувати емоційно релевантні кампанії, а розробникам систем – створювати більш чутливі до емоцій користувачів інтерфейси.

У контексті великих даних медіадані для аналізу емоційного стану поділяються на кілька основних типів залежно від способу отримання, структури та призначення. Найчастіше виділяють чотири великі категорії даних: текстові (вербальні вирази емоцій), візуальні (невербальні сигнали емоцій), аудіальні (паралінгвістичні характеристики мовлення та звукові сигнали) та мультимодальні . Кожна категорія має свої унікальні характеристики, джерела та методи обробки, що визначають їх придатність для різних завдань аналізу емоційного стану. Розглянемо кожне із них.

2.2.1 Текстові джерела медіаданих

Текстові дані є одним з найпоширеніших джерел для аналізу емоцій. Соціальні мережі, форуми, блоги, месенджери та інші платформи генерують великі обсяги текстів, що містять емоційну інформацію. Кожен текст може

містити явні або приховані індикатори емоційного стану через лексику, синтаксис, пунктуацію та стилістичні особливості.

Основними текстовими джерелами даних є:

– соціальні мережі (X/Twitter, Facebook, Instagram). Є одним із найпотужніших джерел даних для навчання моделей, оскільки характеризуються надвисокою динамікою оновлення та специфічним форматом контенту. Обмеження на довжину повідомлень (особливо у Twitter/X) та культура «швидкого споживання» інформації змушують користувачів висловлювати емоції у концентрованому вигляді, часто «в моменті» переживання події. Для текстів із цих платформ характерне використання специфічного сленгу, скорочень, хештегів та неформального синтаксису, що створює певні виклики для класичних NLP-алгоритмів, але водночас надає багатий матеріал для аналізу реальних, невідфільтрованих емоційних реакцій. Величезні обсяги таких даних дозволяють тренувати нейромережі на розпізнавання сарказму, агресії чи радості в умовах зашумленого тексту;

– платформи відгуків (IMDb, Amazon, Google Reviews). Становлять особливу цінність для дослідників завдяки наявності чіткої структурної прив’язки тексту до кількісної оцінки. На відміну від хаотичних постів у соцмережах, відгуки зазвичай мають бімодальну природу: текстовий коментар супроводжується рейтингом (наприклад, зірки від 1 до 5), що слугує природною міткою «істини» для навчання з учителем. Це значно спрощує процес підготовки датасетів, оскільки дозволяє автоматично класифікувати тексти за полярністю (позитивний/негативний досвід). Крім того, такі дані часто містять аспектно-орієнтовані емоції, коли користувач може висловлювати захоплення одним аспектом продукту (наприклад, сценарієм фільму), але розчарування іншим (акторською грою), що дозволяє навчати моделі глибшому розумінню контексту;

– форуми та дискусійні платформи (Reddit, Stack Overflow, Quora). Відрізняються від інших джерел своєю деревовидною структурою комунікації, де емоції розгортаються не в ізольованих повідомленнях, а в ланцюжках взаємодії. На таких платформах, як Reddit, користувачі об’єднуються у тематичні

спільноти (сабреддіти), що дозволяє аналізувати специфічні емоційні патерни, притаманні конкретним групам інтересів. Тексти тут зазвичай є більш розгорнутими та аргументованими, ніж у соцмережах, що дає змогу досліджувати переходи емоційних станів у процесі дискусії, виявляти колективні настрої та аналізувати рівень токсичності або підтримки всередині спільноти. Великий обсяг щоденних публікацій (понад 100 мільйонів коментарів) робить ці платформи ідеальним полігоном для вивчення соціальної динаміки емоцій;

– блогові платформи та довгі текстові форми (Medium, WordPress, LiveJournal). Надають матеріал для аналізу глибинних емоційних станів та складних наративних структур. Формат лонгрідів дозволяє авторам детально описувати свої рефлексії, спогади та переживання, що дає можливість алгоритмам відстежувати динаміку зміни емоційного фону протягом одного тексту – від зав’язки до висновку. Особливістю цього типу даних є сильна залежність емоційної лексики від тематики блогу: наприклад, тревел-блоги частіше насичені маркерами захоплення та здивування, тоді як блоги психологічної спрямованості можуть містити складні комбінації суму, тривоги та надії. Аналіз таких текстів дозволяє виходити за рамки простої класифікації «позитив/негатив» і виявляти стійкі психологічні патерни особистості автора;

– месенджери та чат-боти (Telegram, WhatsApp, Viber). Представляють сегмент найбільш приватної та безпосередньої комунікації, де емоційні прояви часто є найбільш відвертими. Тексти в особистих повідомленнях характеризуються високою щільністю паралінгвістичних маркерів, які компенсують відсутність візуального контакту: активне використання емодзі, стікерів, написання слів великими літерами (Caps Lock) для імітації крику, або специфічна пунктуація (наприклад, множинні знаки оклику). Ці елементи часто несуть більше емоційне навантаження, ніж самі слова. Аналіз діалогів у месенджерах дозволяє досліджувати інтерактивну природу емоцій, коли стан одного співрозмовника впливає на стан іншого, а також виявляти специфічні патерни цифрового етикету та неформального вираження почуттів.

2.2.2 Візуальні джерела медіаданих

Візуальні дані, включаючи зображення та відео, містять багату емоційну інформацію через міміку, жести, позу та інші невербальні сигнали. Фотографії в соціальних мережах часто відображають емоційний стан користувачів безпосередньо через вираз обличчя, або опосередковано через вибір сюжету, кольорової гами, композиції та інших візуальних елементів. Аналіз зображень обличчя для визначення емоцій є класичним завданням комп'ютерного зору, де згорткові нейронні мережі демонструють високі результати завдяки здатності виявляти локальні патерни та просторові залежності.

Основними візуальними джерелами даних є:

– соціальні мережі з акцентом на фотоконтент (Instagram, Pinterest, Flickr). Ці платформи є найбільшими у світі сховищами статичних зображень, що містять обличчя людей (Facial Imagery). Вони надають величезну варіативність даних: від професійних портретів з ідеальним освітленням до побутових «селфі» з низькою якістю та складними ракурсами. Така різноманітність є критично важливою для тренування стійких CNN [16], здатних розпізнавати базові емоції (радість, сум, злість тощо) незалежно від умов зйомки. Крім того, наявність хештегів та описів дозволяє використовувати методи слабкого навчання для попередньої автоматичної розмітки емоцій;

– відеохостинги та платформи користувацького контенту (YouTube, TikTok, Vimeo). Динамічні візуальні дані з цих джерел дозволяють аналізувати емоції в часі. Відеоблоги (влоги), реакції (reaction videos) та стріми є джерелом так званих «спонтанних» емоцій, які відрізняються від награнних «постановочних» виразів обличчя акторів. Відеодані дозволяють фіксувати не лише статичний емоційний стан, а й мікровирази – швидкі, мимовільні рухи м'язів обличчя, що тривають частки секунди і часто видають приховані справжні почуття, які людина намагається замаскувати;

– системи відеоконференцзв'язку та вебінарів (Zoom, Microsoft Teams, Google Meet). Масовий перехід на віддалену роботу перетворив записи онлайн-

зустрічей на цінне джерело візуальних даних для аналізу ділової комунікації. Специфікою цього джерела є фронтальний ракурс зйомки («голова, що говорить») та відносно стабільне положення користувача перед камерою. Аналіз таких даних дозволяє оцінювати рівень залученості, втоми, нудьги або фрустрації учасників під час навчання чи робочих нарад, що є основою для систем Affective Computing у корпоративному та освітньому секторах;

– системи відеоспостереження та аналітики натовпу (CCTV). Камери спостереження у громадських місцях, торгових центрах та на стадіонах генерують дані, що дозволяють аналізувати емоційний стан груп людей або поведінкові патерни на відстані. Тут акцент зміщується з детального аналізу міміки (яку часто не видно через низьку роздільну здатність) на аналіз мови тіла та пози. Нахил голови, хода, жестикуляція рук та загальна динаміка рухів дозволяють класифікувати стани тривоги, агресії або паніки, що є критично важливим для систем безпеки [17];

– кінематографічний та телевізійний контент (фільми, серіали, ток-шоу). Хоча емоції в таких джерелах є «зіграними», вони залишаються еталоном для навчання базових моделей розпізнавання. Професійне освітлення та чітка артикуляція емоцій акторами дозволяють створювати високоякісні анотовані датасети (наприклад, кадри з фільмів, де емоція є очевидною та однозначною). Це джерело часто використовується на початкових етапах тренування моделей перед їх донавчанням на більш складних реальних даних.

2.2.3 Аудіальні джерела медіаданих

Аудіальні дані (мовлення та звукові сигнали) є унікальним джерелом для аналізу емоційного стану, оскільки вони містять інформацію не лише про семантичний зміст (що сказано), але й про просодичні характеристики (як сказано). Тон, тембр, гучність, швидкість мовлення та паузи часто передають справжній емоційний стан точніше, ніж слова, особливо у випадках сарказму або

прихованої агресії.

Основними джерелами аудіоданих для навчання та тестування систем розпізнавання емоцій є:

– записи розмов у кол-центрах та службах підтримки. Це одне з найбільш прагматичних та комерційно цінних джерел «живих» емоцій. Такі записи містять реальні діалоги між клієнтами та операторами, насичені широким спектром станів: від спокою та вдячності до роздратування, гніву та фрустрації. Особливістю цих даних є їхня природність (спонтанність) та наявність чіткого контексту (вирішення проблеми). Аналіз таких аудіопотоків дозволяє навчати моделі виявляти стрес та ескалацію конфлікту на ранніх стадіях, що є критичним для систем автоматизованого контролю якості обслуговування;

– голосові повідомлення у месенджерах. Зі зростанням популярності асинхронної комунікації у Telegram, WhatsApp та Viber, голосові повідомлення стали масовим джерелом аудіоданих. На відміну від телефонних дзвінків, такі повідомлення часто записуються в більш розслабленій, інтимній атмосфері, що робить їх джерелом аутентичних, непідробних емоцій. Вони багаті на невербальні вокалізації – зітхання, сміх, плач, довгі паузи для роздумів, – які є потужними маркерами емоційного стану. Крім того, такі записи часто містять фоновий шум (вулиця, транспорт), що дозволяє тренувати стійкість моделей до реальних акустичних умов;

– подкасти, радіоефіри та інтерв’ю. Це джерело надає доступ до довготривалих записів мовлення, що дозволяє аналізувати динаміку зміни настрою спікера протягом тривалого часу. Подкасти часто характеризуються емоційною експресивністю ведучих та гостей, варіюючись від серйозних аналітичних дискусій до гумористичних шоу. Різноманітність голосів, акцентів та манер мовлення у публічних аудіозаписах робить їх цінним матеріалом для навчання моделей, здатних генералізувати емоції незалежно від індивідуальних особливостей спікера (Speaker Independent SER);

– взаємодія з голосовими асистентами. Запити до Siri, Google Assistant, Alexa або автомобільних систем управління генерують специфічний тип

емоційних даних. Користувачі часто змінюють манеру мовлення при зверненні до «машини» (так звана «комп’ютерно-спрямована мова»), роблячи її більш чіткою та артикульованою. Однак у випадках помилок розпізнавання або нерозуміння команди, у голосі користувача різко проявляються специфічні емоції: роздратування, нетерпіння або гнів. Цей тип даних є критично важливим для покращення адаптивності AI-систем;

– актоване мовлення (фільми, аудіокниги, театральні постановки). Як і у випадку з візуальними даними, професійні аудіозаписи слугують еталоном «чистих» емоцій. Актори навмисно гіперболізують інтонаційні патерни для передачі суму, радості чи страху, що дозволяє створювати високоякісні базові датасети з чітким розділенням класів. Аудіокниги, зокрема, є джерелом емоційно забарвленого мовлення з високою якістю запису та відсутністю шумів, що ідеально підходить для початкового тренування нейронних мереж перед переходом до більш складних «зашумлених» даних.

2.2.4 Мультимодальні джерела медіаданих

Сучасні цифрові платформи часто генерують мультимодальні дані, які поєднують кілька типів медіа в одному контенті. Наприклад, пост у соціальній мережі може містити одночасно текст, зображення, відео та аудіо, створюючи багатовимірний емоційний сигнал. Аналіз мультимодальних даних вимагає інтеграції різних типів обробки та моделей, які здатні об’єднувати інформацію з різних модальностей для більш точного визначення емоційного стану.

Мультимодальні дані також включають контекстуальну інформацію, таку як метадані про час створення контенту, геолокацію, пристрій, через який було створено контент, та інші параметри, які можуть впливати на інтерпретацію емоційного стану. Наприклад, час публікації може вказувати на нічну активність, що може корелювати з певними емоційними станами, а геолокація може надавати контекст про події, які впливають на емоційний стан користувача.

2.2.5 Особливості медіаданих для аналізу емоцій

Різні типи медіаданих мають унікальні характеристики, які впливають на їх придатність для аналізу емоцій та вибір методів обробки. Текстові дані є структурованими та легко доступними для автоматичної обробки, але можуть містити неоднозначності через залежність від контексту, культурних особливостей та індивідуальних стилів спілкування. Візуальні дані надають безпосередню емоційну інформацію через міміку та жести, але вимагають значних обчислювальних ресурсів для обробки та можуть бути вразливі до змін освітлення, кута зйомки та інших факторів. Аудіодані містять багату паралінгвістичну інформацію, але їх обробка є складнішою через необхідність перетворення сигналів у формати, придатні для машинного навчання. Мультимодальні дані надають найбільш повну картину емоційного стану, але вимагають складних архітектур та методів інтеграції різних типів інформації. Кожен тип медіаданих має свої переваги та обмеження, і вибір джерел залежить від конкретного завдання аналізу емоційного стану та доступних ресурсів для обробки.

Поєднання розглянутих текстових, візуальних та аудіальних джерел створює підґрунтя для багатовимірного підходу до аналізу емоційного стану, де висновки базуються не лише на очевидних маркерах (ключові слова чи посмішка), а й на прихованих просодичних та контекстуальних чинниках. Саме така компіляція різнорідних медіаданих перетворює сучасні системи класифікації на унікальний інструмент, здатний нівелювати обмеження окремих модальностей і досягати глибшого розуміння людських реакцій, наближаючи штучний інтелект до реального сприйняття світу.

У результаті якісні та різноманітні навчальні вибірки стають фундаментом для створення більш досконалих прогностичних моделей. Поєднання чистих лабораторних датасетів зі спонтанними даними з реального життя забезпечує моделям необхідну міцність та здатність до генералізації. Завдяки цьому медіадані перетворюються зі звичайного інформаційного ресурсу на стратегічну

основу для побудови емпатичних інтерфейсів, де кожне рішення системи підкріплене фактами, мультимодальним аналізом та розумінням емоційної логіки користувача.

2.3 Методи та алгоритми обробки та підготовки медіаданих для класифікації емоцій

Ефективність роботи алгоритмів машинного та глибокого навчання критично залежить від якості вхідних даних. «Сирі» медіадані (зображення, аудіозаписи, тексти) зазвичай містять значний рівень шуму, мають різний формат та розмірність, що унеможливлює їх безпосередню подачу на вхід нейронних мереж. В цьому розділі розглядаються методи попередньої обробки та вилучення ознак для кожної модальності.

2.3.1 Алгоритм попередньої обробки та аугментації візуальних даних

У задачах комп'ютерного зору, зокрема при класифікації емоцій (Facial Expression Recognition, FER) [20], якість та репрезентативність навчальної вибірки є визначальними факторами ефективності моделі. Оскільки вихідні медіадані характеризуються гетерогенністю форматів та наявністю шумів, розроблено уніфікований конвеєр попередньої обробки.

Перед описом розробленого алгоритму необхідно визначити ключові теоретичні поняття, що лежать в його основі: нормалізацію та аугментацію даних.

Нормалізація – це процедура попередньої обробки вхідних даних, метою якої є приведення значень ознак (у випадку зображень – інтенсивності пікселів) до єдиного діапазону, як правило $[0,1]$ або $[-1,1]$. З математичної точки зору, необхідність нормалізації зумовлена природою алгоритмів оптимізації на основі

градієнтного спуску. «Сирі» значення пікселів у діапазоні $[0, 255]$ призводять до насичення функцій активації та нестабільності градієнтів (exploding gradients), що унеможливлює швидку збіжність моделі. У роботі використовується стандартизація (Z-score normalization), яка центрує дані навколо нуля:

$$x' = \frac{x - \mu}{\sigma}$$

де x – вхідне значення,

μ – середнє значення вибірки,

σ – стандартне відхилення.

Аугментація даних – це методика регуляризації, яка полягає у штучному розширенні навчальної вибірки шляхом створення модифікованих копій вихідних зображень. Цей метод є критично важливим для запобігання перенавчанню у глибоких мережах. Аугментація дозволяє моделювати інваріантність нейромережі до афінних перетворень (поворотів, зміщень), змушуючи модель вивчати семантичні ознаки емоцій, а не запам’ятовувати конкретні піксельні патерни. Математично геометрична аугментація описується матрицею афінного перетворення M:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Загальну структуру розробленого алгоритму підготовки даних для формування датасету наведено на рисунку 2.2.

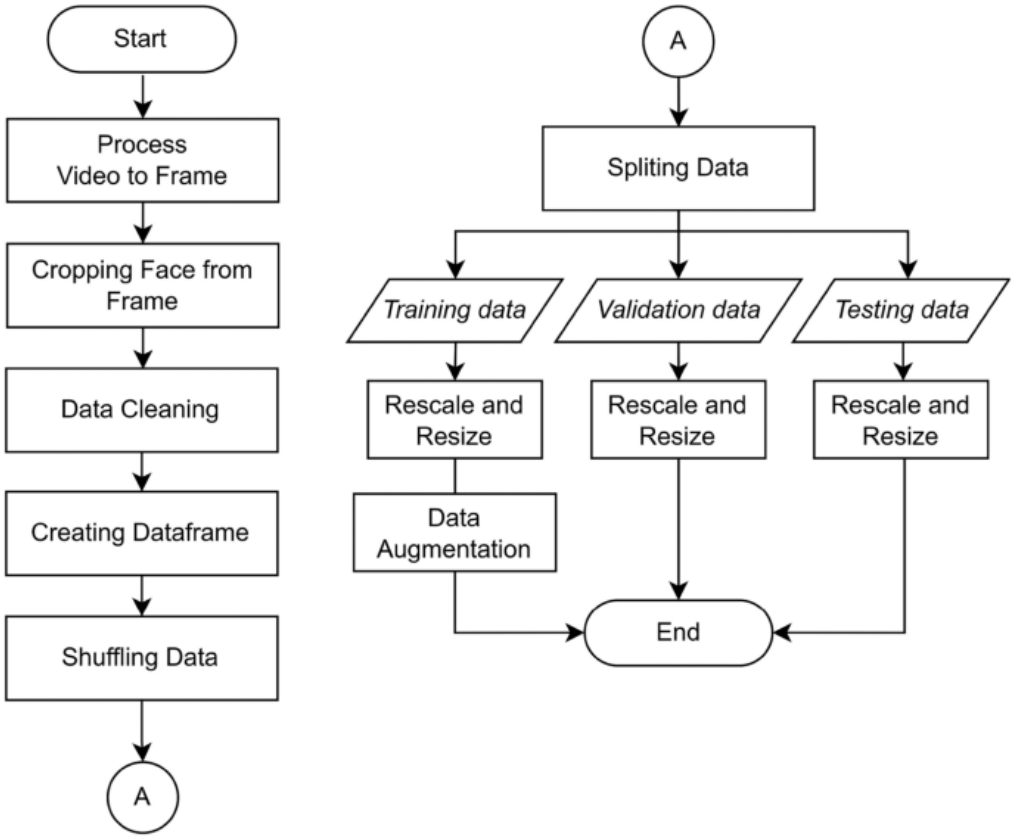

Рисунок 2.2 - Блок-схема алгоритму попередньої обробки відеоданих для формування датасету (англ.)

Алгоритм реалізовано як послідовність етапів трансформації даних, кожен з яких виконує специфічну функцію очищення або збагачення інформації. Розглянемо детально функціональне призначення кожного блоку, зображеного на схемі.

Етап 1. Первинна обробка та вилучення ознак.

– дискретизація відеопотоку (process video to frame). На вхід системи подаються відеофайли, які представляють собою часові послідовності. На цьому етапі відбувається дискретизація неперервного відеопотоку на окремі статичні зображення (фрейми). Частота вибірки кадрів налаштовується таким чином, щоб забезпечити баланс між надлишковістю даних та достатньою варіативністю емоційних станів;

– локалізація та екстракція обличчя (cropping face from frame). Оскільки фон зображення є високоентропійним шумом для задачі класифікації емоцій, критичним кроком є виділення області інтересу (Region of Interest, RoI).

Застосовуються алгоритми виявлення обличчя (наприклад, на базі каскадів Хаара або нейромережевих детекторів MTCNN/RetinaFace), які повертають координати обмежувальної рамки. Зображення обрізається по цих координатах, залишаючи лише інформативну частину – обличчя суб'єкта;

– фільтрація та очищення даних (data cleaning). Цей блок відповідає за забезпечення якості даних. Відбувається автоматична або напівавтоматична перевірка отриманих зображень. Кадри, на яких детектор не спрацював, або зображення з критичними артефактами (розмиття, перекриття обличчя сторонніми об'єктами, екстремальні кути повороту голови), вилучаються з вибірки для запобігання внесенню шумів у навчальний процес;

– структурування метаданих (creating dataframe). Після отримання масиву «чистих» зображень відбувається їх індексація. Створюється структурований табличний об'єкт, де кожному зображенню ставиться у відповідність його шлях у файловій системі та мітка класу емоції. Це дозволяє абстрагуватися від фізичного розміщення файлів на наступних етапах;

– стохастичне перемішування (shuffling data). Відеодані за своєю природою мають сильну часову кореляцію (сусідні кадри майже ідентичні). Подача таких даних у нейромережу може призвести до нестабільності градієнтного спуску та локальних оптимумів. Блок перемішування порушує цей порядок, забезпечуючи незалежність розподілу прикладів у навчальних пакетах.

Етап 2. Формування вибірок та аугментація.

Після переходу через конектор «А» алгоритм фокусується на підготовці тензорів для навчання.

а) стратифіковане розбиття даних (Splitting Data). Загальний масив даних розділяється на три незалежні підмножини для забезпечення об'єктивної валідації моделі: дані для тренування (Trainin data) як основний масив для навчання ваг (зазвичай 70-80%); дані для валідації (Validation data), що використовуються для налаштування гіперпараметрів та ранньої зупинки; дані для тестування (testing data) у вигляді відкладеної вибірки для фінальної перевірки узагальнюючої здатності.

б) геометрична та піксельна нормалізація (rescale and resize). Ця операція виконується паралельно для всіх трьох підмножин:

1) зміна розміру (resize). Зображення масштабуються до фіксованої розмірності вхідного шару нейронної мережі (наприклад, $48 \times 48 \times 1$ для grayscale або $224 \times 224 \times 3$ для RGB);

2) зміна масштабу (rescale). Значення інтенсивності пікселів нормалізуються (приводяться до діапазону $[0,1]$ або стандартизуються через Z-score), що є необхідною умовою для швидкої збіжності алгоритмів оптимізації.

в) синтетична аугментація даних (data augmentation) Це ключовий етап регуляризації, який застосовується виключно до навчальної вибірки (training data). Метою блоку є штучне розширення простору ознак та запобігання перенавчанню (рис. 2.3).

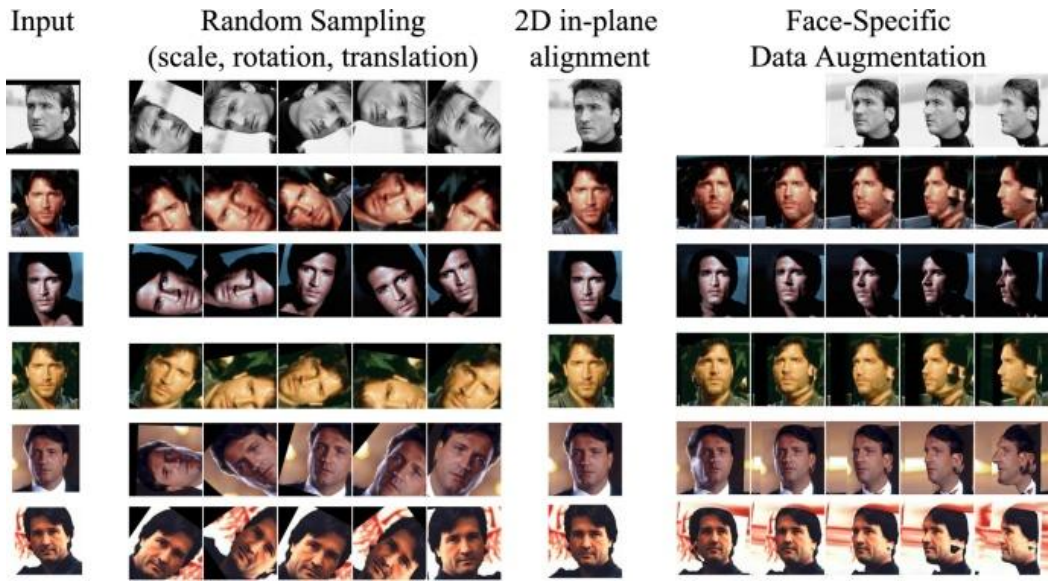

Рисунок 2.3 – Приклади геометричних трансформацій та вирівнювання обличчя при аугментації даних (англ.)

Як продемонстровано на рисунку 2.2, аугментація включає ряд специфічних перетворень:

– геометричні трансформації: випадкові повороти, зсуви по осях X та Y, масштабування. Це емулює варіативність положення голови користувача перед камерою;

– 2D вирівнювання в площині: вирівнювання обличчя відносно лінії очей для зменшення внутрішньокласової дисперсії;

– аугментація конкретних ділянок обличчя: використання складніших методів, наприклад, зміни освітлення або додавання шумів, що дозволяє моделі вивчати інваріантні ознаки емоцій, а не запам’ятовувати конкретні піксельні патерни.

Завершальний блок End символізує формування готових тензорів, готових до подачі на вхід згорткової нейронної мережі.

2.3.2 Методи токенізації та векторного представлення текстової інформації

Текстові дані, отримані з соціальних мереж або транскрибації аудіо, є неструктурованими послідовностями символів. Для їх використання у нейронних мережах необхідно виконати перетворення природної мови (Natural Language Processing, NLP) у числовий векторний простір [21]. Цей процес є багатокроковим і вимагає чіткого визначення базових понять.

Токенізація – це процес сегментації суцільного тексту на елементарні дискретні одиниці (токени), якими можуть виступати слова, підслова або окремі символи. Це фундаментальний етап перетворення «сирих» даних у формат, придатний для обробки машиною.

Векторизація – загальний процес відображення токенів у числові вектори фіксованої довжини.

Ембедінг – це специфічний тип векторизації, що створює щільні вектори у багатовимірному просторі, де семантично близькі слова розташовуються поруч. На відміну від простих ідентифікаторів, ембедінги захоплюють контекстуальне значення слова.

Алгоритм попередньої обробки тексту.

Етап 1. Очищення та нормалізація тексту. Типовий конвеєр (pipeline) попередньої обробки представлений на рисунку 2.4.

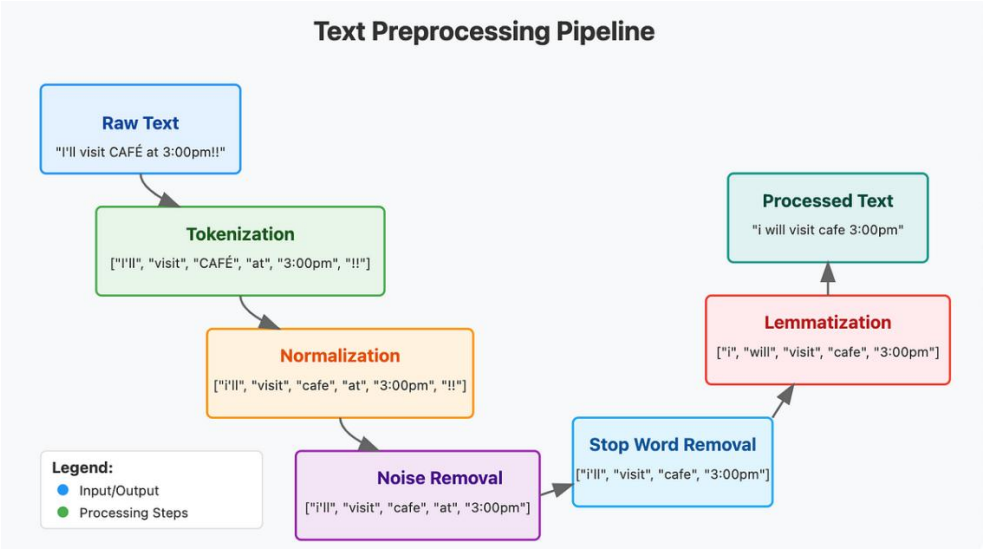

Рисунок 2.4 – Конвеєр попередньої обробки тексту (англ.)

Як показано на схемі, процес трансформації вхідного повідомлення (Raw Text) у готовий для аналізу вигляд включає наступні кроки:

– токенізація (tokenization). Вхідний рядок розбивається на список окремих елементів. На прикладі з рисунка, фраза "I’ll visit CAFÉ..." розкладається на [«I’ll», «visit», «CAFÉ», ...];

– нормалізація (normalization). Приведення всіх символів до нижнього регістру (lowercase). Це необхідно для зменшення розмірності словника, щоб слова «Cafe» і «CAFÉ» сприймалися моделлю як ідентичні;

– видалення шуму (noise removal). Вилучення спеціальних символів, розділових знаків та інших елементів, що не несуть семантичного навантаження в контексті аналізу емоцій (наприклад, видалення «!!»);

– видалення стоп-слів (stop word removal). Фільтрація службових частин мови (артиклів, прийменників, сполучників), які зустрічаються часто, але мають низьку інформативність;

– лематизація (lemmatization). Приведення слів до їхньої словникової (початкової) форми (леми). Наприклад, дієслово майбутнього часу трансформується у базові форми. Це дозволяє об’єднати різні граматичні форми одного слова в один токен.

Результатом цього етапу є Processed Text – очищена послідовність значущих токенів.

Етап 2. Векторизація та побудова моделей вбудовування. Після очищення тексту настає етап перетворення дискретних токенів у неперервні числові представлення. Важливо розуміти відмінність між простою токенізацією та створенням моделей вбудовування: токенізація перетворює дані в унікальні ідентифікатори (ID), тоді як моделі перетворюють ці ідентифікатори в вектори значень, що описують сутність об’єкта.

Архітектуру процесу отримання векторних представлень наведено на рисунку 2.5.

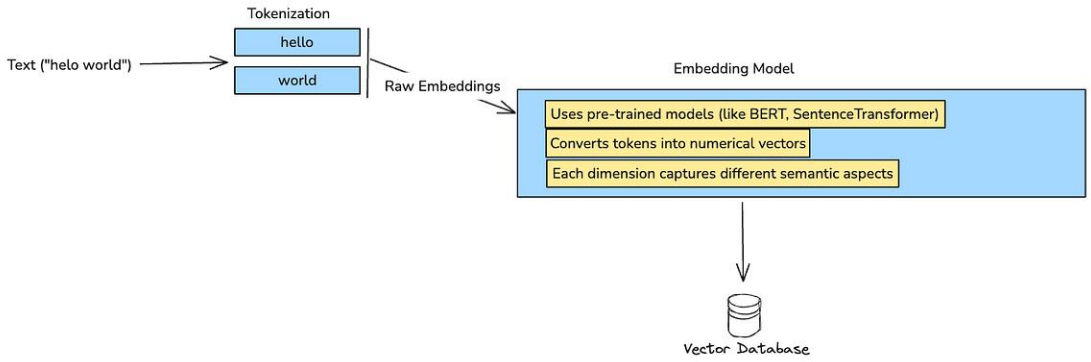

Рисунок 2.5 – Схема процесу токенізації та генерації векторних ембедінгів (англ.)

Процес, відображений на схемі, складається з таких етапів:

– токенізація (tokenization). Очищений текст знову розглядається як послідовність токенів (наприклад, «hello», «world»);

– модель вбудовування (embedding model). Токени подаються на вхід попередньо навченої мовної моделі. У сучасних NLP-задачах стандартом є використання трансформерних архітектур, таких як BERT (Bidirectional Encoder Representations from Transformers) або спеціалізованих моделей SentenceTransformers (див. блок "Embedding Model" на рис. 2.4);

– генерація векторів (vector generation). Модель конвертує токени у числові

вектори. Ключовою особливістю є те, що кожен вимір цього вектора відповідає за певний семантичний аспект слова;

– векторна база даних (vector database). Отримані вектори зберігаються у векторній базі даних для подальшого використання у класифікаторі.

Використання вбудованих моделей дозволяє моделі розуміти синонімію та контекст. Наприклад, у векторному просторі R^d вектори слів «чудовий» та «прекрасний» матимуть високу косинусну подібність (будуть розташовані поруч), що критично важливо для коректної класифікації емоційного забарвлення тексту.

Для демонстрації роботи конвеєра обробки тексту розглянемо реальний відгук користувача про відеогру, який містить складну емоційну конструкцію (змішані почуття та умовний спосіб).

Вхідний текст: "its still fun game to play but only if they fixed the anti-cheat system it would be fantastic".

Розглянемо детально перетворення даних на кожному етапі:

Етап 1. Попередня обробка. Алгоритм виконує очищення та нормалізацію вхідного рядка:

– виправлення помилок (correction): «its» → «it is»;

– токенізація (tokenization): розбиття на семантичні одиниці: [«it», «is», «still», «fun», «game», «to», «play», «but», «only», «if», «they», «fixed», «the», «anti-cheat», «system», «it», «would», «be», «fantastic»];

– лематизація (lemmatization): приведення слів до базової форми («fixed» → «fix»);

– фільтрація стоп-слів: у даному випадку критично важливо не видаляти сполучники «but» та «if», оскільки вони кардинально змінюють емоційний вектор речення. Прості моделі (BoW) часто ігнорують їх, що призводить до помилок, тоді як трансформери (BERT) враховують їх як маркери контексту [22].

Етап 2. Векторизація та механізм уваги (vectorization & attention). Текст перетворюється на ембедінги. Механізм уваги моделі аналізує зв’язки між словами в декілька кроків:

– модель фіксує позитивні маркери: «fun game» (вага: +0.6), «fantastic» (вага: +0.8);

– модель фіксує умовний маркер (contrastive conjunction): «but only if» (інверсія контексту);

– модель ідентифікує джерело негативу (aspect extraction): «anti-cheat system» → «fix» (імплікація: система наразі не працює).

Етап 3: Класифікація (classification). Нейронна мережа агрегує ваги векторів. Хоча у тексті присутні сильні позитивні слова («fun», «fantastic»), конструкція «but only if» сигналізує про невдоволення поточним станом.

Таблиця 2.4 – Результати використання NLP-аналізу емоційного забарвлення

Клас емоції	**Ймовірність**	**Пояснення**
Frustration (Фрустрація)	65%	Домінує через невдоволення системою анти-чіту.
Joy (Радість)	20%	Знижена через умовний характер («було б», а не «є»).
Hope/Anticipation (Надія)	15%	Очікування виправлення проблеми.

Цей приклад ілюструє важливість використання глибоких мовних моделей (Deep Learning), здатних розуміти контекст, на відміну від простих словникових методів, які класифікували б цей відгук як «Позитивний» через наявність слів «fun» та «fantastic».

2.3.3 Методи попередньої обробки та спектральної параметризації аудіосигналів

Фізична природа аудіосигналу як одновимірного часового ряду створює обмеження для алгоритмів машинного навчання. Висока варіативність та зашумленість часового представлення ускладнюють пошук стійких патернів, що

корелюють з емоційним станом мовця. Для подолання цієї проблеми та ефективної екстракції ознак необхідно виконати перетворення сигналу з часового простору у частотний, що дозволяє отримати компактний та інформативний опис акустичних характеристик мовлення. Визначимо поняття.

Спектрограма – це двовимірне візуальне представлення спектру частот сигналу, що змінюється у часі. Математично вона є квадратом амплітуди короткочасного перетворення Фур'є (STFT). На спектрограмі вісь *X* відповідає часу, вісь *Y* – частоті, а інтенсивність кольору – амплітуді звуку.

MFCC (Mel-Frequency Cepstral Coefficients) – це набір коефіцієнтів, що формують «короткостроковий спектр потужності» звуку [23]. Особливістю MFCC є використання шкали Мел, яка апроксимує психоакустичні особливості людського слуху (людина краще розрізняє зміни низьких частот, ніж високих).

Архітектура процесу вилучення ознак.

Загальну схему класифікації емоцій на основі спектральних ознак наведено на рисунку 2.6.

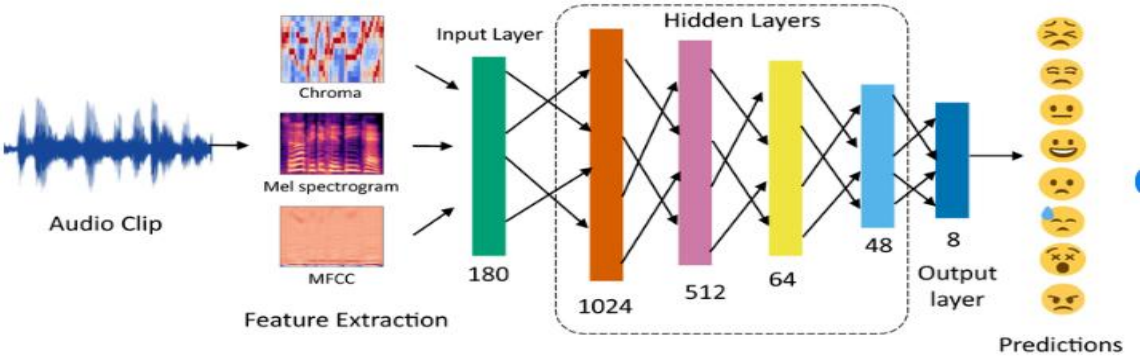

Рисунок 2.6 – Архітектура нейронної мережі для класифікації емоцій з використанням різних типів спектральних ознак (англ.)

Як показано на схемі, процес починається з вхідного аудіокліпу (Audio Clip). На етапі Feature Extraction система генерує три типи представлень:

– Chroma – представлення, що відображає гармонічний зміст сигналу (розподіл енергії за 12 тонами музичної октави);

– Mel-spectrogram – спектрограма, частотна вісь якої трансформована за

шкалою Мел;

– MFCC – компактний вектор кепстральних коефіцієнтів (зазвичай 13-40 коефіцієнтів).

Отримані ознаки подаються на вхід повнозв'язної нейронної мережі (DNN). У наведеному прикладі вхідний шар має розмірність 180 нейронів, за яким слідують приховані шари зі зменшенням розмірності (1024 → 512 → 64 → 48), що виконують роль класифікатора. Вихідний шар генерує розподіл ймовірностей для 8 класів емоцій.

Алгоритмічна реалізація отримання Mel-спектрограми.

Процес перетворення «сирого» сигналу у Mel-спектрограму є складною послідовністю операцій цифрової обробки сигналів (DSP). Деталізований алгоритм цього перетворення зображено на рисунку 2.7.

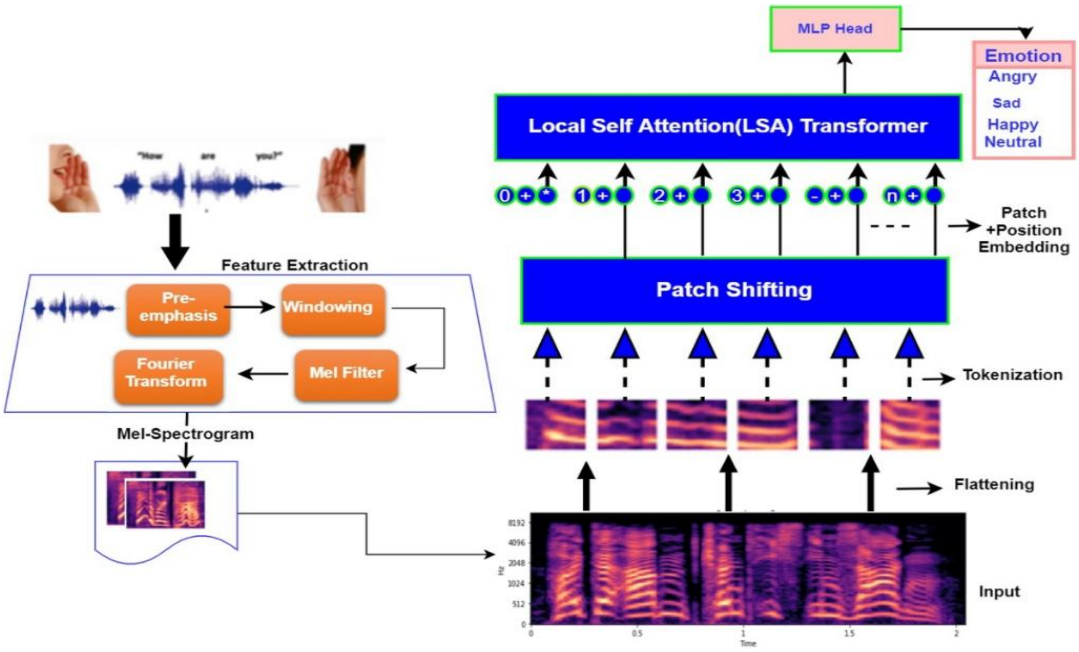

Рисунок 2.7 – Деталізована схема вилучення ознак та токенізації спектрограми для трансформерних моделей (англ.)

Блок Feature Extraction на цій схемі розкриває внутрішню логіку процесу в декілька етапів:

Етап 1. Попереднє підсилення (Pre-emphasis) – проходження сигналу через фільтр високих частот для компенсації природного загасання високочастотних

компонент мовлення.

Етап 2. Віконне перетворення (Windowing) – розбиття сигналу на короткі фрагменти (фрейми) довжиною 20-40 мс із застосуванням віконної функції (наприклад, вікна Хеммінга) для зменшення спектрального витоку на краях фрагментів.

Етап 3. Перетворення Фур'є: застосування швидкого перетворення Фур'є (FFT) для переходу в частотну область.

Етап 4. Mel Filter: застосування банку трикутних фільтрів для перерахунку спектру енергії у шкалу Мел.

Важливою особливістю схеми на рисунку 2.7 є подальша підготовка даних для архітектури Transformer. Отримана Mel-спектрограма не подається цілком, а проходить етап Flattening та розбиття на патчі (Patch Shifting → Tokenization). Спектрограма розрізається на вертикальні смуги (токени), до яких додаються позиційні ембедінги (Position Embedding), що дозволяє механізму уваги (Local Self Attention Transformer) аналізувати часові залежності між різними частинами висловлювання.

Сегментація та просторово-часовий аналіз.

Для аналізу довгих аудіозаписів використовується підхід, що поєднує згорткові (CNN) та рекурентні (RNN/BLSTM) мережі в декілька етапів (рис. 2.8).

Етап 1. Сегментація. Вхідний сигнал (Speech signal) розбивається на сегменти фіксованої довжини (наприклад, 265 мс) з перекриттям.

Етап 2. Генерація спектрограм: Для кожного моменту часу $t = 1, t = 2, \ldots, N$ формується окрема спектрограма, яка розглядається як «зображення» звуку.

Етап 3. Feature Extraction (CNN). Кожна спектрограма проходить через ідентичні блоки згорткової нейронної мережі (CNN), які вилучають просторові ознаки (форманти, різкі зміни енергії).

Етап 4. Часова агрегація (Decision). Вектори ознак (Feature 1...N), отримані від CNN, подаються як послідовність у BLSTM (Bidirectional Long Short-Term Memory). Це дозволяє врахувати емоційну динаміку, де значення поточного фрагмента залежить від попереднього та наступного контексту.

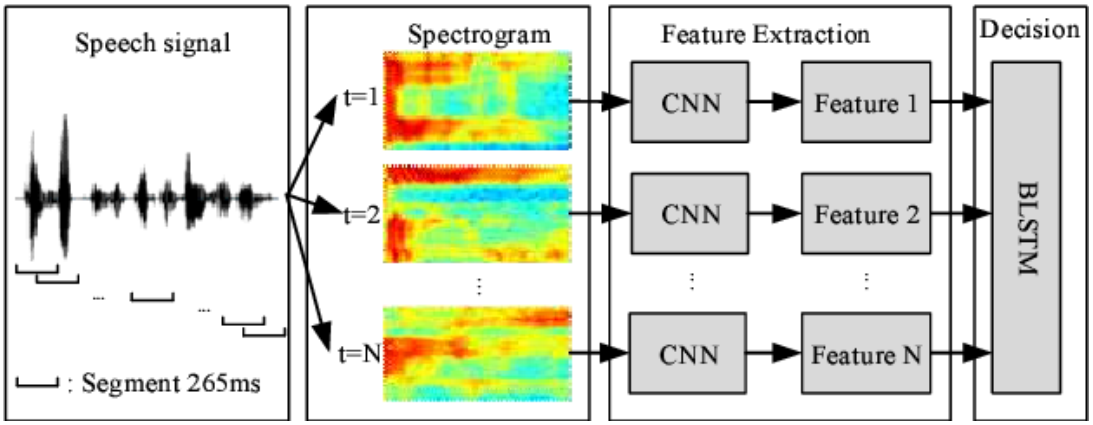

Рисунок 2.8 – Схема обробки сегментованого мовного сигналу гібридною мережею CNN-BLSTM (англ.)

Такий багаторівневий підхід дозволяє досягти найвищої точності класифікації, оскільки використовує переваги як спектрального аналізу, так і глибокого навчання для виявлення складних нелінійних залежностей у мовленні [24].

2.3.4 Використання машинного навчання для класифікації емоцій у медіаданих

Після етапів попередньої обробки та вилучення ознак, описаних у попередніх розділах, отримані векторні представлення (тензори зображень, ембедінги тексту або MFCC-коефіцієнти аудіо) подаються на вхід алгоритмів машинного навчання. Метою цього етапу є побудова відображення $f: X \rightarrow Y$, де X – простір вхідних ознак, а Y – простір цільових міток емоцій [25].

У сучасних системах афективних обчислень домінуючу роль відіграють методи глибокого навчання, здатні автоматично виявляти складні нелінійні залежності у даних. Однак, перш ніж розглядати конкретні архітектури мереж, необхідно визначити постановку задачі класифікації, оскільки вона диктує структуру вихідного шару нейронної мережі та вибір функції втрат.

2.3.4.1 Типологія задач класифікації емоцій

Залежно від складності емоційної моделі та вимог до системи, задачу розпізнавання можна сформулювати трьома основними способами (рис. 2.9).

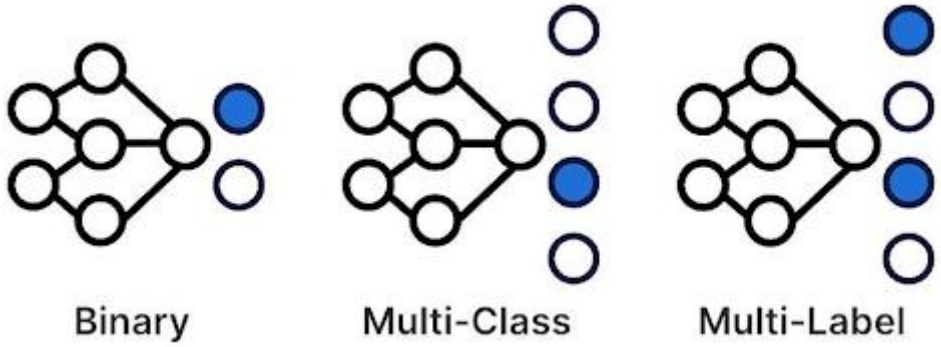

Рисунок 2.9 – Архітектурні відмінності вихідних шарів нейронної мережі для бінарної, багатокласової та багатоміткової класифікації (англ.)

Розглянемо детально кожен із підходів, зображених на схемі:

– бінарна класифікація (Binary). На лівій частині рисунка 2.9 зображено модель, де вихідний шар приймає одне з двох можливих рішень (активний вузол виділено синім кольором). Задача розподілу об’єктів на два класи, які є протилежними один одному. Найчастіше використовується в аналізі тональності тексту для визначення полярності: позитивний чи негативний.

Математична реалізація: вихідний нейрон зазвичай використовує функцію активації Sigmoid, яка повертає ймовірність $p \in [0,1]$ належності до позитивного класу. Якщо $p > 0.5$, об’єкт класифікується як «позитивний»;

– багатокласова класифікація (Multi-Class). Центральна схема на рисунку 2.9 демонструє ситуацію, де з набору можливих класів (чотири вузли) обирається лише один активний (синій вузол). Задача, в якій кожен вхідний зразок може належати тільки до одного класу з N можливих ($N > 2$). Використовується для

розпізнавання дискретних базових емоцій (наприклад, модель Пола Екмана: радість, сум, злість, страх, здивування, огида). Система припускає, що людина в конкретний момент часу не може відчувати «радість» і «сум» одночасно.

Математична реалізація: на вихідному шарі використовується функція активації Softmax. Вона перетворює вихідні значення нейронів (логіти) у розподіл ймовірностей, сума яких дорівнює одиниці ($\sum p_i = 1$). Класом об'єкта вважається той, що має найвищу ймовірність;

– багатоміткова класифікація (Multi-Label). Права частина рисунка 2.9 ілюструє найбільш складний сценарій, де для одного вхідного зразка активовано відразу кілька вихідних вузлів. Задача, в якій об'єкту може бути присвоєно довільну кількість міток класів одночасно (від 0 до N). Критично важлива для розпізнавання складних, змішаних емоційних станів (наприклад, «ностальгія» може бути комбінацією суму та радості). Це дозволяє системі фіксувати нюанси людської психології, які ігноруються у жорсткій багатокласовій моделі.

Математична реалізація: замість softmax використовується функція sigmoid для кожного нейрона вихідного шару окремо. Кожен вихідний вузол діє як незалежний бінарний класифікатор, що вирішує, чи присутня дана емоція в сигналі.

2.3.4.2 Архітектурні підходи до навчання моделей.

У підсумку після визначення типу класифікації обирається базова архітектура нейронної мережі, яка найкраще відповідає типу вхідних медіаданих:

CNN переважно використовуються для просторового аналізу:

– в зображеннях: виділяють ієрархію візуальних патернів – від простих геометричних примітивів (ліній, кутів) на перших шарах до складних семантичних об'єктів (очі, губи, мімічні зморшки) на глибоких шарах;

– в аудіо: застосовуються до спектрограм як до двовимірних зображень,

дозволяючи ідентифікувати візуальні патерни на частотно-часовій карті (наприклад, формантні треки або різкі сплески енергії);

– в тексті: одновимірні згорткові мережі (1D-CNN) ефективні для виявлення локальних n-грам (стійких словосполучень) у послідовності ембедінгів, ігноруючи глобальний синтаксис на користь ключових емоційних маркерів [26].

RNN та їх модифікації (LSTM/GRU) спеціалізуються на секвенційному (часовому) аналізі, обробляючи дані як послідовність, де поточний стан залежить від попередніх. Є незамінними для обробки потокових даних (аудіо, відео, текст), де емоція розгортається динамічно у часі. Приклад: підвищення інтонації в кінці аудіозапису або зміна порядку слів у реченні може кардинально змінити класифікацію з «нейтрального твердження» на «здивування» або «питання», що неможливо відстежити статичними методами.

Трансформери представляють собою сучасний стандарт у більшості задач. Завдяки механізму Self-Attention (самоуваги), вони здатні враховувати глобальний контекст всього повідомлення або відеофрагменту, визначаючи, які частини вхідних даних є найбільш значущими для формування конкретної емоції. Аналізують весь контекст повідомлення або відеофрагменту одночасно (паралельно), визначаючи вагу (значущість) кожної частини вхідних даних для формування конкретної емоції. Це дозволяє враховувати зв’язки між віддаленими елементами, які втрачаються у класичних RNN [26].

2.3.4.3 Навчання та оптимізація

Процес навчання моделі полягає у мінімізації функції втрат, яка оцінює розбіжність між передбаченими ймовірностями та істинними мітками. Отже для бінарної класифікації використовується бінарна крос-ентропія, для багатокласової категоріальна кросс-ентропія, для багатоміткової – сума бінарних крос-ентропій для кожного класу.

Таким чином, вибір методу машинного навчання є не просто вибором алгоритму, а комплексним рішенням, що включає визначення парадигми класифікації, архітектури мережі та стратегії оптимізації, виходячи з природи медіаданих.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ У ВІДЕОПОТОЦІ

3.1 Обґрунтування вибору мови програмування для програмної реалізації

Для розробки системи розпізнавання емоційного стану у відеопотоці критично важливим є вибір інструментарію, що забезпечує баланс між швидкістю розробки та продуктивністю обчислень. Найбільш раціональним вибором для реалізації завдань комп'ютерного зору та глибокого навчання на сьогодні є мова програмування Python.

Вибір Python зумовлений наступними факторами:

– Python є де-факто стандартом у галузі штучного інтелекту, що забезпечує доступ до найсучасніших архітектур нейронних мереж;

– наявність оптимізованих модулів для обробки зображень та матричних обчислень дозволяє реалізувати складні алгоритми (згортку, пулінг, нормалізацію) з мінімальними витратами часу на написання низькорівневого коду;

– інтерпретована природа мови дозволяє запускати розроблене програмне забезпечення на різних операційних системах (Windows, Linux, macOS) без необхідності перекомпіляції, що є важливим для універсальності системи TrackEmotion.

3.1.1 Вибір технологій та бібліотек

Аналіз технічних вимог до системи, зокрема необхідності роботи в режимі реального часу та використання CNN, обумовив вибір наступного технологічного стеку. На основі аналізу залежностей проєкту (requirements.txt), було використано такі бібліотеки:

– OpenCV-Python – фундаментальна бібліотека комп’ютерного зору. У розробленій системі вона використовується для захоплення відеопотоку з веб-камери, попередньої обробки кадрів (зміна кольорової схеми на Grayscale) та виявлення облич за допомогою каскадів Хаара. Це забезпечує виділення обличчя людини перед подачею даних у нейронну мережу;

– TensorFlow – відкрита програмна бібліотека для машинного навчання, розроблена компанією Google, що стала фундаментом для побудови обчислювального ядра системи. У межах даної роботи TensorFlow використовується як бекенд для виконання високонавантажених тензорних операцій. Ключовою причиною вибору цього фреймворку є його здатність ефективно розпаралелювати обчислення при тренуванні глибоких нейронних мереж, а також вбудована підтримка збереження та завантаження навчених моделей у форматі HDF5, що є критичним для забезпечення персистентності розробленого класифікатора емоцій;

– keras – високорівневий API нейронних мереж, що працює поверх TensorFlow. Вибір Keras зумовлений його модульністю та простотою прототипування архітектур глибокого навчання. У програмній реалізації системи саме засобами Keras описано шари згорткової нейронної мережі (Conv2D, MaxPooling2D, Flatten, Dense), налаштовано функцію втрат та оптимізатор. Це дозволило абстрагуватися від низькорівневих математичних перетворень і зосередитися на налаштуванні гіперпараметрів моделі розпізнавання емоцій;

– numpy – бібліотека для високопродуктивних наукових обчислень. Застосовується для представлення зображень у вигляді багатовимірних масивів (тензорів), нормалізації значень пікселів та виконання матричних операцій, необхідних для підготовки вхідних даних для CNN;

– pillow (PIL) – бібліотека обробки зображень, що використовується для додаткових маніпуляцій з графічними даними, конвертації форматів та зміни розмірності зображень відповідно до вимог вхідного шару нейромережі;

– matplotlib – інструмент для візуалізації даних. У межах проєкту

використовується для побудови графіків розподілу емоцій, відображення результатів навчання моделі (кривих точності та втрат) та аналізу статистики сесії;

– scikit-learn – бібліотека для класичного машинного навчання. Використовується для розрахунку метрик якості класифікації, побудови матриці помилок та попередньої обробки даних.

3.1.2 Вибір середовища розробки та інструментів

Середовище розробки на мові Python повинно забезпечувати стабільну роботу з бібліотеками комп'ютерного зору та глибокого навчання, надавати інструменти для зручного налагодження коду в режимі реального часу та ефективного керування залежностями проєкту. Для реалізації програмного комплексу було обрано редактор коду Visual Studio Code та хмарну платформу Google Colab, комбінація яких дозволила розділити задачі написання коду та ресурсоємного навчання нейронних мереж.

Основна розробка програмних модулів здійснювалася у кросплатформенному редакторі коду Visual Studio Code, що характеризується гнучкістю налаштувань та широкою підтримкою розширень для Python. Вибір цього середовища обумовлений необхідністю прямої взаємодії з апаратним забезпеченням комп'ютера, зокрема веб-камерою, що є критичним для коректної роботи бібліотеки OpenCV при захопленні відеопотоку. Visual Studio Code забезпечує зручну роботу з інтегрованим терміналом, що дозволило ефективно керувати віртуальними середовищами, вирішувати конфлікти версій бібліотек TensorFlow і Keras та відстежувати логи виконання програми в реальному часі. Використання інструментів IntelliSense та Pylance прискорило процес написання коду завдяки автодоповненню та статичному аналізу, а вбудовані засоби налагодження дозволили швидко виявляти помилки при завантаженні каскадів Хаара та ваг моделі.

Для задач, пов’язаних із верифікацією та тестуванням попередньо навченої згорткової нейронної мережі, використовувалося хмарне середовище Google Colab. Доступ до обчислювальних ресурсів цієї платформи дозволив виконати швидку перевірку коректності завантаження ваг моделі та проаналізувати її архітектуру без необхідності складного налаштування локального середовища CUDA на початкових етапах. Google Colab також застосовувався для візуалізації структури мережі та тестового прогону алгоритму класифікації на статичних зображеннях, що слугувало етапом валідації перед перенесенням рішення в основну програму для роботи з відеопотоком. Для забезпечення відтворюваності проєкту та ізоляції залежностей використовувалося віртуальне середовище venv із фіксацією точних версій пакетів у файлі requirements.txt. Контроль версій та резервне копіювання вихідного коду здійснювалися за допомогою системи Git та репозиторію GitHub, що забезпечило цілісність розробки та можливість швидкого розгортання системи на інших робочих станціях.

3.2 Архітектура та програмна реалізація системи

Алгоритм роботи розробленої системи передбачає циклічну обробку вхідного відеопотоку в режимі реального часу з метою виявлення обличчя, класифікації емоційного стану та візуалізації аналітичних даних. Архітектурно рішення побудоване на взаємодії модуля захоплення відео, модуля попередньої обробки кадрів та нейромережевого класифікатора.

3.2.1 Особливості інтерфейсу для взаємодії з користувачем

Інтерфейс користувача системи розпізнавання емоцій реалізовано з використанням бібліотеки OpenCV, що забезпечує обробку відеопотоку у режимі реального часу та інтерактивну візуалізацію результатів виявлення. Для

організації взаємодії користувача з модулями програми розроблено клас UIManager, який управляє елементами керування. Центральним компонентом системи є клас EmotionDetector, який координує основні етапи конвеєру, а саме: захоплення відеокадру з камери, виявлення осіб за допомогою каскадного класифікатора Хаара [27], класифікацію емоцій за допомогою навченої нейронної мережі, збереження результатів та запис сесії (рис. 3.1). Таким чином, OpenCV виступає основним посередником між користувачем і програмною логікою, забезпечуючи оптимальну продуктивність для обчислень у режимі реального часу.

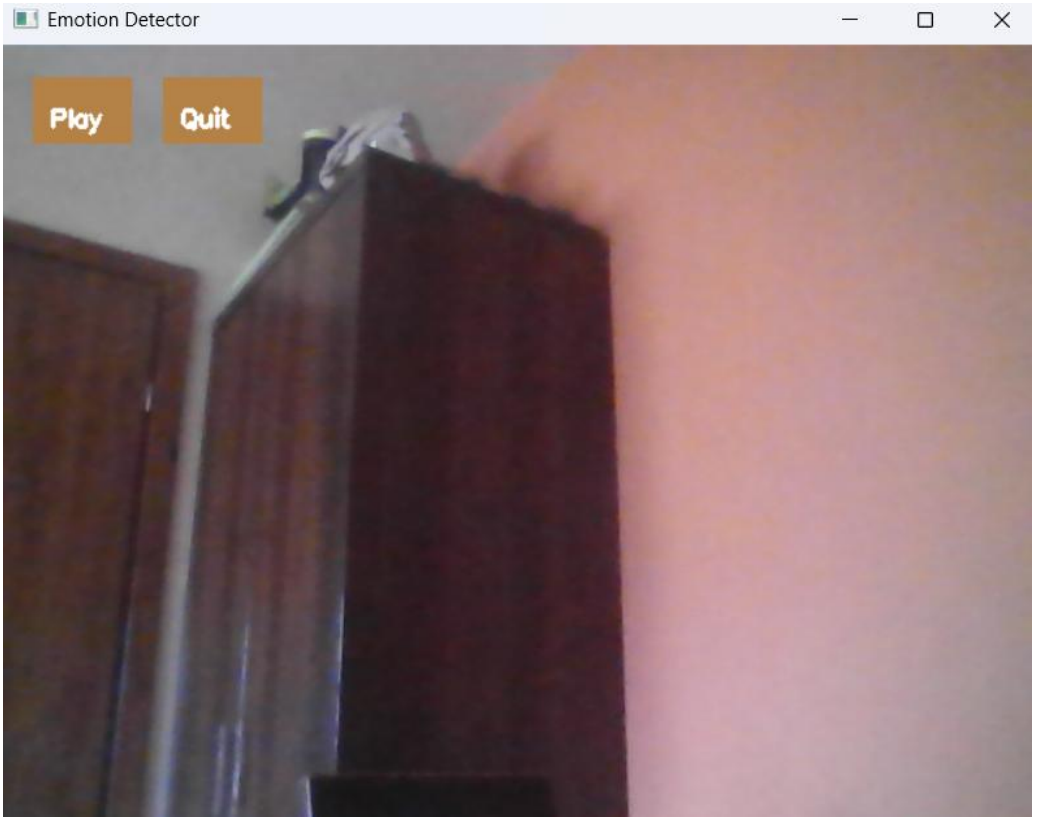

Рисунок 3.1 – Головне вікно програми із увімкненою вебкамерою

Інтерфейс містить мінімалістичні елементи керування, розміщені у верхній частині вікна, що забезпечують інтуїтивну взаємодію з системою. Користувач має можливість:

– управління процесом виявлення за допомогою кнопки «Play» (початок/припинення обробки);

– завершення роботи програми за допомогою кнопки «Quit» (закриття додатку);

– спостереження за ефектами виявлення в режимі реального часу на головному екрані.

Для виявлення осіб використовується каскадний класифікатор Хаара, який забезпечує швидкий і надійний пошук облич у видеопотоці. Система налаштована з наступними параметрами оптимізації (налаштування у src/config.py файлі):

CASCADE_SCALE_FACTOR = 1.3 – коефіцієнт масштабування для піраміди зображень (нижче значення = знаходить більше облич, вище = суворіше критерії);

CASCADE_MIN_NEIGHBORS = 5 – мінімальна кількість сусідніх прямокутників для валідного виявлення (вище значення = суворіша фільтрація помилок);

MIN_FACE_SIZE = 0 – мінімальний розмір обличчя в пікселях (0 = без обмежень, тобто оригінальна поведінка).

Для розпізнавання емоцій використовується вже навчена згорткова нейронна мережа, збережена у форматі HDF5. Мережа класифікує виділене обличчя на одну з семи категорій: гнів, відразу, страх, радість, нейтрально, сум, здивування. Система надає фільтрацію результатів за порогом впевненості:

CONFIDENCE_THRESHOLD = 0.0 – мінімальна впевненість класифікатора для прийняття результату (0% = приймаються всі передбачення).

Основні компоненти інтерфейсу включають:

а) елементи керування (UI контролери):

1) «Play» кнопка розташована в позиції (20, 20) розміром 60×40 пікселів; активна кнопка відображається світло-блакитним кольором (135, 206, 250), неактивна – сталевим синім (70, 130, 180);

2) «Quit» кнопка розташована в позиції (100, 20) розміром 60×40 пікселів; дизайн аналогічний «Play» кнопці.

б) візуалізація результатів:

1) прямокутники облич – жовтого кольору (0, 255, 255) з товщиною лінії 2 пікселі, малюються навколо кожного виявленого обличчя для наочної вказівки на область обробки;

2) текстові мітки – над кожним обличчям відображається найменування емоції та відсоток впевненості (наприклад, "Happy (95%)");

3) лічильник емоцій – у нижній частині вікна показується накопичений підрахунок кожної категорії емоції за сесію.

Виявлені кадри автоматично зберігаються в структурованих папках за типами емоцій (results/angry/, results/happy/, results/sad/ тощо) у форматі JPG з мітками часу для подальшого аналізу та валідації результатів. Крім того, вся сесія записується як відеофайл у форматі MP4 з частотою 10 кадрів за секунду, що дозволяє користувачеві переглянути повний перебіг виявлення.

Система містить фоновий потік обробки для збереження зображень, що запобігає блокуванню основного потоку виявлення. Це забезпечує плавну роботу програми без затримок, оскільки операції вводу-виводу виконуються паралельно. Таке архітектурне рішення гарантує коректне завершення роботи – остання виявлена емоція автоматично зберігається перед закриттям програми.

Користувач може здійснювати взаємодію з програмою, використовуючи миш (клік по кнопкам) або клавіатуру (клавіша ‘q’ для виходу). Усі операції виконуються локально на комп’ютері користувача, що гарантує конфіденційність медіаданих та не потребує інтернет-з’єднання. Загальна архітектура інтерфейсу спрямована на мінімізацію когнітивного навантаження користувача, дозволяючи йому сфокусуватися на якості виявлення емоцій.

Усі сесії виявлення емоцій автоматично записуються у вигляді відеофайлів. Система організована таким чином, щоб забезпечити структурованість і зручність управління даними. Ці відеозаписи зберігаються в окремій папці «session_streams/», розташованій в кореневій директорії проекту. Це архітектурне рішення дозволяє:

– відділити відеоматеріали від статичних зображень емоцій;

– відстежити всі проведені сесії за часовими мітками;

– організувати довгострокове зберігання даних дослідження;

– уникнути перемішування даних різних типів.

Параметри відеозапису. Відеозаписи створюються з наступними технічними параметрами:

– кодек: MPEG-4 – забезпечує сумісність з більшістю медіаплеєрів;

– частота кадрів (FPS): 10 кадрів за секунду – оптимальна для аналізу емоцій без надмірної ваги файлу;

– розширення файлу: MP4 – відкритий стандарт контейнера, підтримується всіма сучасними операційними системами;

– мітка часу: кожний файл має унікальне найменування формату session_DDMmmYYYY_HHMMSS.mp4 (наприклад, session_17Dec2025_122226.mp4), що відображає точний час та дату початку сесії.

Наявність заданої архітектури проєкту дозволяє зручно орієнтуватись в папках з результатами роботи програми.

3.2.2 Алгоритм обробки емоційних даних

В основі реалізації системи розпізнавання емоцій лежить поетапна обробка медіаданих за допомогою бібліотек комп’ютерного зору, алгоритмів виявлення облич та глибоких нейронних мереж для класифікації емоційних станів. Кожен етап реалізовано у вигляді окремого модуля, який обробляє проміжні результати та передає їх наступному компоненту. Така модульна архітектура забезпечує гнучкість системи й дозволяє змінювати параметри без необхідності редагування основного коду. Модульна структура проекту включає: DataGenerator – для завантаження та аугментації даних; CNNModel – для побудови та навчання архітектури; ModelEvaluator – для оцінки якості та аналізу результатів; EmotionDetector – для виявлення у режимі реального часу та класифікації; та UIManager – для управління інтерфейсом користувача. (рис. 3.2).

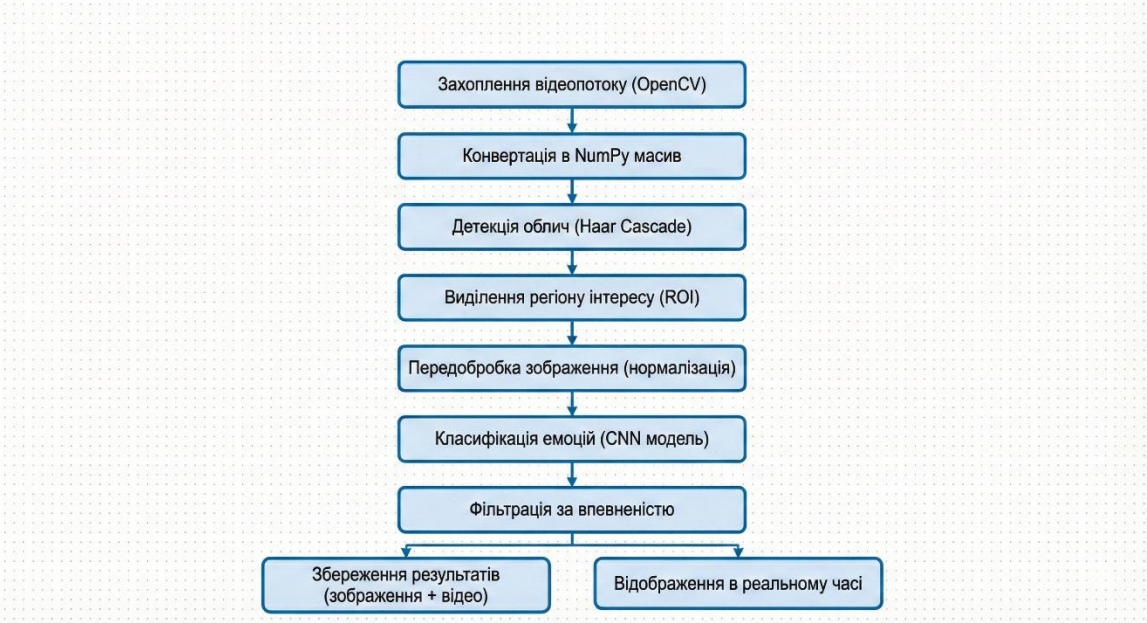

Рисунок 3.2 – Схема алгоритму розпізнавання емоцій у реальному часі

Розберемо поетапно роботу реалізації системи.

Етап 1. Захоплення та конвертація даних. На початковому етапі система захоплює відеопотік з вебкамери за допомогою бібліотеки OpenCV. Кожен кадр отримується функцією VideoCapture та автоматично конвертується у NumPy-масив, придатний для подальшої обробки. Здійснюється операція ображення кадру (cv2.flip) для отримання дзеркального зображення, що відповідає природній уяві користувача. Параметри захоплення налаштовано таким чином: CAMERA_INDEX = 0 для первинної вебкамери пристрою, розширення кадру 640 × 480 пікселів за замовчуванням, формат BGR (Blue-Green-Red) для сумісності з OpenCV.

Етап 2. Детекція (виявлення) облич. Виявлення облич виконується за допомогою каскадного класифікатора Haar Cascade (haarcascade_frontalface_default.xml). Спочатку кадр конвертується в градацію сірого (grayscale), що забезпечує оптимальну роботу класифікатора та зменшує обчислювальні витрати. Система використовує параметр CASCADE_SCALE_FACTOR = 1.3 як коефіцієнт масштабування для піраміди зображень, де менше значення знаходить більше облич, але з більшою кількістю

помилок. CASCADE_MIN_NEIGHBORS = 5 служить мінімальною кількістю сусідніх детекцій для валідації, де вище значення означає суворішу фільтрацію. MIN_FACE_SIZE = 0 дозволяє виявляти обличчя без обмежень за розміром. Результатом етапу є координати прямокутників, що обмежують виявлені обличчя в форматі у ширині та висоті.

Етап 3. Передобробка облич. Для кожного виявленого обличчя виділяється регіон інтересу з зображенням в градаціях сірого. Далі здійснюється зміна розміру всіх регіонів до уніфікованого розміру 48×48 пікселів, відповідно до архітектури навченої моделі. Нормалізація пікселів здійснюється діленням значень на 255 для отримання діапазону [0, 1], що прискорює конвергенцію мережі. Дані переводяться у формат (1, 48, 48, 1) для сумісності з модельним входом, що забезпечує правильну розмірність для обробки.

Етап 4. Класифікація емоцій за допомогою CNN. Оброблені зображення подаються на вхід навченої згорткової нейронної мережі, архітектура якої складається з шести конволюційних блоків. Перший блок містить Conv2D із 32 фільтрами, BatchNormalization, ReLU активацію, MaxPooling(2×2) та Dropout(0.2). Другий блок розширює до 64 фільтрів з аналогічною структурою, третій – до 128 фільтрів, четвертий – до 256 фільтрів, п’ятий – до 512 фільтрів, і шостий – до 1024 фільтрів. Після конволюційних блоків здійснюється розгортання з наступними повносвязними шарами: Dense(512) із BatchNormalization, ReLU та Dropout(0.2), потім Dense(256) з аналогічною структурою, і нарешті вихідний шар Dense(7) з Softmax активацією для семи категорій емоцій. Модель скомпільована з оптимізатором Adam та швидкістю навчання LEARNING_RATE = 0.001 із впровадженням адаптивного зменшення швидкості навчання (ReduceLROnPlateau) з параметрами LR_DECAY_FACTOR = 0.2 як коефіцієнт множника при зменшенні, LR_DECAY_PATIENCE = 6 епох без поліпшення перед зменшенням, та MIN_LEARNING_RATE = 0.00001 як мінімальна допустима швидкість.

Етап 5. Фільтрація за впевненістю. Модель повертає вектор ймовірностей для всіх семи категорій емоцій. Система обирає клас з найвищою ймовірністю та

порівнює її з порогом впевненості CONFIDENCE_THRESHOLD = 0.0, який означає 0% мінімальної впевненості для прийняття результату, тобто приймаються всі передбачення без фільтрації. При потребі цей параметр може бути збільшений для фільтрації низькоякісних передбачень та підвищення точності виявлення.

Етап 6. Збереження результатів. Система автоматично зберігає розпізнані емоції у структурованому форматі. Кожне виявлене обличчя з мітками емоцій зберігається у відповідну папку (results/angry/, results/happy/ тощо) у форматі JPG з якісним стисненням. Назва файлу формується як {емоція}{дата}{час}.jpg (наприклад, angry_17Dec2025_120703.jpg), а зберігання кадрів із виявленою емоцією відбувається кожну секунду відповідно до параметра PICTURE_INTERVAL = 1. Вся сесія записується як відеофайл у форматі MP4 з кодеком MPEG-4 у вигляді 10 кадрів за секунду (VIDEO_FPS = 10) для оптимізації розміру файлу. Всі записи зберігаються в окремій папці SESSION_STREAMS_DIR. Для уникнення блокування основного циклу виявлення використовується фоновий потік (background worker thread), який обробляє чергу операцій збереження, що забезпечує гладку роботу програми без затримок, оскільки операції вводу-виводу виконуються паралельно з розпізнаванням емоції.

Етап 7. Відображення результатів. Результати виводяться на екран у реальному часі. Прямокутники облич малюються жовтим кольором (0, 255, 255) з товщиною 2 пікселі, над кожним обличчям показується емоція та відсоток впевненості. У нижній частині вікна накопичується лічильник кількості виявлених кожної категорії емоції протягом сесії (рис. 3.3) [28-30].

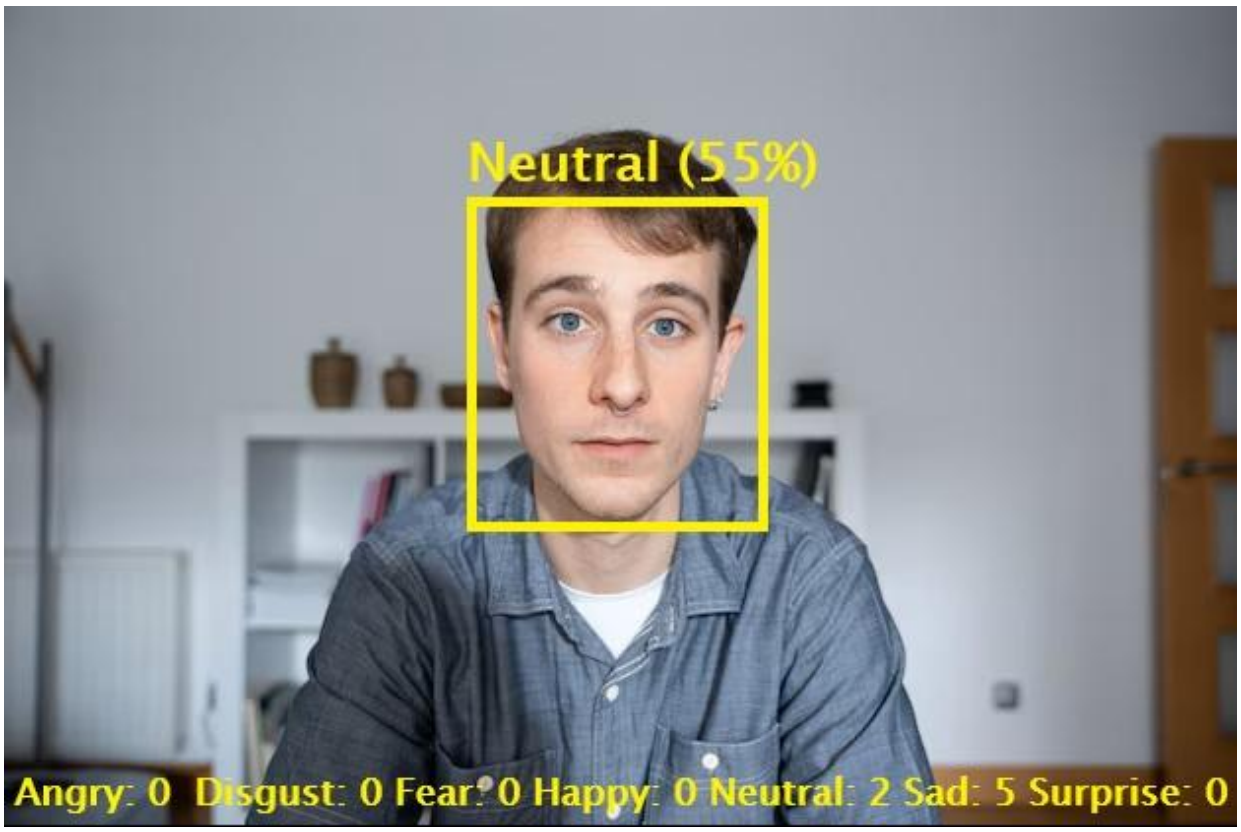

Рисунок 3.3 – Збережене зображення в ході роботи системи виявлення емоцій (англ.)

Користувач може управляти процесом за допомогою кнопки «Play» для активації/деактивації виявлення та кнопки «Quit» для закриття програми. Всі операції виконуються локально на пристрої користувача, що гарантує конфіденційність медіаданих та забезпечує відсутність залежності від інтернет-з'єднання. Виявлення облич виконується на CPU, оскільки каскад Хаара працює оптимально на центральному процесорі та не вимагає графічного прискорення. Модель CNN потенційно може бути оптимізована для запуску на GPU за допомогою CUDA, що прискорило б передбачення у 10-100 разів залежно від апаратури.

3.3 Навчання моделі

Основою аналітичного модуля системи є глибока згорткова нейронна

мережа, завданням якої є класифікація вхідних зображень облич за сімома категоріями емоцій: гнів, відраза, страх, радість, нейтрально, сум та здивування. На відміну від детекції облич, яка виконується каскадом Хаара, класифікація емоцій реалізована через емпіричне навчання на великому масиві даних.

Для навчання моделі використано датасет, розділений на три частини: навчальну, валідаційну та тестову. Структура даних організована у директоріях archive/images/1.train, archive/images/2.validation та archive/images/3.test, де кожна піддиректорія відповідає окремому класу емоцій. Вхідні дані представлені уніфікованими монохромними зображеннями розміром 48×48 пікселів. Перед подачею в мережу застосовується нормалізація (приведення значень пікселів до діапазону $[\mathbf{0}, \mathbf{1}]$) та аугментація навчальної вибірки (геометричні трансформації: зміщення, ротація, масштабування) для підвищення стійкості моделі до варіативності вхідних даних.

Архітектура CNN складається з шести послідовних конволюційних блоків із прогресуючим збільшенням кількості фільтрів (від 32 до 1024). Кожен блок включає шари згортки (Conv2D), нормалізації (BatchNormalization), активації (ReLU), субдискретизації (MaxPooling) та регуляризації (Dropout 0.2). Класифікаційна частина мережі реалізована через повнозв’язні шари (Dense) з вихідним шаром Softmax, що формує розподіл ймовірностей для 7 класів.

Процес навчання керується оптимізатором Adam (Learning Rate = 0.001) з функцією втрат категоріальної кросс-ентропії. Для адаптивного налаштування параметрів використовуються callback-функції: ReduceLROnPlateau для динамічного зменшення швидкості навчання при стагнації та ModelCheckpoint, який автоматично зберігає найкращу версію ваг моделі у файл model_weights.h5.

Оцінка якості навчання здійснюється модулем ModelEvaluator на тестовій вибірці з розрахунком метрик Accuracy, Precision, Recall та F1-score. Фінальний файл model_weights.h5 містить оптимізовані параметри мережі у форматі HDF5 і завантажується класом EmotionDetector під час ініціалізації системи для роботи в режимі реального часу.

ВИСНОВКИ

Таким чином, у кваліфікаційній роботі досліджено методи класифікації емоцій у медіаданих для аналізу емоційного стану користувачів та вирішено такі завдання: проведено аналіз літературних джерел та сучасних підходів у сфері афективних обчислень, що дало можливість визначити стан дослідженої проблематики, а також переваги та обмеження методів комп'ютерного зору при розпізнаванні міміки; проведено аналіз архітектур нейронних мереж (CNN, RNN, Transformers), що дало змогу детально вивчити їх сильні сторони для обробки статичних зображень та відеопотоків.

Сформовано цілісний процес попередньої обробки візуальних даних, що включає детекцію облич за допомогою каскада Хаара, конвертацію у відтінки сірого, нормалізацію пікселів та геометричну аугментацію, що дало можливість побудувати узгоджений датасет для якісного навчання моделі. Побудовано та реалізовано архітектуру глибокої згорткової нейронної мережі з шести блоків із використанням шарів Dropout та BatchNormalization, візуалізовано структуру процесу навчання, що дозволило наочно продемонструвати всі етапи: підготовку батчів, оптимізацію ваг алгоритмом Adam, моніторинг функції втрат та збереження найкращої моделі.

Разроблено програмну систему мовою Python із використанням бібліотек OpenCV та TensorFlow/Keras, що дало можливість створити повнофункціональний застосунок для детекції емоцій у режимі реального часу через веб-камеру. Реалізовано механізм візуалізації результатів шляхом накладання графічних примітивів (Bounding Box) та текстових міток на відеоряд, а також модуль побудови графіків динаміки емоційних станів (Valence/Arousal), що дозволило підвищити інформативність системи для кінцевого користувача. Створено модуль оцінки якості моделі із застосуванням метрик Accuracy, Precision, Recall та F1-score, що дало можливість об'єктивно перевірити ефективність класифікації на тестовій вибірці.

Проведено експериментальне дослідження роботи системи, яке показало,

що поєднання класичних методів детекції облич (каскад Хаара) та глибокого навчання забезпечує оптимальний баланс між швидкодією та точністю. Визначено, що розроблена модель демонструє високу стійкість до варіативності вхідних даних, а реалізована архітектура здатна ефективно працювати на локальному обладнанні без затримок у відеопотоці, забезпечуючи конфіденційність даних користувача.

У рамках кваліфікаційної роботи було реалізовано модульну архітектуру системи з класом EmotionDetector, який централізує ініціалізацію обладнання, інференс нейромережі та збереження результатів, що дозволило автоматизувати повний цикл роботи алгоритмів у єдиному середовищі. Візуалізовано ключові процеси у вигляді блок-схем, що дозволило оптимізувати логіку обробки подій та структуру програмного забезпечення.

Наукова новизна роботи полягає у вдосконаленні підходу до побудови систем розпізнавання емоцій шляхом комбінування каскадних класифікаторів для швидкої локалізації об’єктів та глибоких згорткових мереж для семантичного аналізу, що дозволило отримати нові висновки щодо поведінки таких систем у динаміці реального часу. Такий підхід сприяє глибшому розумінню особливостей впровадження технологій штучного інтелекту у прикладні системи моніторингу психоемоційного стану.

У результаті дослідження розроблено працездатний програмний модуль з можливістю запису сесій, генерації аналітичних звітів та роботи з відеопотоком, що дозволило повністю задовольнити мету кваліфікаційної роботи.

Результати роботи апробовано у вигляді 2 тез доповідей під час IX Міжнародної студентської наукової конференції в місті Черкаси від 7 листопада 2025 року [31] та IX Міжнародної студентської наукової конференції в місті Житомир від 14 листопада 2025 року [32].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Noldus (2024). Human behavior application: Emotion analysis. Noldus. URL: https://noldus.com/applications/emotion-analysis (дата звернення 15.10.2025).

2. Li, C. et al. (2023). Emotion recognition of social media users based on deep learning. PMC. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10280477/ (дата звернення 02.11.2025).

3. Gong, Z. et al. (2024). A Mapping on Current Classifying Categories of Emotions. Columbia University. URL: https://www.columbia.edu/2024/emotion_paper.pdf (дата звернення 20.10.2025).

4. Emotion Classification in Short English Texts (2024). ArXiv. URL: https://arxiv.org/html/2402.16034v1 (дата звернення 12.11.2025).

5. Kher, D. et al. (2022). Multi-label Emotion Classification using Machine Learning. SCITEPRESS. URL: https://www.scitepress.org/Papers/2022/115324/115324.pdf (дата звернення 05.10.2025).

6. EEG emotion recognition attention-based (2023). ScienceDirect. URL: https://www.sciencedirect.com/science/article/abs/pii/S1746809423002689 (дата звернення 18.11.2025).

7. Facial Expression Recognition VGG (2024). ACM Digital Library. URL: https://dl.acm.org/doi/10.1145/3627341.3630376 (дата звернення 25.10.2025).

8. Wang, Y., Song, W., Tao, W., et al. (2022). A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. ArXiv preprint arXiv: 2203.06935.

9. Gadetska, S. V., & Gorokhovatsky, V. O. (2018). Statistical measures for computation of the image relevance of visual objects. Telecommunications and Radio Engineering, 77(12), pp. 1041–1053.

10. Gorokhovatskyi, V., Gadetska, S., & Stiahlyk, N. (2023). Accelerating

image classification based on a model for estimating descriptor-to-class distance. International Journal of Computing, 22(4), 485-492.

11. Zigpoll (2024). How Real-Time User Emotion Tracking Can Revolutionize Usability Testing in UX Research. URL: http://www.zigpoll.com (дата звернення 01.12.2025).

12. Gorokhovatskyi, V. O., & Gadetska, S. V. (2020). Statistical processing and data mining in structural image classification methods. Kharkiv: FLP Panov A. N.

13. Deep learning framework for subject-independent emotion detection (2021). PLOS ONE. URL: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0242946 (дата звернення 09.10.2025).

14. Кобилін, О. А., & Творошенко, І. С. (2021). Методи цифрової обробки зображень: навчальний посібник. Харків: ХНУРЕ, 124 с.

15. Imentiv AI (2024). Exploring Emotion Patterns in User Testing with Emotion AI. URL: https://imentiv.ai/blog/ (дата звернення 22.11.2025).

16. Atulapra (2017). Real-time Facial Emotion Detection using deep learning. GitHub repository. URL: https://github.com/atulapra/Emotion-detection (дата звернення 14.10.2025).

17. Pomazan, V., Tvoroshenko, I., & Gorokhovatskyi, V. (2023). Handwritten character recognition models based on CNNs. International Journal of Academic Engineering Research, 7(9), pp. 64–72.

18. Sükei, E. et al. (2021). Predicting Emotional States Using Behavioral Markers. PMC. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8088855/ (дата звернення 30.11.2025).

19. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference pp. 51–56.

20. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

21. Viso AI (2024). AI Emotion Recognition and Sentiment Analysis. URL: https://viso.ai/deep-learning/visual-emotion-ai-recognition/ (дата звернення 11.10.2025).

22. Delve AI (2024). Emotion Analysis: Definition, Models and Use-cases. URL: https://www.delve.ai/blog/emotion-analysis (дата звернення 27.10.2025).

23. Systematic Review of Emotion Detection (2024). Frontiers. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11175284/ (дата звернення 16.11.2025).

24. Bisong, E. (2019). Building ML & DL models on Google Colab. In Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress.

25. Du, Y., et al. (2020). PaddleOCR: An industrial OCR system. arXiv.

26. Lu, X. (2022). Deep Learning Based Emotion Recognition and Visualization. Frontiers in Psychology. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.818833/full (дата звернення 04.11.2025).

27. Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). URL: https://ieeexplore.ieee.org/document/990517 (дата звернення 14.11.2025).

28. Mentionlytics (2024). Emotion Analysis: What it is & How to Do it with AI. URL: https://www.mentionlytics.com/blog/emotion-analysis/ (дата звернення 28.11.2025).

29. Deep learning-based EEG emotion recognition (2023). Frontiers. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1126994/full (дата звернення 03.10.2025).

30. Shehu, H.A. et al. (2025). Emotion categorization from facial expressions. ScienceDirect. URL: https://www.sciencedirect.com/science/article/pii/S0925231225000396 (дата звернення 01.12.2025).

31. Цісаренко, О. І. (2025). Застосування глибоких згорткових нейронних мереж до задачі класифікації емоцій в зображеннях. Збірник наукових праць з матеріалами IX Міжнародної студентської конференції (с. 303–306). Черкаси, Україна.

32. Цісаренко, О. І. (2025). Модернізація та сучасні українські і світові наукові дослідження. Збірник наукових праць з матеріалами IX Міжнародної студентської конференції (с. 309–311). Житомир, Україна.