Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Покращення прозорості процесу прийняття рішень у моделях на базі трансформера в задачах динамічного розпізнавання виразів обличчя в реальних умовах
(тема)

Виконав:
здобувач другого року навчання,
групи СШІм-23-2
Анастасія Кочкіна
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник проф. Кирило Смеляков
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Олег ЗОЛОТУХІН
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

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

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Кочкіній Анастасії Петрівні
(прізвище, ім'я, по батькові)

1. Тема роботи Покращення прозорості процесу прийняття рішень у моделях на базі трансформера в задачах динамічного розпізнавання виразів обличчя в реальних умовах

затверджена наказом університету від 21 квітня 20 25 р. № 295Ст

2. Термін подання студентом роботи до екзаменаційної комісії 3 червня 20 25 р.

3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел та відомих наукових проектів, Python, TensorFlow, PyTorch, MediaPipe, XAI documentation, набір даних MAFW для формування набору даних для тренування і тестування моделі.

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Методи і матеріали
3) Експерименти та результати

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	21.04.2025	виконано
2	Аналіз предметної галузі		виконано
3	Огляд існуючих методів динамічного	26.04.2025	виконано
	розпізнавання виразів облич в реальних умовах		
4	Огляд існуючих методів інтеграції	29.04.2025	виконано
	пояснювального штучного інтелекту		
5	Аналіз підходів до вирішення задачі	01.05.2025	виконано
6	Експериментальне моделювання датасету	05.05.2025	виконано
7	Експериментальне моделювання та навчання	09.05.2025	виконано
	моделі		
8	Написання пояснювальної записки	13.05.2025	виконано
9	Перевірка на академічний плагіат	26.05.2025	виконано
10	Нормоконтроль	27.05.2025	виконано
11	Підготовка презентації та доповіді	28.05.2025	виконано
12	Попередній захист	29.05.2025	виконано
13	Рецензування	30.05.2025	виконано
14	Захист перед ЕК	03.06.2025	

Дата видачі завдання 21 квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Кирило Смеляков
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 68 с., 6 рис., 2 табл., 1 дод., 27 джерел.

ГІБРИДНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ, ГРАФОФІ ТРАНСФОРМЕРИ, ДИНАМІЧНЕ РОЗПІЗНАВАННЯ ВИРАЗІВ ОБЛИЧЧЯ, ПРОЗОРИЙ ШІ, GRAD-CAM, MEDIAPIPE, TEXT TO SPEECH SYSTEMS.

Об’єкт дослідження – системи розпізнавання виразів обличчя в умовах реального світу.

Предмет дослідження – методи покращення точності та пояснюваності моделей розпізнавання емоцій на основі графових і трансформерних архітектур.

Мета роботи – розробити інтерпретовану модель розпізнавання виразів обличчя з використанням Spatio-Temporal Graph Transformer та інтегрувати її з системою емоційно-залежного синтезу мовлення Llasa.

Методи дослідження – методи комп’ютерного зору, графових нейронних мереж, трансформерних моделей, а також підходи пояснювального ШІ для візуалізації важливих ознак і рішень моделі.

У роботі розроблено модель Spatio-Temporal Graph Transformer для розпізнавання емоцій у відео на основі орієнтирів обличчя. Реалізовано обробку просторово-часових даних та інтегровано XAI-методи (Grad-CAM, attention attribution), що підвищують інтерпретованість. Проведено експерименти на послідовностях різної довжини. Результати демонструють високу точність і прозорість моделі.

ABSTRACT

Master’s thesis contains: 68 pp., 6 fig., 2 tabl., 1 ann., 27 references.

DYNAMIC FACIAL EXPRESSION RECOGNITION, GRAD-CAM, GRAPH TRANSFORMERS, HYBRID INTELLIGENT SYSTEMS, MEDIAPIPE, TEXT TO SPEECH SYSTEMS, TRANSPARENT AI.

Object of research – facial expression recognition systems in real-world conditions.

Subject of research – methods for improving the accuracy and explainability of emotion recognition models based on graph and transformer architectures.

Purpose of research – develop an interpreted facial expression recognition model using Spatio-Temporal Graph Transformer and integrate it with the Llasa emotion-dependent speech synthesis system.

Methods of research – computer vision methods, graph neural networks, transformer models, as well as explanatory AI approaches for visualizing important features and model solutions.

The paper develops a Spatio-Temporal Graph Transformer model for emotion recognition in video based on facial landmarks. Spatio-temporal data processing is implemented and XAI methods (Grad-CAM, attention attribution) are integrated, which increase interpretability. Experiments are conducted on sequences of different lengths. The results demonstrate high accuracy and transparency of the model.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів 8
Вступ 10
1 Аналіз предметної галузі та постановка задачі 12
1.1 Еволюція моделей розпізнавання виразів обличчя у відео 12
1.1.1 Класичні методи: ручні ознаки (до 2010 року) 12
1.1.2 Поява глибоких CNN (2012–2016) 13
1.1.3 Графові моделі (2017–2021) 13
1.1.4 Трансформери у FER (2021–дотепер) 14
1.2 Методи графів для динамічного розпізнавання виразу обличчя 14
1.3 Впровадження мереж на основі трансформерів для DFER 17
1.4 Пояснюваний AI (XAI) для DFER 19
1.5 Постановка задачі 21
2 Методи і матеріали 23
2.1 Набір даних і попередня обробка 23
2.2 Архітектура моделі 26
2.3 Порівняння STGT із сучасними моделями 27
2.3.1 CNN + LSTM: базовий гібридний підхід 28
2.3.2 ViViT: трансформери для відео 28
2.3.3 GCN + RNN: графові структури з рецидитивними модулями .. 29
2.3.4 DFER-Net: доменна архітектура для емоцій 30
2.3.5 MMA-DFER: мультимодальне узгодження 31
2.4 Реалізація XAI 33
2.5 Метрики 35
2.6 Інтеграція моделі до гібридної інтелектуальної системи 38
3 Експерименти та результати 40
3.1 Розробка архітектури моделі та тренування 40
3.2 Методи інтеграції пояснювального штучного інтелекту 42
3.3 Альтернативні підходи до XAI у DFER 46

3.3.1 LIME (Local Interpretable Model-agnostic Explanations) 47
3.3.2 SHAP (SHapley Additive exPlanations) 48
3.3.3 Guided Backpropagation 48
3.3.4 LIME (Local Interpretable Model-agnostic Explanations) 49
3.3.5 Visual Prompting + XAI: перспективний напрямок 50
3.4 Аналіз результатів 50
3.4.1 Порівняльний аналіз ознак для емоцій «щастя» і «сум» 50
3.4.2 Вплив довжини відео на якість класифікації 53
3.4.3 Помилки класифікації та їх інтерпретація 55
3.4.4 Аналіз складних випадків 57
3.4.5 Аналіз attention maps у часовій динаміці 59
Висновки 61
Перелік джерел посилання 63
Додаток А Відомість кваліфікаційної роботи 68

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Датасет – впорядкована сукупність даних, що використовується для навчання, тестування та аналізу моделей у процесі обробки інформації;

Патчі – невеликі фрагменти зображення або даних, вирізані з більших об’єктів, які використовуються для локального аналізу, навчання або підвищення ефективності моделі;

AI – Artificial Intelligence – штучний інтелект;

Attention map – карта ваг самоуваги, що показує, на які частини входу фокусується модель;

DFER – Dynamic Facial Expression Recognition – розпізнавання динамічних виразів обличчя;

Feature ablation – метод пояснення, який передбачає послідовне виключення окремих ознак для оцінки їх впливу;

GCN – Graph Convolutional Network – графова згорткова нейронна мережа;

Grad-CAM – Gradient-weighted Class Activation Mapping – метод XAI для побудови теплових карт важливості ознак на основі градієнтів;

LSTM – Long Short-Term Memory – тип рекурентної нейронної мережі, що здатен моделювати довготривалі залежності у даних;

MediaPipe Face Mesh – бібліотека для виявлення й відстеження ключових точок обличчя у відео;

STGT – Spatio-Temporal Graph Transformer – архітектура, що поєднує графову обробку просторових даних та трансформерну обробку часових залежностей;

Transformer – нейронна архітектура з механізмом самоуваги, призначена для обробки послідовностей;

XAI – Explainable Artificial Intelligence – пояснюваний штучний інтелект; підхід до побудови моделей, які можуть інтерпретувати свої рішення.

ВСТУП

Розпізнавання обличчя в динамічних умовах реального світу все ще залишається однією з найбільш складних і вимогливих областей комп'ютерного зору. Розпізнавання обличчя має широкий спектр практичних шляхів розвитку, від технологій громадської безпеки до більш детального аналізу особистих емоцій і взаємодії в соціальній робототехніці. Ми бачимо значний прогрес за останні роки; однак завдання залишається складним через дикі динамічні умови, зміну світла, пози голови та наявність часткових оклюзій.

У комп'ютерному зорі, зокрема, у полі розпізнавання виразу обличчя (FER), згорткові нейронні мережі (CNN) найчастіше використовуються для задач статичного та динамічного двовимірного розпізнавання. Однак CNN може втратити продуктивність у 3D-завданнях через зміну положення об'єкта. Застосування графових згорткових мереж (GCN) [26] для таких завдань є доцільним, оскільки вони ефективно фіксують просторові співвідношення між ключовими точками обличчя, дозволяючи точніше виявляти зміни. У той же час мережі на основі Transformer [22], спочатку створені для завдань обробки природної мови, мають великий потенціал для обробки тимчасової послідовності, фіксуючи динамічні зміни з високою складністю.

Включення в цю структуру керованих моделями НЛП, таких як Llama робить можливим мультимодальний підхід, коли вирази обличчя не тільки розпізнаються, але й перетворюються на відповідні контексту мовні відповіді. Інтеграція Llasa забезпечує експресивний синтез мовлення шляхом зіставлення мовних шаблонів [14] з емоційними виразами, гарантуючи, що виявлені вирази обличчя можуть імітувати людське мовлення, крім того, це покращить ефективні обчислення, роблячи інтеграцію людини та ШІ більш емоційно обізнаною. Використовуючи масштабованість архітектур на основі Llama, Llasa забезпечує високоточний

синтез мовлення, який фіксує не лише лексичний вміст, але й емоційний підтекст виявлених шаблонів виразу обличчя.

Проте мережі зі складною архітектурою залишаються «чорними скриньками», які впливають не тільки на продуктивність, але й на довіру кінцевих користувачів. Цей другий аспект надзвичайно важливий у конкретних сферах, таких як медицина, безпека та обробка персональних даних.

Питання інтерпретації набуло все більшого значення через зростаючий інтерес до прозорості та пояснюваності в процесі прийняття рішень за допомогою ШІ. Користувачі та розробники повинні розуміти, які функції та кадри є вирішальними в прогнозах моделі. Ось чому інтеграція підходів Explainable AI (XAI) у моделі розпізнавання облич є ключовим напрямком для підвищення їх практичної цінності.

Це дослідження спрямоване на розробку та інтеграцію спеціалізованих методів XAI в архітектури на основі графів і трансформаторів, щоб підвищити точність розпізнавання та значно покращити інтерпретативність процесів прийняття рішень. Зокрема, увага зосереджена на аналізі та візуалізації найбільш відповідних орієнтирів обличчя та ключових відеокадрів, які суттєво впливають на кінцевий результат класифікації. Крім того, це дослідження вивчає інтеграцію SpatioTemporal Graph Transformer (STGT) з моделлю синтезу тексту в мову Llasa [13] для створення гібридної мультимодальної системи, що забезпечує емоційно адаптивну систему ШІ [19].

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Еволюція моделей розпізнавання виразів обличчя у відео

Розпізнавання виразів обличчя (Facial Expression Recognition, FER) є однією з найдавніших та водночас найдинамічніших підгалузей комп'ютерного зору. Його історичний розвиток відображає загальні тенденції у штучному інтелекті – від ручного інженерії ознак до сучасних глибоких архітектур, що навчаються на великих масивах даних. У контексті динамічного FER (DFER), де обробляється послідовність кадрів, еволюція моделей відбулася особливо швидко, охоплюючи кілька ключових етапів.

1.1.1 Класичні методи: ручні ознаки (до 2010 року)

Перші підходи до FER у відео ґрунтувалися на ручному виділенні ознак та класичних машинних алгоритмах класифікації. Найбільш відомими з них були:

– HOG (Histogram of Oriented Gradients) – метод опису локальної структури зображення, що дозволяв виділяти контури та градієнти міміки;

– LBP (Local Binary Patterns) – ефективний дескриптор текстур, який застосовували до областей обличчя для опису дрібних змін у виразах;

– SVM (Support Vector Machines) – використовувався як основний класифікатор.

У таких системах відео розглядалося як набір окремих кадрів, а тимчасовий контекст ігнорувався або моделювався через прості фільтри. Основною перевагою цих методів була обчислювальна легкість, однак вони були дуже чутливі до зміни освітлення, ракурсу та шумів.

1.1.2 Поява глибоких CNN (2012–2016)

З проривом моделей глибокого навчання, зокрема AlexNet і VGG, стало можливим автоматично вивчати ознаки без ручного втручання. У FER почали використовувати Convolutional Neural Networks (CNN), які демонстрували високу точність на статичних зображеннях. У відео-контексті ці мережі застосовували або до кожного кадру окремо, або з агрегуванням ознак через середнє або max-pooling.

Однак CNN обмежено враховують часовий контекст, тому на цьому етапі з’являються перші гібридні архітектури:

– CNN + LSTM, де CNN вилучає просторові ознаки, а LSTM (Long Short-Term Memory) обробляє послідовність у часі;

– 3D-CNN, що обчислюють згортки у тривимірному просторі (висота, ширина, час).

Ці моделі покращили якість динамічного FER, але мали недоліки: складність навчання, великий обсяг параметрів і обмежена прозорість.

1.1.3 Графові моделі (2017–2021)

Новий етап у розвитку FER став можливим завдяки появі Graph Neural Networks (GNN). Замість обробки всього зображення, у відео почали використовувати ключові точки обличчя (landmarks) як вузли графа, що з’єднані ребрами відповідно до анатомічної структури обличчя.

GNN забезпечили такі переваги:

– зменшення розмірності вхідних даних;

– стійкість до оклюзій і зміни фону;

– природне моделювання міміки як зміни конфігурації вузлів у графі.

Реалізації типу GCN + LSTM дозволяли враховувати як просторову, так і часову інформацію, проте залишалися обмеженими у здатності масштабуватися до складних відео.

1.1.4 Трансформери у FER (2021–дотепер)

Зі стрімким розвитком трансформерних архітектур у комп’ютерному зорі, зокрема моделей ViT (Vision Transformer), TimeSformer, ViViT, та інших, розпізнавання виразів обличчя у відео отримало новий імпульс. Ці архітектури реалізують механізм self-attention, що дозволяє моделі фокусуватися на найважливіших регіонах обличчя та ключових моментах відеопослідовності без потреби в рекурентних блоках або згортках.

На відміну від CNN або RNN, трансформери:

– одночасно обробляють усю послідовність кадрів, без втрати контексту чи залежностей;

– не залежать від порядку обробки кадрів, що усуває проблему градієнтного згасання, типову для LSTM;

– надають прозорість через attention maps, які можна візуалізувати та інтерпретувати.

У випадку FER це дозволяє моделі виявляти мікро-вирази, поступові емоційні зміни та зв’язки між різними частинами обличчя у просторі й часі. Наприклад, ViViT розділяє обробку відео на просторову та часову [20], завдяки чому досягає високої точності навіть у складних сценаріях: при змінах освітлення, часткових оклюзіях та індивідуальних мімічних відмінностях. Сучасні трансформери довели свою ефективність у багатьох задачах динамічного FER, демонструючи високу масштабованість, стійкість до шуму, а також можливість використання у реальному часі. Завдяки цим властивостям вони дедалі частіше інтегруються в системи розпізнавання емоцій у телемедицині, освіті, водінні та цифровій етиці.

1.2 Методи графів для динамічного розпізнавання виразу обличчя

Методи на основі графів відіграють ключову роль у сучасному підході до динамічного розпізнавання виразів обличчя (DFER), оскільки

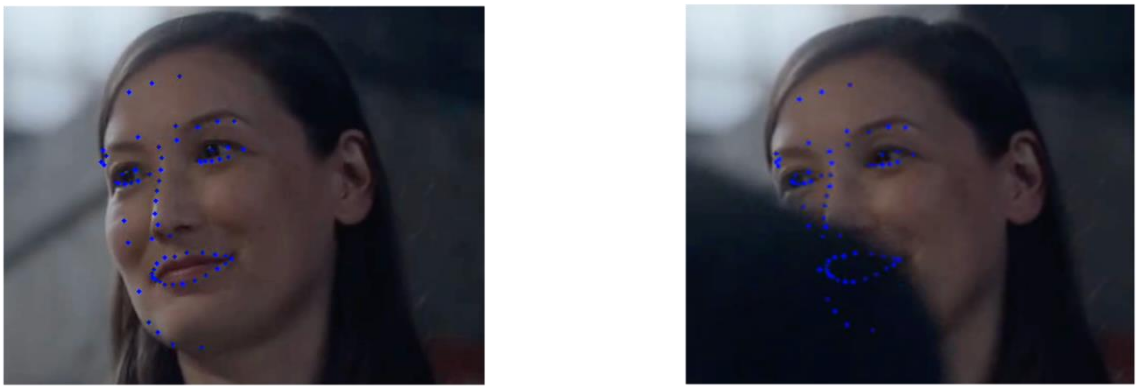
дозволяють моделювати взаємозв’язки між локальними регіонами обличчя у просторово-часовому контексті. У графовій репрезентації обличчя орієнтири (ключові точки) виступають як вузли графа, а ребра визначають логічні або евклідові зв’язки між ними – наприклад, на основі топології обличчя або к-найближчих сусідів (k-NN).

На відміну від традиційних методів, які представляють обличчя як 2D-зображення, графова модель дозволяє явно кодувати структуру обличчя у вигляді графу з 3D-координатами, що суттєво покращує точність у задачах з високою міжособовою варіативністю, частковими перекриттями та складними міміками. Перехід до трьох-вимірного простору (3D) дозволяє моделі краще захоплювати тонкі зміни глибини, що особливо важливо для розпізнавання мікро-виразів, характерних для таких емоцій, як страх, сором або сумнів.

Одним із найефективніших інструментів для автоматичного виділення 3D-орієнтирів обличчя є MediaPipe Face Mesh, розроблений Google. Цей фреймворк дозволяє у реальному часі визначати 468 ключових точок обличчя з високою точністю навіть за умов часткових перекриттів, поворотів голови або варіацій освітлення. Зокрема, координати X та Y нормалізуються у діапазоні [0, 1], що дозволяє працювати незалежно від роздільної здатності відео, тоді як координата Z розраховується відносно осі X, відображаючи глибину у просторі, що дозволяє моделювати тривимірну форму обличчя через проекційну перспективну модель.

Обчислювальна ефективність MediaPipe робить його ідеальним інструментом для інтеграції в реальному часі, навіть на пристроях з обмеженими ресурсами, зокрема мобільних. Це дає можливість поєднувати його з графовими нейронними мережами (GNN), які можуть обробляти кожне відео як динамічну послідовність графів у часі. Таке представлення забезпечує глибше просторове розуміння структури обличчя, виявлення ключових змін між кадрами, та більш стійке виявлення емоційних сигналів навіть у складних умовах. Цей підхід забезпечує послідовне та стабільне

представлення глибини, що є критичним для аналізу мікро-виразів і тонких рухів обличчя в динамічних послідовностях. MediaPipe демонструє потужну здатність виявляти орієнтири на кадрах із перешкодами, як показано на рисунку 1.1.



(A) (B)

Рисунок 1.1 – MediaPipe Face Mesh [16] визначає ключові точки на зразку MAFW [15]. (A) точки визначені на чистому кадрі без поміх; (B) точки визначені на кадрі з поміхами

Обчислювальна ефективність MediaPipe робить його ідеальним інструментом для інтеграції в реальному часі, навіть на пристроях з обмеженими ресурсами, зокрема мобільних. Це дає можливість поєднувати його з графовими нейронними мережами (GNN), які можуть обробляти кожне відео як динамічну послідовність графів у часі. Таке представлення забезпечує глибше просторове розуміння структури обличчя, виявлення ключових змін між кадрами, та більш стійке виявлення емоційних сигналів навіть у складних умовах. Крім того, інтеграція MediaPipe з трансформерними підходами, орієнтованими на self-attention та temporal modeling, дозволяє поєднати найкраще з обох світів: точну локалізацію регіонів інтересу на рівні вузлів графа [27] з одночасним моделюванням довготривалих залежностей між емоційними змінами у часі.

Таким чином, графові методи в комбінації з MediaPipe формують фундамент для побудови продуктивних, точних та пояснюваних моделей DFER, що можуть працювати в реальному часі в задачах охорони здоров'я, дистанційної освіти, емоційної аналітики та ін.

1.3 Впровадження мереж на основі трансформерів для DFER

Останніми роками архітектури на основі трансформерів стали передовим інструментом у сфері динамічного розпізнавання виразів обличчя (DFER), завдяки своїй здатності ефективно моделювати довготривалі залежності у послідовних даних. На відміну від класичних моделей – таких як згорткові нейронні мережі (CNN), які переважно фокусуються на просторовому аналізі, або рекурентні архітектури (RNN, LSTM), що мають обмежену здатність до запам'ятовування довготривалих змін, трансформери реалізують механізм самоуваги (self-attention). Це дозволяє їм паралельно обробляти всю послідовність відео та фокусуватися на критичних фреймах і регіонах обличчя.

У задачі DFER трансформерні моделі виявили себе як потужні засоби для інтегрованої просторово-часової обробки, забезпечуючи комплексне представлення емоційних змін на обличчі. Одним із провідних прикладів є модель ViViT (Video Vision Transformer), яка демонструє інноваційний підхід до розділення обробки відео на просторову та часову складові через факторизовану увагу. На відміну від гібридних архітектур CNN-RNN, які поєднують окремі модулі для простору та часу, ViViT реалізує єдиний механізм уваги, здатний охопити взаємозалежності між піксельними патчами в межах фрейму і між фреймами у послідовності.

Модель працює за принципом поділу кожного відео на патчі (patches), які потім перетворюються у векторні представлення за допомогою лінійного вбудовування. Отримані патч-ембедінги подаються на вхід просторовим та

часовим шарам трансформера, які окремо моделюють зв’язки у межах фрейму та між кадрами. Це дозволяє системі детально вивчати мікро-вирази, дрібні мімічні зміни, а також затримані реакції, характерні для природного спілкування.

Крім того, ViViT [21] показує високу масштабованість до довгих відеопослідовностей, що робить його ідеальним для реальних застосувань, таких як моніторинг емоцій в освіті, телемедицині, психотерапії або системах безпеки. Модель також має вбудовану інваріантність до позиції та механізми компенсації часткових оклюзій, що забезпечує її стабільну роботу навіть за умов зміщень камери, змін освітлення чи часткового перекриття обличчя. Окрему увагу в DFER привертають архітектури, що поєднують трансформери з графовими уявленнями. У цьому підході кожен кадр подається у вигляді графа, де вузлами є орієнтири обличчя, а просторові та часові залежності обробляються через відповідні блоки трансформера:

– Spatial Transformer Block фокусується на важливих регіонах обличчя за допомогою self-attention між орієнтирами;

– Temporal Transformer Block моделює часову динаміку – зміну положення вузлів у часі, виявляючи поступові або раптові зміни виразів.

Ця спільна обробка простору та часу дозволяє моделі захоплювати контекстуальні шаблони, які важко вловити за допомогою класичних мереж [25]. Архітектура виявляється особливо ефективною для задач із високим рівнем складності, таких як:

– аналіз тонких емоційних сигналів (наприклад, сарказму, стриманого задоволення чи внутрішньої тривоги);

– виявлення оклюзій і компенсування втрати частини інформації;

– адаптація до змін поз, ракурсів і глибини, що часто трапляються у відео з реального світу.

Останні експериментальні дослідження [5] підтверджують, що ViViT і його варіанти систематично перевершують архітектури типу CNN-LSTM,

демонструючи вищу точність, стабільність та пояснюваність у широкому спектрі сценаріїв застосування в афективному обчисленні.

1.4 Пояснюваний AI (XAI) для DFER

У контексті динамічного розпізнавання емоцій за виразом обличчя (DFER) пояснюваний штучний інтелект (XAI) відіграє ключову роль у підвищенні прозорості, надійності та довіри до моделей глибинного навчання. Оскільки DFER оперує послідовностями кадрів у реальному часі, він стикається з низкою складних викликів: темпоральними залежностями, варіаціями пози, мікро-виразами, змінами освітлення та індивідуальними анатомічними відмінностями. В умовах реального світу ці чинники можуть суттєво ускладнити класифікацію та знизити довіру користувача до «чорної скриньки» нейронної мережі.

Традиційні моделі, зокрема CNN та RNN, хоч і показують високу точність, залишаються погано інтерпретованими. У зв'язку з цим у DFER активно впроваджуються методи XAI, які дозволяють дослідникам і кінцевим користувачам зрозуміти, які саме просторові та часові ознаки впливають на рішення моделі. Популярні методи, такі як LIME (Local Interpretable Model-agnostic Explanations) та SHAP (SHapley Additive exPlanations), використовуються для визначення впливових ознак на рівні окремого прикладу. Grad-CAM(Gradient-weighted Class Activation Mapping) є ефективним інструментом візуалізації, який виділяє області обличчя, що найбільше сприяють класифікації емоцій.

З появою трансформерів, які використовують механізми уваги, відкрилися нові можливості для XAI у DFER. Завдяки механізму self-attention, трансформери здатні визначати критичні часові фрейми та релевантні просторові області без необхідності жорсткої локалізації ознак. Вони надають інтерпретованість через attention maps, які можна накладати на відео для ідентифікації ключових моментів у міміці.

Графові нейронні мережі (GNN), які моделюють ключові точки обличчя як вузли графа, також забезпечують високий рівень пояснюваності. Завдяки використанню графових attention-механізмів, модель може акцентувати увагу на найбільш релевантних вузлах, що дає змогу створювати інтерпретовані карти важливості орієнтирів. Адаптації Grad-CAM для графових і часово-просторових структур (наприклад, Temporal Grad-CAM) дозволяють генерувати теплові карти, які еволюціонують у часі, наочно демонструючи, які зміни міміки призвели до класифікації певної емоції.

Окрім технічних методів, також розглядаються підходи на основі прототипів (наприклад, ProtoPNet), що забезпечують логіку «розпізнавання через схожість»: модель обґрунтовує рішення, посилаючись на найбільш схожі приклади з навчального набору. Методи контрастивного пояснення (CEM – Contrastive Explanation Method) дозволяють виявити особливості, які відрізняють один клас емоції від іншого, акцентуючи увагу на тонких відмінностях у виразах.

Оцінка якості XAI у DFER виходить за межі простого перегляду теплових карт. Вона включає дослідження з людиною в циклі (human – in – the – loop), експерименти на вірність (faithfulness) пояснень, а також аналіз узгодженості результатів з психологічними моделями емоційного сприйняття. Попри досягнутий прогрес, залишаються відкриті виклики: інтерпретація темпоральних залежностей, подолання упередженості в даних, забезпечення XAI в режимі реального часу для сфер охорони здоров’я, освіти чи епіднагляду.

У перспективі майбутні дослідження мають бути зосереджені на розробці масштабованих, інтерактивних та реалістичних методів XAI, які дозволятимуть не лише пояснювати, але й адаптивно коригувати поведінку моделі відповідно до вимог прозорості, надійності та етичності в критично важливих застосуваннях.

1.5 Постановка задачі

Метою даної кваліфікаційної роботи є поглиблене вивчення підходів до покращення інтерпретованості (explainability) сучасних моделей глибинного навчання, які застосовуються для розпізнавання емоцій на основі виразів обличчя в умовах реального світу. У центрі уваги дослідження знаходяться графові згорткові мережі (GCN) та трансформерні архітектури, як найбільш перспективні підходи для моделювання структурованих і послідовних даних, зокрема ключових точок обличчя у відео.

Об’єктом дослідження є модель Graph Transformer, орієнтована на відеоаналіз емоцій. Вхідними даними для неї слугує послідовність кадрів, у кожному з яких визначено 68 ключових точок обличчя з координатами у форматі (x, y, z). Така структура дозволяє подати відео фрагмент як спостереження за еволюцією графа у часі – тобто динамічну графову послідовність, обробку якої здійснює спеціально адаптована просторово-часова архітектура.

Для досягнення поставленої мети було визначено низку науково-практичних завдань:

– огляд сучасного стану галузі: аналіз літератури та технічної документації з метою вивчення існуючих методів розпізнавання емоцій на основі відеоданих. Особлива увага приділялася підходам, що поєднують графові нейронні мережі, які ефективно працюють із структурованими даними, та трансформери, здатні моделювати складні часові залежності;

– формування навчального набору даних: збір і попередня обробка відео з виразами обличчя, з подальшою екстракцією ключових точок обличчя за допомогою бібліотеки MediaPipe Face Mesh. Усі кадри були перетворені у структурований формат, придатний для обробки графовими та трансформерними моделями;

– розробка архітектури Spatio-Temporal Graph Transformer (STGT): побудова моделі, яка поєднує просторову обробку орієнтирів обличчя у межах кадру (як графа) з часовою обробкою їх змін у динаміці. Основна мета цієї моделі – точна класифікація емоційного стану на основі всієї відео послідовності;

– інтеграція методів пояснення (Explainable AI): застосування таких методів, як Grad-CAM, attention attribution, attention rollout та feature ablation, до побудованої архітектури. Це дозволяє виявити, які просторові та часові особливості найбільше впливають на рішення моделі, підвищуючи довіру до її прогнозів;

– проведення експериментальної оцінки: тестування моделі на сформованому датасеті, з вимірюванням точності класифікації та глибоким аналізом результатів explainability-методів. Це дозволяє оцінити як ефективність моделі, так і інформативність візуалізацій рішень моделі для кінцевого користувача або дослідника.

Таким чином, дослідження має на меті не лише досягнення високої точності у задачі розпізнавання емоцій, але й забезпечення інтерпретованості та прозорості у функціонуванні сучасних AI-систем, що є критично важливим для їх етичного застосування в реальному світі.

2 МЕТОДИ І МАТЕРІАЛИ

2.1 Набір даних і попередня обробка

У цьому дослідженні як основа для створення навчального набору даних було використано датасет MAFW (Multi-modal Affect-in-the-Wild). Основна мета полягала у створенні високоякісного структурованого масиву для навчання трансформерної моделі, здатної ефективно розпізнавати емоційні стани за виразами обличчя у відео.

Початково всі відеофайли були оброблені з використанням MediaPipe Face Mesh – високоточним інструментом, який забезпечує виділення 468 ключових точок обличчя у кожному кадрі. Оскільки робота трансформерної архітектури потребує оптимального балансу між розміром вхідних даних та ефективністю обчислень, було реалізовано скорочення до 68 найбільш інформативних ключових точок, які здатні адекватно передати міміку обличчя.

Подальша обробка включала фреймування відео – поділ на фрагменти фіксованої довжини (T = 50, 100 або 150 кадрів). У межах кожного фрагменту координати (x, y, z) для всіх 68 точок зберігались у форматі (T, N, 3), що повністю відповідало очікуванням графової моделі. Ці дані зберігались у структурованому вигляді – кожен фрейм мав відповідне ім’я кліпу, номер кадру, метку класу (за наявності) та координати точок.

На основі цієї обробки було сформовано три окремі під-набори, що відповідали довжині відео у 50, 100 та 150 кадрів. Сукупно було оброблено 8120 відео-фрагментів, однак частина з них не мала відповідної емоційної мітки, що спричинило зменшення остаточного об’єму до 7706 дійсних кліпів.

Класова нерівновага – важлива характеристика датасету, яка була врахована на етапі тренування моделі. Рисунок 2.1 демонструє розподіл класів по всіх трьох під-наборах. Структура кінцевого масиву даних,

сформованого для навчання моделі, зображена на рисунку 2.2. Вона представляє конкатенацію всіх відеокліпів у вигляді єдиного масиву, що містить координати ключових точок у часовій послідовності та пов’язану метаінформацію.

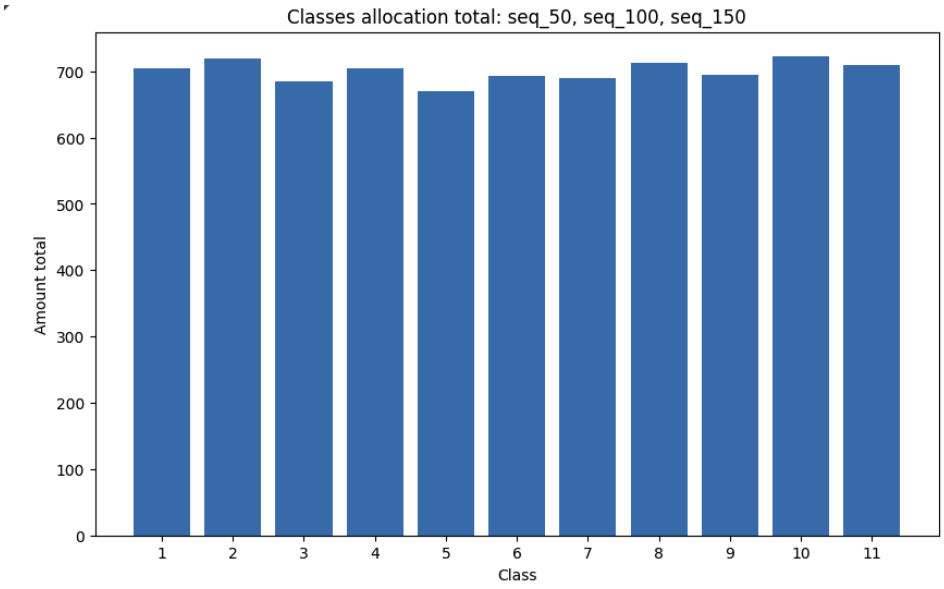


Рисунок 2.1 – Таблиця розподілу класів в отриманому датасеті

Розподіл класів у всіх трьох варіантах підготовлених наборів даних (із довжиною відеопослідовностей 50, 100 та 150 кадрів відповідно) детально проаналізовано та представлено на рисунку 2.1. Ця діаграма забезпечує наочне візуальне відображення нерівномірного балансу між емоційними категоріями, що є характерною рисою реальних емоційних баз даних. Зокрема, можна спостерігати суттєве переважання таких емоцій, як «нейтральність» та «радість», тоді як емоції «відраза», «страх» або «сюрприз» зустрічаються набагато рідше. Такий дисбаланс класів суттєво впливає на процес навчання моделі: вона може переорієнтуватися на домінуючі категорії, ігноруючи менш представлені, що знижує загальну узгодженість результатів.

Візуалізація на рисунку 2.1 дозволяє не лише побачити ці розбіжності, але й порівняти, як змінюється частка кожної емоції залежно від довжини

послідовності. Наприклад, коротші послідовності (T=50) частіше представлені нейтральними станами, тоді як у довших фрагментах (T=150) збільшується частка емоцій з поступовим розвитком (наприклад, «сум», «страх»), що є логічним результатом тривалішого часу спостереження.

Ще один важливий аспект – структура підготовлених даних, яка відіграє ключову роль у коректному навчанні та тестуванні моделі. На рисунку 2.2 наведено схему формату даних, що подаються на вхід моделі STGT. Кожен фрагмент відео представлений як послідовність із T кадрів, у яких збережені координати 68 орієнтирів обличчя (по три просторові координати: x, y, z на кожен). Така структура дозволяє розглядати обличчя як графову структуру у просторі та часі, де кожен кадр відповідає окремому графу, а вся послідовність – багатовимірному тензору з форматом (T, 68, 3).

	clip_name	T_frame	label	x0	y0	z0	x1	\
0	00019.json	0	5	0.398745	0.203495	-0.041064	0.425867	
1	00019.json	1	5	0.398172	0.199947	-0.037238	0.425291	
2	00019.json	2	5	0.397061	0.199594	-0.037346	0.424351	
3	00019.json	3	5	0.396368	0.201792	-0.037634	0.423466	
4	00019.json	4	5	0.394634	0.200031	-0.037633	0.421918	

	y1	z1	x2	...	z64	x65	y65	z65	\
0	0.199946	-0.039751	0.450780	...	0.000267	0.421870	0.523082	0.001159	
1	0.195821	-0.035886	0.450128	...	-0.001719	0.419704	0.522611	-0.000866	
2	0.195394	-0.035746	0.449256	...	-0.002179	0.417977	0.522486	-0.001349	
3	0.197278	-0.036459	0.448318	...	-0.001542	0.419543	0.521060	-0.000600	
4	0.195525	-0.036299	0.446926	...	-0.001748	0.417088	0.522978	-0.000863	

	x66	y66	z66	x67	y67	z67
0	0.414129	0.514805	0.003303	0.408782	0.507868	0.006442
1	0.412001	0.514819	0.001314	0.406835	0.508386	0.004602
2	0.410188	0.514938	0.000820	0.404951	0.508738	0.004112
3	0.411840	0.513410	0.001594	0.406664	0.507003	0.004830
4	0.409384	0.515294	0.001322	0.404251	0.508980	0.004598

[5 rows x 207 columns]

Рисунок 2.2 – Дані відформатовані та структуровані у єдиний масив даних

Кожна така одиниця доповнюється міткою класу (label), що використовується під час навчання моделі для супервізії. Всі записи структуровано у формат, придатний до ефективної пакетної обробки (batching) у фреймворку PyTorch. Завдяки цьому досягається

висока ефективність при обробці великих наборів даних і забезпечується гнучкість під час експериментів із різною довжиною входу.

Таким чином, обидва рисунки – 2.1 і 2.2 – відіграють важливу роль у методології дослідження. Перший – демонструє виклики, пов’язані з дисбалансом емоцій, другий – ілюструє технічну реалізацію структури даних, на базі якої реалізовано просторово-часову обробку виразів обличчя.

2.2 Архітектура моделі

Запропонована архітектура Spatio-Temporal Graph Transformer (STGT) є модульною і спеціально адаптованою для задачі динамічного розпізнавання емоцій за виразом обличчя у відео. Її основною метою є моделювання як просторових залежностей між орієнтирами обличчя в межах окремого кадру, так і часових взаємозв’язків між різними кадрами відео послідовності. Архітектура складається з кількох послідовних компонентів:

– Landmark Embedding Layer: на цьому етапі необроблені тривимірні координати кожного з 68 ключових точок обличчя перетворюються у векторні представлення фіксованої розмірності (наприклад, 64). Це реалізується через лінійний шар, який працює з вхідним тензором розмірності (B, T, N, 3) – де B – розмір пакету (batch), T – кількість кадрів, N=68 – кількість орієнтирів. Результатом є векторний формат (B, T, N, D), де D – розмірність простору ознак. Цей шар дозволяє моделі краще захоплювати складні шаблони розташування ключових точок;

– Spatial Transformer Block: просторовий блок відповідає за вивчення структурних взаємозв’язків між орієнтирами в межах одного кадру. Тут кожен орієнтир розглядається як вузол графа, і між вузлова взаємодія моделюється через багато головковий механізм уваги (Multi- head Attention). Обробка проводиться окремо для кожного кадру, у форматі (B, N, D) (використовуючи N як послідовність), після чого

додається нормалізація та feedforward-мережа для поглиблення представлень. У результаті утворюється покращене просторове представлення кожного кадру;

– Temporal Transformer Block: для моделювання часової динаміки обличчя у відео використовується Temporal Transformer, який застосовується до послідовності кадрів. Після агрегування просторових представлень у вектор кожного кадру, формується послідовність (B, T, D) – де кожен кадр описується як один вектор ознак. Цей блок складається з кількох рівнів Transformer Encoder, кожен з яких застосовує механізм уваги до всієї часової послідовності. Він дозволяє моделі виявляти як короткострокові, так і довгострокові залежності між кадрами. Додатково вбудовані гачки для захоплення градієнтів (hook) забезпечують можливість подальшої інтерпретації на основі Grad-CAM або attention attribution;

– Classification Head: вихідні часові представлення обробляються через глобальне середнє об’єднання за часовим виміром, зменшуючи послідовність до одного вектора ознак розмірності D для кожного відеофрагменту. Цей агрегований вектор передається у повнозв’язний шар, який виконує остаточну класифікацію у простір міток (емоційних станів). Кінцевий результат – логіти розмірності (B, C), де C – кількість класів.

У сукупності, архітектура STGT поєднує в собі переваги графових структур для просторового аналізу та потужність трансформерів для моделювання динаміки, що робить її ефективним рішенням для задач емоційного аналізу на відео.

2.3 Порівняння STGT із сучасними моделями

Запропонована модель Spatio-Temporal Graph Transformer (STGT) поєднує в собі дві сучасні ідеї в області розпізнавання динамічних емоцій – використання графових представлень обличчя та механізмів уваги трансформера для обробки просторово-часових патернів. Для

обґрунтування її ефективності було проведено порівняльний аналіз із популярними архітектурами, що застосовуються в задачах DFER (Dynamic Facial Expression Recognition).

2.3.1 CNN + LSTM: базовий гібридний підхід

Одним з традиційних рішень є комбінація згорткових нейронних мереж (CNN) для вилучення просторових ознак із кожного кадру та довготривалої короткочасної пам’яті (LSTM) для моделювання динаміки емоцій у часі. Такі архітектури добре підходять для обробки відео, але мають кілька суттєвих обмежень:

– обмежене паралельне обчислення: LSTM обробляє послідовність послідовно, що знижує ефективність обчислень на великих відео;

– втрати довготривалих залежностей: попри пам’ять, LSTM схильні до згасання градієнтів;

– складність інтерпретації: LSTM не мають вбудованої уваги, тому важко виявити, які кадри чи ознаки були вирішальними.

2.3.2 ViViT: трансформери для відео

ViViT (Video Vision Transformer) є провідною архітектурою, яка застосовує факторизовану увагу до просторових і часових аспектів відео.

Основні переваги ViViT:

– висока точність: особливо при роботі з довгими послідовностями;

– глобальний контекст: трансформери враховують взаємозв’язки між усіма кадрами та областями одночасно;

– недоліки: потреба у великих обчислювальних ресурсах та складність застосування XAI через непрозорість патчів.

На відміну від ViViT, модель STGT працює не з піксельними патчами, а з 3D-орієнтирами обличчя – це дає змогу краще узгоджувати архітектуру з анатомією людини та зменшує розмірність вхідних даних.

2.3.3 GCN + RNN: графові структури з рецидитивними модулями

Одним із популярних підходів у розпізнаванні динамічних виразів обличчя є поєднання графових згорткових мереж (Graph Convolutional Networks, GCN) та рекурентних нейронних мереж (RNN), зокрема LSTM (Long Short-Term Memory) або GRU (Gated Recurrent Unit). Такий підхід дозволяє обробляти як структуру обличчя у кожному кадрі (через GCN), так і послідовність змін у часі (через RNN), що є базовою вимогою для систем DFER.

У цій архітектурі GCN виступає просторовим модулем, який працює на рівні окремого кадру. Обличчя представляється у вигляді графа, де кожен орієнтир – це вузол, а зв’язки між ними формуються на основі анатомічної або евклідової близькості. GCN дає змогу агрегувати інформацію між сусідніми орієнтирами, що дозволяє виявити локальні патерни міміки: наприклад, підняття брів, стиснення губ або асиметричне напруження щік. Просторова згортка зберігає топологію обличчя, що є суттєвою перевагою у порівнянні зі звичайними CNN, які не враховують структурованість вхідних точок.

Після обробки кожного кадру GCN-модулем, отримані ознаки подаються в часовий модуль – LSTM або GRU, який виконує обробку послідовності ознак у часі.

Рекурентна мережа намагається відстежити еволюцію виразів, вловлюючи зміни у стані обличчя протягом декількох десятків або сотень кадрів. Таким чином, ця комбінація дозволяє виявляти не лише статичні ознаки, але й динамічні прояви емоцій, що розгортаються у часі.

Однак незважаючи на свою функціональність, підхід GCN + RNN має низку суттєвих обмежень. Найбільш критичною є вбудована послідовна природа LSTM, що унеможливлює повноцінну паралельну обробку даних і призводить до довгого часу навчання на великих послідовностях.

Крім того, RNN-модулі часто страждають від згасання або вибуху градієнтів при роботі з довгими відеофрагментами, що ускладнює захоплення довготривалих залежностей.

Навіть при використанні розширених рекурентних блоків з механізмами забування (forget gates), ефективність моделювання складної динаміки емоцій може бути обмеженою.

Ще одним недоліком є обмежена пояснюваність такої архітектури. Оскільки LSTM не має вбудованого механізму self-attention, важко визначити, на які саме кадри модель орієнтується при прийнятті рішення, а також які частини обличчя були найбільш значущими у процесі класифікації. Це ускладнює застосування XAI-методів – таких як attention attribution чи temporal saliency – і знижує прозорість моделі.

2.3.4 DFER-Net: доменна архітектура для емоцій

DFER-Net – це спеціалізована CNN-архітектура, оптимізована під задачі емоційного аналізу. Вона має вбудовану багаторівневу агрегацію ознак, проте:

- не моделює графову структуру обличчя;
- не використовує self-attention;
- має обмежену пояснюваність.

DFER-Net є ефективною у простих сценаріях, але втрачає стабільність у динамічному середовищі.

2.3.5 MMA-DFER: мультимодальне узгодження

Модель MMA- DFER (MultiModal Adaptation for DFER) [4] була запропонована як відповідь на потребу об’єднання різних модальностей даних – зокрема, відео, аудіо та, у деяких реалізаціях, фізіологічних сигналів. Основна ідея моделі полягає у використанні унімодальних попередньо навчених моделей, які спеціалізуються на певному типі інформації (наприклад, CNN для зображень, RNN для аудіо), з подальшою адаптацією їх ознак до спільного латентного простору.

Такий підхід забезпечує високу гнучкість: модель може адаптуватися до наявності чи відсутності певного каналу вхідної інформації, що є важливим у реальних умовах, де якість відео чи звуку може варіюватися. Наприклад, якщо аудіосигнал є зашумленим або відсутнім, система автоматично фокусуватиметься на зорових ознаках, і навпаки. Цей принцип перетину модальностей (cross-modal adaptation) є однією з ключових переваг MMA-DFER.

Варто зазначити, що графовий компонент у MMA-DFER є обмеженим за функціональністю. Хоча вхідні дані можуть бути подані у вигляді ключових точок обличчя, їх подальша обробка не використовує повноцінну графову нейронну архітектуру з урахуванням топології обличчя. Як наслідок, модель не здатна повною мірою враховувати просторові зв’язки між мімічними ділянками, що може призводити до втрати локальних мікропатернів у міміці.

Ще одним суттєвим обмеженням MMA-DFER є відсутність вбудованих засобів пояснюваності (XAI). Модель функціонує як «чорна скринька», а її рішення не можуть бути безпосередньо проінтерпретовані з використанням attention maps, Grad-CAM або інших методів пояснення. Це обмежує її застосування в контекстах, де критично важлива довіра до рішення системи – зокрема, в освіті, психодіагностиці чи криміналістиці.

Таким чином, MMA-DFER є потужним мультимодальним рішенням, особливо у випадках, коли доступні синхронізовані дані з кількох каналів. Однак її обмежена підтримка графової структури обличчя та відсутність пояснюваності знижують її цінність у завданнях, де необхідна інтерпретованість або глибокий аналіз просторових залежностей між мімічними ознаками (таблиця 2.1).

Таблиця 2.1 – Переваги STGT у порівняльному контексті

Критерій	CNN-LTSM	ViViT	GCN-RNN	DFER-Net	STGT
Просторове представлення	Пікселі	Патчі	Орієнтири	Пікселі	Орієнтири
Часова обробка	LTSM	Transformer	LTSM	CNN-based	Transformer
Self-attention	Ні	Так	Ні	Ні	Так
Пояснюваність(XAI)	Ні	Обмежено	Ні	Обмежено	Вбудована(Grad-CAM, Attention Attribution)
Обчислювальна ефективність	Середня	Низька	Середня	Висока	Висока
Придатність до реального часу	Обмежено	Ні	Обмежено	Так	Так
Структуроване уявлення обличчя	Ні	Ні	Так	Ні	Так

Таким чином, модель STGT демонструє найкращий баланс між точністю, поясненістю та придатністю до реального застосування. Вона особливо ефективна в умовах, де важливе розуміння того, чому модель прийняла те чи інше рішення – це особливо критично в етичних, медичних та освітніх системах.

2.4 Реалізація XAI

У задачі динамічного розпізнавання виразів обличчя (DFER) за допомогою моделей графового Трансформера (Graph Transformer) пояснюваність (Explainable Artificial Intelligence, XAI) є ключовим аспектом для підвищення довіри до системи, валідації її рішень та забезпечення інтерпретативності поведінки моделі [3]. У порівнянні з класичними «чорними скриньками», графо-трансформерна модель дозволяє об'єктивно аналізувати, як вона використовує просторову структуру обличчя і його часову динаміку для класифікації емоцій.

У рамках цієї роботи було реалізовано та адаптовано кілька XAI-методів для графо-трансформерної архітектури:

– Attention Attribution [11] – цей метод ґрунтується на аналізі вагових коефіцієнтів уваги (attention weights), що формуються у шарах трансформера під час обробки графа орієнтирів. У контексті Graph Transformer для DFER кожен із 68 орієнтирів обличчя є вузлом, який взаємодіє з іншими через attention-механізм. Attention Attribution дозволяє визначити, які саме вузли (наприклад, очі, кутики рота, брови) отримують найбільшу вагу під час класифікації певної емоції. Це дає змогу створити аналітичну карту значущості орієнтирів, що дозволяє інтерпретувати, наприклад, які області обличчя найбільш важливі для виявлення страху, радості чи презирства. Завдяки часовому компоненту трансформера цей метод також дає можливість виявити зміни важливості вузлів у динаміці, підкреслюючи ті моменти відео, де спостерігаються ключові мімічні зміни;

– абляція функцій – метод абляції ознак (feature ablation) реалізується шляхом маскування або виключення певних вхідних даних з моделі – окремих вузлів (орієнтирів) або часових фреймів. Це дозволяє емпірично визначити, які саме частини обличчя чи моменти відео є критично важливими для прийняття модельного рішення.[SEP]У нашій реалізації було протестовано як просторову абляцію (наприклад, повне виключення

лобової або нижньої частини обличчя), так і часову абляцію (виключення або затемнення окремих кадрів), щоб оцінити стійкість класифікації до відсутності окремих сигналів. Цей підхід дозволяє моделі бути перевіреною на наявність надмірної залежності від обмежених регіонів, ідентифікуючи потенційні упередження;

– Grad-CAM (Gradient-weighted Class Activation Mapping) є популярним методом XAI у CNN-моделях комп'ютерного бачення, але у випадку DFER з Graph Transformer його було адаптовано до графової структури входу. Замість звичайних пікселів Grad-CAM застосовується до вузлів графа, генеруючи теплові карти значущості у тривимірному обличчі. Кожному орієнтиру призначається коефіцієнт впливу, обчислений на основі зворотного поширення градієнта через attention-блоки, що дозволяє візуалізувати «гарячі точки» моделі у просторі обличчя. Це особливо цінно при дослідженні динамічних патернів, оскільки можна простежити, як змінюється фокус моделі у часі – наприклад, чи концентрується вона на очах під час гніву, чи на роті під час радості;

– Attention Rollout дозволяє простежити еволюцію інформації по глибинах трансформера, поєднуючи attention-матриці з усіх шарів у єдину мета-карту. У графо-трансформерній моделі це надає ієрархічне розуміння того, як модель переходить від локальних зв'язків між вузлами до глобального контексту всієї послідовності. Ця методика особливо цінна в гібридних підходах, де графова локальна увага поєднується з глобальним temporal self-attention. Attention Rollout демонструє, як короткочасні фрагменти емоцій (мікро-вирази) накопичуються і формують високорівневий часовий шаблон, який модель використовує для остаточного рішення.

Завдяки інтеграції цих методів у модель Graph Transformer для DFER, ми отримали не лише точні результати класифікації, а й глибоке розуміння внутрішніх механізмів прийняття рішень. Це має вирішальне значення для практичних застосувань у сфері охорони здоров'я, психодіагностики,

відеоспостереження та етичного використання AI у соціально чутливих доменах.

2.5 Метрики

Оцінка ефективності моделі Spatio-Temporal Graph Transformer (STGT) у задачі динамічного розпізнавання виразів обличчя (Dynamic Facial Expression Recognition, DFER) ґрунтується на багаторівневому аналітичному підході, що враховує не лише точність класифікації, але й здатність моделі до узагальнення, інтерпретованість рішень та стабільність при роботі з незбалансованими даними.

З формальної точки зору, базовими метриками виступають точність (accuracy), повнота (recall), точність класифікації по класах (per-class precision), а також F1-score, які забезпечують загальне уявлення про здатність моделі правильно класифікувати кожен тип емоції. Для моніторингу процесу навчання також обчислюється величина функції втрат (loss) як на тренувальній, так і на валідаційній вибірках. Динаміка втрат дозволяє виявляти явища перенавчання (overfitting) або недонавчання, а також оцінити ефективність налаштувань гіперпараметрів, таких як швидкість навчання, розмір батча або регуляризація.

Однак у контексті розпізнавання емоційної динаміки цього недостатньо. Важливо не лише знати, яку емоцію було передбачено, а й зрозуміти, які саме просторово-часові компоненти (тобто які рухи обличчя і в які моменти часу) призвели до такого результату. З цією метою в оцінку було інтегровано методи пояснюваного штучного інтелекту (XAI), зокрема Grad-CAM, Attention Attribution та Attention Rollout. Вони дозволяють візуалізувати, на які вузли графа (орієнтири обличчя) і на які кадри відео модель звертала найбільше уваги при прийнятті рішення. Такий підхід відкриває можливість інтерпретації моделі не як «чорної скриньки», а як

структури з внутрішньою логікою уваги, що особливо важливо для роботи в медичних, освітніх чи етичних системах.

Окрему увагу приділено дисбалансу класів у навчальному датасеті, де деякі емоції (наприклад, «радість» або «нейтральність») представлені значно частіше за інші («страх», «відраза» тощо). Така нерівномірність може призвести до упередженого навчання, коли модель переважно «вгадує» найбільш часті класи. Для компенсації цього ефекту застосовується зважена функція втрат на основі крос-ентропії, у якій ваги класів розраховуються динамічно на основі статистики вибірки. Це дозволяє «підсилити голос» менш представлених емоцій і змусити модель враховувати їх з не меншою важливістю. Такий підхід значно покращує збалансованість класифікації та дозволяє досягти вищої стабільності у прогнозах на складних або мало представлених прикладах.

Загалом, така комплексна стратегія оцінювання дозволяє отримати не лише числові показники якості моделі, але й глибоке розуміння її поведінки, сильних та слабких сторін, а також зон потенційного вдосконалення. Формула втрат визначається як:

$$L_{CE} = -\sum_{i=1}^{C} w_i y_i log_{\hat{y}_i}, \tag{2.1}$$

де C – кількість класів емоцій;

w_i – коефіцієнт ваги присвоєний до класу i;

y_i – основна мітка істинності (одночасне кодування);

$\hat{y}_i$ – передбачувана вірогідність для класу i.

Такий підхід дозволяє моделі уникати упередженості до найчастіше представлених класів та глибше навчатися на менш поширених прикладах.

У процесі тренування відслідковуються втрати як на тренувальному, так і на валідаційній підмножинах. Це дає змогу контролювати збіжність і своєчасно виявляти переобладнання, тоді як динамічний планувальник

швидкості навчання зменшує темп у разі стагнації валідаційної функції втрат. Для оцінки класифікаційної продуктивності використовується матриця плутанини, яка демонструє співвідношення між фактичними та передбаченими класами. Побудова теплових карт із використанням бібліотеки Seaborn допомагає візуально виявити закономірності помилкової класифікації та направити зусилля на вдосконалення просторово-часової обробки.

Grad-CAM генерує теплові карти уваги, обчислюючи градієнт найбільшої оцінки класу щодо карт функцій. Активація Grad-CAM для даного місця (x,y) є:

$$L_{Grad-CAM}^{(c)}(x,y) = ReLU\left(\sum_{k} \alpha_k^c A_k(x,y)\right), \tag{2.2}$$

де $A_k(x,y)$ – карта активації k-ої згорткової карти ознак;

α_k^c – вага важливості, обчислена як:

$$\alpha_k^c = \frac{1}{Z}\sum_{x,y}\frac{\partial S^c}{\partial A_k(x,y)}, \tag{2.3}$$

де S^c – прогнозований бал для класу c;

Z – загальна кількість просторових точок.

Для надійності класифікації ми використовуємо Grad-CAM [24] для аналізу просторової та часової уваги. Grad-CAM використовується для візуалізації, які орієнтири обличчя є найбільш впливовими на класифікацію. Ми створимо два типи теплових карт для просторової Grad-CAM (підсвічує ключові риси обличчя, які сприяють прогнозуванню) і часової Grad- CAM (визначає критичні часові рамки, у яких

відбуваються емоційні переходи). Ці візуалізації забезпечують пояснюваність, роблячи рішення моделі більш зрозумілими.

Для часового Grad-CAM цей процес подовжується в часі шляхом обчислення градієнтів між послідовними кадрами.

Для вибору та оптимізації моделі ми надаємо контрольні точки з найкращими втратами перевірки. Система автоматично зберігає найкращу модель на основі втрати перевірки. Модель зберігається лише в тому випадку, якщо вона покращує втрати підтвердження та підтримує розумні втрати навчання (>0,4), щоб запобігти передчасній конвергенції.

Інтегруючи метрики класифікації, аналіз втрат, візуалізацію матриці плутанини та методи інтерпретації, ми забезпечуємо комплексну оцінку гібридного Spatio-Temporal Graph Transformer і системи синтезу мовлення Llasa. Поєднання всіх перелічених вище гарантує не тільки точність, але й надійність і можливість інтерпретації системи, що робить її придатною для ефективних обчислювальних програм реального світу.

2.6 Інтеграція моделі до гібридної інтелектуальної системи

Щоб побудувати гібридну інтелектуальну систему, ми інтегруємо трансформатор просторово-часового графіка (STGT) із моделлю синтезу мовлення Llasa на основі Llama [13], яка покладається на мовну обробку [8] для забезпечення природного та відповідного контексту емоційного вираження. Використовуючи механізми самоуважності, STGT призначає важливість ключовим шаблонам, ідентифікуючи такі прояви, як гнів, щастя, смуток, здивування, страх тощо.

Виявлені емоційні класи відображаються в лінгвістичній просодії [9] і транслюються в лінгвістичні атрибути, такі як варіація висоти, темп мовлення та вставки пауз.

Llasa включає текстові трансформатори (архітектура на основі Llama) для інтерпретації семантичного значення, що стоїть за створеним або

попередньо визначеним мовним вмістом [1]. Токенізатор мови гарантує, що виявлена емоція з виразу обличчя узгоджується з інтонацією та ритмом синтезованого мовлення.

Кожне емоційно позначене речення зазнає трансформації, щоб відповідати фонетичним і просодичним атрибутам експресивного мовлення. Наприклад, при гнівних виразах мова звучить різко і має підвищену гучність [12].

Llasa використовує кодеки векторного квантування (VQ) для перетворення текстових токенів у виразні мовні сигнали. Використовуючи Process Reward Models (PRM), система інтерактивно вдосконалює синтез мовлення, регулюючи артикуляцію, тон і ритм відповідно до візуальних емоційних сигналів і мовного контексту [7]. Остаточне мовлення тимчасово синхронізується з розпізнаною мімікою, що гарантує, що вимовлені слова, тон і рухи обличчя виглядають разом природно.

3 ЕКСПЕРИМЕНТИ ТА РЕЗУЛЬТАТИ

3.1 Розробка архітектури моделі та тренування

Вихідні дані були об’єднані в табличний формат, що детально описує кожен кадр у кожному відеофрагменті. Кожен запис містив ідентифікатор кліпу, номер кадру у послідовності, відповідну мітку емоційного стану (за наявності), а також координати 68 ключових точок обличчя у форматі (x, y, z). Такий формат дозволив сформувати структурований датасет, оптимізований для графової обробки в режимі міні-батчів.

На етапі вхідної обробки дані надходили до шару вбудовування, який перетворював координати орієнтирів у вектори вищої розмірності за допомогою лінійного перетворення. Це дозволяло моделі навчитися більш абстрактним представленням просторових конфігурацій обличчя, які не залежать від конкретної пози чи освітлення. Кожен кадр інтерпретувався як окремий граф, де вузлами виступали орієнтири обличчя, а між ними створювались неявні зв’язки через багатоголовковий механізм самоуваги (multi-head attention), що давав змогу моделі самостійно визначити релевантні взаємозв’язки.

Після просторової обробки кадрів, векторні представлення подавались у часовий блок трансформера, який агрегував інформацію з усього відеофрагменту. Цей блок обробляв послідовність кадрів як єдину темпоральну сутність, виявляючи закономірності, що виникали протягом часу, наприклад, поступові зміни виразу або мікро-рухи мімічних м’язів. Для отримання фінального прогнозу використовувався класифікаційний шар (classifier head), який поєднував отримані ознаки у єдине представлення і виконував багатокласову класифікацію на 11 емоцій.

Процес навчання моделі проходив поетапно. На першому етапі використовувався набір даних із довжиною послідовностей 50 кадрів, що дозволило швидко протестувати архітектуру і сформувати базову точність.

На другому етапі модель тренувалась на фрагментах по 100 кадрів, що покращило здатність до виявлення довших патернів емоційної динаміки. Нарешті, для досягнення повної темпоральної глибини було застосовано послідовності довжиною 150 кадрів, які максимально розкривали потенціал архітектури в контексті складних, протяжних емоційних реакцій.

Для забезпечення ефективного навчання використовувались сучасні техніки оптимізації: динамічне зменшення швидкості навчання при стагнації втрат, контроль за переобладнанням через валідаційні метрики, а також регулярне збереження найкращої моделі на основі перевірки. У процесі було враховано дисбаланс класів шляхом використання зваженої функції втрат (таблиця 3.1).

Таблиця 3.1 – Параметри натренованої моделі STGT

Classes	Precision	Recall	F1-Score	Support
1 – Anger	0.78	0.79	0.78	543
2 – Disgust	0.90	0.93	0.91	566
3 – Fear	0.92	0.61	0.76	545
4 – Happiness	0.88	0.85	0.86	561
5 – Neutral	0.99	0.97	0.98	572
6 – Sadness	0.92	0.89	0.90	564
7 – Surprise	0.83	0.96	0.89	566
8 – Contempt	0.90	0.80	0.85	558
9 – Anxiety	0.86	0.92	0.88	562
10 – Helplessness	0.94	0.99	0.96	565
11 – Disappointment	0.98	1.00	0.99	563
Accuracy			0.90	6165
Macro avg	0.89	0.89	0.89	6165
Weighted avg	0.91	0.90	0.91	6165

Вихідні показники моделі STGT наведено в таблиці 3.1, що свідчать про її високу продуктивність і здатність узагальнювати як просторові, так і часові ознаки. Точність навченої моделі наведено у рисунку 3.1.

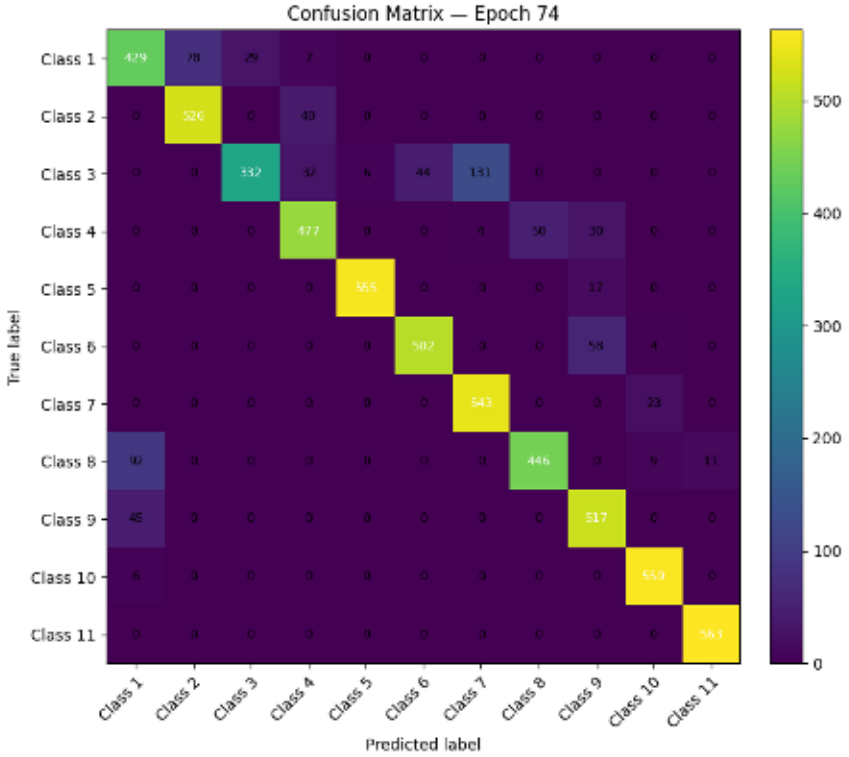


Рисунок 3.1 – Матриця заплутаності для STGT після 74 епох навчання

3.2 Методи інтеграції пояснювального штучного інтелекту

Щоб підкреслити аспекти, які використовує модель у процесі прийняття рішень, ми запровадили методи XAI для чіткішого аналізу. Атрибуцію уваги було реалізовано шляхом вилучення вагових коефіцієнтів уваги безпосередньо з шарів уваги моделі Multi-Head Attention. У просторовому трансформаторному блоці ми отримали вагові коефіцієнти уваги для кожного кадру, показуючи, наскільки важливе значення модель надала кожному орієнтиру обличчя під час прийняття рішень. Подібним чином у Temporal Transformer Block ваги уваги були витягнуті між кадрами,

щоб визначити ключові часові сегменти, що впливають на результати класифікації. Вагові коефіцієнти уваги показані як теплові карти орієнтирів обличчя, чітко вказуючи, які точки є найважливішими для точної класифікації виразів обличчя. Це забезпечує більш чітке розуміння процесу прийняття рішень у моделі.

Для візуалізації та аналізу механізму просторової уваги було реалізовано видобування вагових коефіцієнтів з багатоголовкового шару самоуваги (Multi-Head Attention), інтегрованого в модель Graph Transformer. Вхідні дані – це векторизовані орієнтири обличчя для одного кадру відео, представлені у вигляді тензора розмірності (batch_size, num_nodes, embed_dim). Після приведення тензора до формату, очікуваного модулем nn.MultiheadAttention, було отримано тензор attn_weights, що містить коефіцієнти уваги для кожної пари вузлів. З метою отримання узагальненої карти значущості було виконано усереднення по всіх головах уваги, результатом чого стала квадратна матриця (68 × 68), де кожен елемент відображає, скільки уваги орієнтир i приділяв орієнтиру j. Така матриця дозволяє інтерпретувати внутрішню логіку просторової обробки в моделі, зокрема, виявити ключові зони фокусування при класифікації емоцій, як зображено в лістингу 3.1.

Лістинг 3.1 – Програмний код з атрибуцією просторової уваги

```
import torch
import torch.nn as nn
multihead_attn = nn.MultiheadAttention(embed_dim=64,
num_heads=4, batch_first=True)
x = torch.rand(1, 68, 64)
x = x.transpose(0, 1)
attn_output, attn_weights = multihead_attn(x, x, x,
need_weights=True, average_attn_weights=False)
mean_attention = attn_weights.mean(dim=0).squeeze(0)
```

Для перевірки значущості ознак було застосовано систематичну стратегію абляції ознак. Певні підгрупи (підгрупи) орієнтирів обличчя поступово видалялися, і спостерігалася результуюча зміна точності класифікації. Це дослідження дозволило нам визначити, які риси обличчя найбільше вплинули на процес прийняття рішень у моделі. Кількісно оцінюючи ці ефекти, ми отримуємо впевненість в інтерпретованості та надійності моделі.

З метою оцінки впливу окремих ділянок обличчя на прийняття рішень моделлю було реалізовано експеримент із просторовою абляцією орієнтирів. Суть підходу полягає у вибірковому маскуванні певних груп ключових точок (наприклад, очей, рота, брів), після чого модель виконує повторну класифікацію на модифікованому прикладі. Якщо після видалення певної групи ознак модель змінює свій прогноз, це свідчить про високу важливість цієї ділянки у процесі класифікації емоції.

У коді, що наведений у лістингу 3.2, було сформовано п’ять груп орієнтирів: рот, ліве око, праве око, брови та ніс. Для кожної з них проводиться поетапне маскування шляхом занулення відповідних координат у вхідному тензорі. Результати кожної абляції фіксуються у вигляді словника, де ключем виступає назва ділянки, а значенням – клас емоції, передбачений після видалення відповідної області обличчя. Такий аналіз дозволяє зробити висновки щодо просторової залежності моделі та її чутливості до локальних мімічних ознак, що має важливе значення для підвищення прозорості та довіри до моделі.

Лістинг 3.2 – Програмний код просторової абляції орієнтирів обличчя

```
landmark_groups = {
    'mouth': list(range(48, 68)),
    'left_eye': list(range(36, 42)),
    'right_eye': list(range(42, 48)),
    'eyebrows': list(range(17, 27)),
    'nose': list(range(27, 36))}
```

Продовження лістингу 3.2

```
def ablate_landmarks(x_original, group_idx):
    x_ab = x_original.clone()
    x_ab[:, group_idx, :] = 0
    return x_ab
ablation_results = {}
with torch.no_grad():
    for group, indices in landmark_groups.items():
        x_ab = ablate_landmarks(x_original, indices)
        logits = model(x_ab.unsqueeze(0))
        predicted_class = torch.argmax(logits,
dim=1).item()
        ablation_results[group] = predicted_class
```

Grad-CAM було інтегровано для подальшого покращення інтерпретації. Grad-CAM дозволив візуалізувати градієнти, що переходять в останній згортковий шар (адаптований до нашого підходу на основі трансформатора), виділяючи конкретні орієнтири та області обличчя, які суттєво впливають на результати класифікації. Це дозволило нам візуально перевірити надійність механізмів уваги моделі та підтвердити правильність ідентифікації критичних областей обличчя, пов'язаних з певними емоційними виразами.

Тимчасова Grad-CAM, реалізована в проекті, забезпечує візуалізацію змін уваги моделі для різних кадрів у послідовностях. Модель епохи 10 показує непов'язану дифузну увагу, але на 60-й епосі модель збирає довші періоди, коли увага висока.

Пошарове поширення релевантності [10] широко використовувалося для кількісної оцінки та візуалізації внеску окремих орієнтирів і конкретних фреймів у рішення щодо класифікації. Поширюючи оцінки релевантності з вихідних даних назад до вхідних ознак, LRP [2] дозволив детально проаналізувати вплив кожного орієнтира з плином часу, що полегшило розуміння того, як часова динаміка впливає на прогнози моделі.

3.3 Альтернативні підходи до XAI у DFER

У сучасних системах штучного інтелекту дедалі більшої ваги набуває аспект пояснюваності – здатності моделі не лише видавати результат, а й демонструвати, чому саме було прийнято те чи інше рішення. Такий підхід отримав назву пояснюваний штучний інтелект (Explainable AI, або XAI) і став особливо актуальним у критично важливих галузях – медицині, освіті, соціальному захисті, безпеці.

У задачах динамічного розпізнавання виразів обличчя (DFER) пояснюваність відіграє подвійно важливу роль.

По-перше, емоційні стани самі по собі є складними для інтерпретації – навіть для людини. Вони можуть бути змішаними, частково вираженими, маскованими або суб'єктивно трактованими.

По-друге, сучасні моделі DFER, зокрема трансформери та графові нейронні мережі, є складними багаторівневими структурами з великою кількістю параметрів, що ускладнює їх інтерпретацію «на око».

У таких умовах XAI є не просто додатковою функцією, а необхідною складовою, яка:

– підвищує довіру користувача до результатів моделі;

– дозволяє розробникам виявляти і коригувати помилки або упередження;

– забезпечує етичну відповідальність у застосуваннях, пов'язаних з людськими емоціями.

Особливо це стосується відеоаналізу обличчя, де рішення базується не на одному кадрі, а на послідовності змін у просторово-часових патернах міміки.

Відтак, ефективні XAI-методи повинні враховувати як просторову взаємодію між ключовими точками обличчя, так і темпоральну динаміку емоцій.

У цьому розділі розглянуто сучасні альтернативні підходи до пояснюваності, які не були реалізовані безпосередньо в межах даного проєкту, але є потенційно придатними для його розширення. Вони включають як універсальні XAI-методи, так і специфічні до глибоких моделей (наприклад, SHAP, LIME, Guided Backpropagation), що можуть бути адаптовані до графових або трансформерних архітектур.

Також аналізуються їх переваги, недоліки та можливості інтеграції у задачі розпізнавання емоцій в динаміці. Таким чином, розгляд альтернатив XAI не лише розширює аналітичну частину дослідження, а й закладає основу для подальшого вдосконалення моделі в напрямку прозорості, інтерпретованості та довіри в реальному застосуванні.

3.3.1 LIME (Local Interpretable Model-agnostic Explanations)

Метод LIME базується на побудові локальних лінійних моделей для пояснення поведінки «чорної скриньки» поблизу конкретного прикладу. Для кожного вхідного відеофрагменту LIME генерує набір злегка модифікованих версій (наприклад, з випадковими виключеннями окремих фреймів або ознак), обчислює для них прогнози моделі та будує локальну апроксимацію.

Переваги:

– незалежність від типу моделі – можна застосовувати до CNN, GNN, Transformers;

– інтуїтивність результату – локальна лінійна модель легко інтерпретується.

Недоліки:

– дуже повільний у випадку роботи з відео – високий обсяг пертурбацій;

– не враховує взаємозв’язки між фреймами – порушення темпоральної цілісності;

– погано масштабується до графових структур з багатьма вузлами.

У задачах DFER LIME потенційно може бути застосований для пояснення рішень на окремих кадрах, але не підходить для аналізу довготривалої динаміки або моделей зі складною структурою вхідних ознак.

3.3.2 SHAP (SHapley Additive exPlanations)

SHAP використовує теоретико-ігровий підхід для оцінки внеску кожної вхідної ознаки у прийняття рішення. Він розраховує Shapley values для ознак, що дозволяє будувати як локальні, так і глобальні пояснення.

Переваги:

– строгі математичні основи – на основі теорії кооперативних ігор;

– можна оцінити вплив як окремих орієнтирів, так і цілих регіонів.

Недоліки:

– потребує експоненціального часу при великій кількості вхідних ознак – в нашому випадку – 204 координати на кадр × T кадрів;

– не підтримує графову структуру «із коробки»;

– мало інструментів для візуалізації результатів у часово-просторовій формі.

SHAP більше підходить для табличних моделей або для CNN із відомою структурою ознак. У Graph Transformer архітектурі потрібна серйозна адаптація SHAP до графових представлень, що робить реалізацію складною, але перспективною.

3.3.3 Guided Backpropagation

Цей метод використовує модифіковану зворотну поширену похибку (backpropagation), щоб показати, які входи викликають сильні активації у вихідному шарі. Він добре працює в CNN і дозволяє візуалізувати найбільш важливі області зображення.

Переваги:

– надає зрозумілу теплову карту на рівні пікселів;

– простий у реалізації.

Недоліки:

– не застосовується до графів чи трансформерів без модифікацій;

– не підтримує часовий контекст;

– обмежена пояснюваність при роботі з координатами замість пікселів.

Метод можна адаптувати до Graph Transformer лише після формалізації просторових ознак як «псевдо-зображення», що значно знижує його практичність у поточному контексті.

3.3.4 LIME (Local Interpretable Model-agnostic Explanations)

Ці пояснення показують, які зміни потрібно внести у вхідні дані, щоб модель змінила рішення. Наприклад: «Якби кутики рота не опустилися – модель не класифікувала б це як “сум”».

Переваги:

– забезпечують сильну інтуїтивну пояснюваність для користувача;

– добре підходять для перевірки етичних обґрунтувань рішення.

Недоліки:

– складність генерації валідних змін для координат орієнтирів;

– високі ризики створення «незаконних» (нереальних) обличчя;

– потреба в додаткових моделях-генераторах для зміни обличчя.

У DFER контрфактичні приклади можуть бути корисні для побудови інтерактивних інтерфейсів для психологів або терапевтів, але потребують окремого фреймворку.

3.3.5 Visual Prompting + XAI: перспективний напрямок

У контексті LLMs та Multimodal AI з’явився підхід «visual prompting» – подача зображення або патерну, що спрямовує увагу моделі.

У поєднанні з XAI це може виглядати так: користувач показує ділянку обличчя, а модель пояснює, як вона використовує цю частину у своїх рішеннях.

Переваги:

– можна реалізувати інтерактивні сценарії;

– підходить для навчання та перевірки.

Недоліки:

– поки що експериментальний напрямок;

– потрібна інтеграція з зовнішніми підказками.

У поєднанні з Attention Attribution та Grad-CAM ця ідея може суттєво підвищити рівень пояснюваності та довіри в реальному використанні. Хоча модель STGT вже інтегрує передові XAI-підходи, такі як Grad-CAM та Attention Rollout, розширення системи через альтернативні методи XAI дозволить покращити взаємодію з кінцевим користувачем, забезпечити прозорість у високоризикових доменах, створити адаптивну, гуманно-орієнтовану систему. Майбутня робота повинна бути зосереджена на адаптації SHAP до графових структур, генерації контрфактичних прикладів через автоенкодери або diffusion models, а також інтеграції XAI у режимі «людина в циклі» (Human-in-the-loop AI).

3.4 Аналіз результатів

3.4.1 Порівняльний аналіз ознак для емоцій «щастя» і «сум»

У межах аналізу пояснюваності (XAI) роботи моделі Graph Transformer для DFER було проведено порівняння між двома психологічно

контрастними емоціями – «щастям» та «сумом». Для цього було використано метод Grad-CAM, адаптований до графових структур, з метою візуалізації просторово-часової уваги моделі. Кожен відео фрагмент представлявся у вигляді послідовності орієнтирів обличчя, що можна побачити на рисунку 3.2, на які модель накладала теплові карти уваги, що відображають значущість конкретних вузлів (орієнтирів) для класифікації.

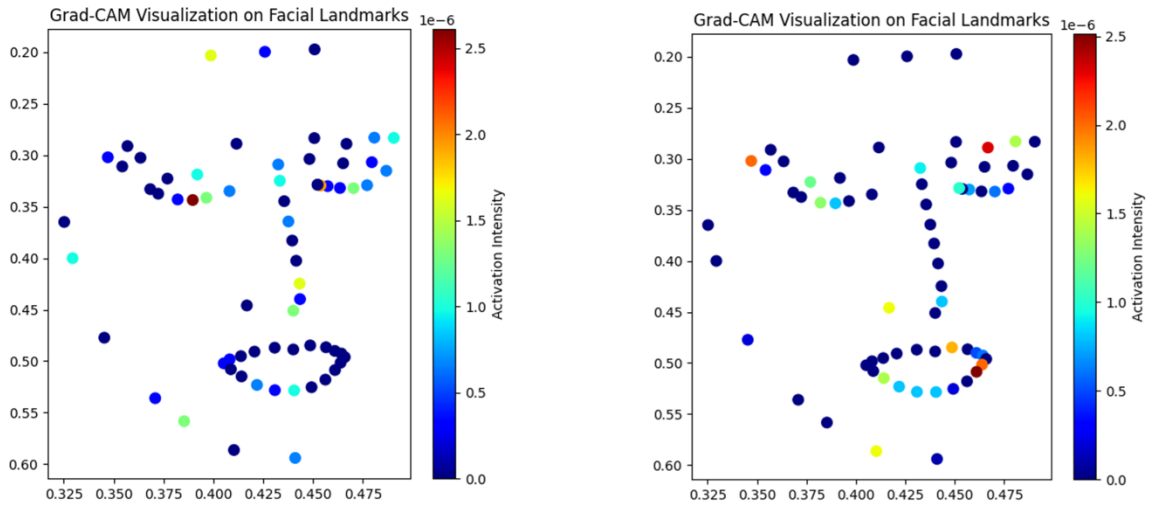


Рисунок 3.2 – Динаміка вдосконалення зображення точок найвищої уваги, цільовий клас 2 – «Відраза». Різниця між моделлю після 10 епох навчання і моделлю після 60 епох навчання.

У випадку «щастя» як зображено на рисунку 3.3 модель стабільно фокусує увагу на:

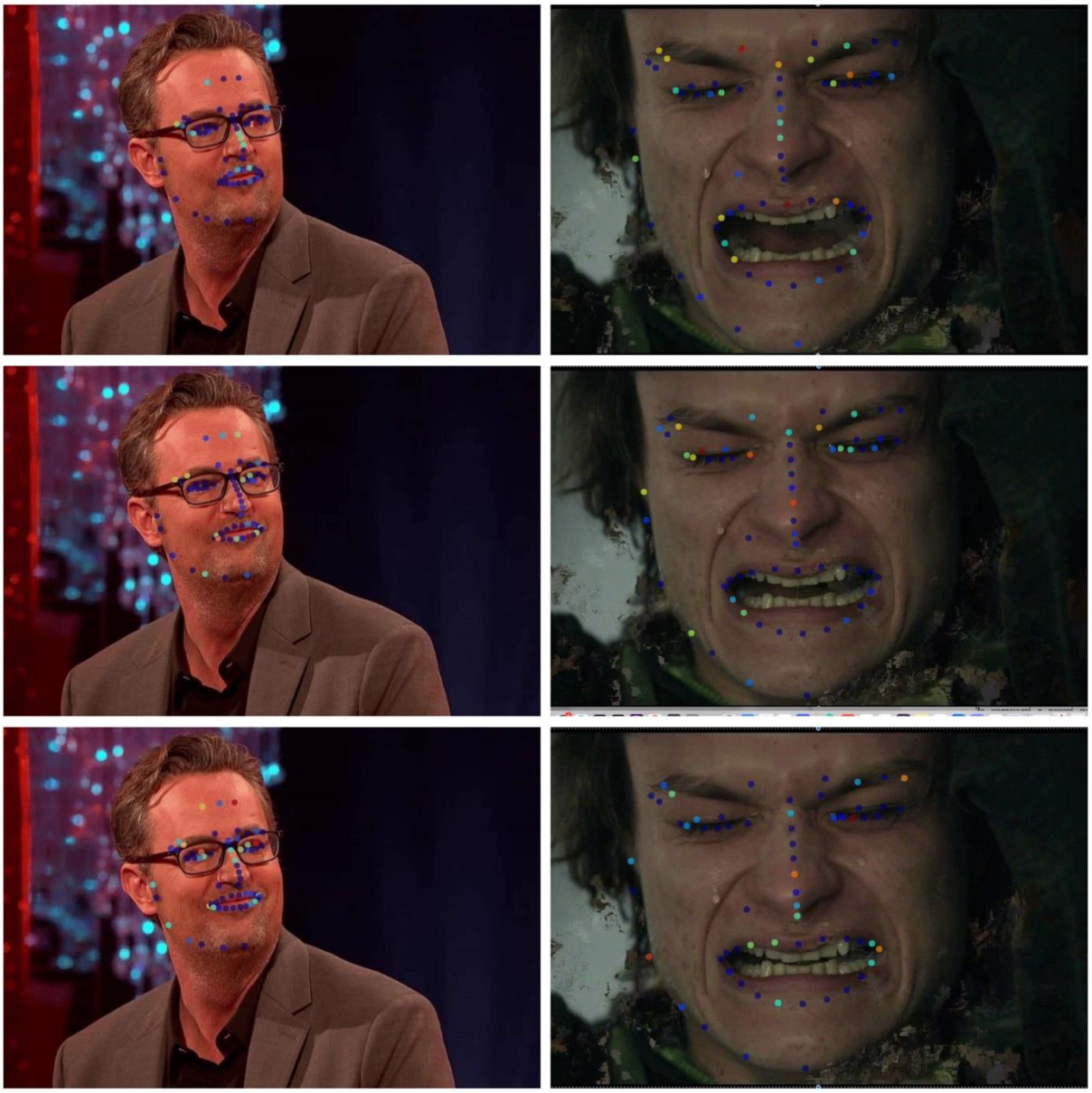
– кутках рота, що підіймаються при посмішці;

– вилицях, що підтягуються вгору;

– області біля очей, зокрема зовнішніх куточків, де з’являються так звані «гусячі лапки».

Активація є інтенсивною, з чітким фокусом у нижній половині обличчя. Це візуально підтверджує фізіологічні маркери радості, що описані

в теорії базових емоцій. Модель демонструє підвищену увагу до моментів, де губи різко змінюють форму або коли з’являється широка посмішка.



ъ

Рисунок 3.3 – Зображення активації уваги в динаміці на двох цільових класах. Зліва для цільового класу 4 – «Щастя», справа – для цільового класу 6 – «Сум».

Для емоції «сум» що продемонстровано на рисунку 3.3, навпаки, активність моделі концентрується на:

– центральній частині брів, що зазвичай нахмурені чи опущені;

– куточках рота, які мають тенденцію опускатися;

– підборідді – через легке напруження м’язів у цій області.

У тепловій карті видно м’якший, але стабільний фокус моделі на нижній частині обличчя. На відміну від «щастя», активність моделі під час «суму» є більш розосередженою, менш симетричною і часто триває протягом довших фрагментів відео, що відображає поступову зміну виразу. Це порівняння демонструє здатність моделі не лише до точної класифікації, але й до глибокого диференціювання між емоційними патернами. У разі «щастя» увага моделі є фокусованою, різкою та пов’язаною активною мімікою, тоді як при «сумі» – дифузною, плавною та асоційоване з м’язовим розслабленням.

Grad-CAM-візуалізації підтверджують, що модель Graph Transformer успішно ідентифікує ключові регіони обличчя, характерні для кожної емоції, і змінює свою увагу відповідно до динаміки міміки в часі. Це дозволяє зробити висновок про пояснюваність і довіру до моделі при використанні її в реальних системах емоційного аналізу.

3.4.2 Вплив довжини відео на якість класифікації

Одним із ключових факторів, що впливають на ефективність моделі динамічного розпізнавання виразів обличчя, є довжина вхідної відеопослідовності, тобто кількість кадрів, що обробляються одночасно. У процесі дослідження було проаналізовано три варіанти: короткі (T = 50), середні (T = 100) та довгі (T = 150) відеофрагменти. Зміна довжини послідовності безпосередньо впливає на те, наскільки повно модель здатна охопити емоційну динаміку у часі.

Модель, яка працює з короткими послідовностями (T = 50), має обмежене уявлення про контекст. Вона здатна вловити лише фрагменти емоційних проявів, що часто призводить до неповного або хибного розпізнавання. Наприклад, у випадках, де емоція розгортається

поступово (типово для «суму» або «відрази»), короткий фрагмент може містити лише нейтральні або перехідні фрейми. Це знижує якість прогнозу і підвищує кількість помилок.

Збільшення довжини послідовності до 100 кадрів суттєво покращує ситуацію: модель отримує більше прикладів мімічних змін і краще відслідковує тривалі емоції, які розгортаються повільно. Це дає змогу точніше виділити ключові фази виразу: початок, пік і завершення. Такі патерни особливо важливі для розрізнення схожих емоцій – наприклад, «нейтральності» та «суму».

Найкращі результати були досягнуті при обробці фрагментів довжиною 150 кадрів. У цьому випадку модель отримує максимальний часовий контекст, що дозволяє не лише фіксувати пік емоції, але й помічати її зародження – мікро-рухи брів, напруження м’язів підборіддя, поступове розширення очей. Такі ознаки особливо важливі для виявлення мікроемоцій, які є короткими, але дуже інформативними проявами внутрішнього стану людини.

Згідно з експериментальними результатами, точність моделі при переході від 50 до 150 кадрів зросла приблизно на 8–10% (для більшості емоцій), що є статистично значущим покращенням. Модель демонструвала вищу впевненість у своїх прогнозах, а attention maps вказували на стабільніші регіони фокусування.

Окремо слід зазначити, що довші послідовності також забезпечують більшу стійкість до шумів – таких як тремтіння камери або часткове перекриття обличчя. Завдяки накопиченню контексту, модель здатна ігнорувати поодинокі артефакти і фокусуватись на глобальній емоційній картині.

У підсумку, довжина послідовності є важливим гіперпараметром у системах DFER. Вона впливає не лише на точність класифікації, але й на глибину інтерпретації міміки. Рекомендується використовувати довгі послідовності (100–150 кадрів) у тих випадках, коли важливо досягти

максимальної точності та пояснюваності – наприклад, у медичних, освітніх або емоційно-чутливих застосуваннях.

3.4.3 Помилки класифікації та їх інтерпретація

Попри високу загальну точність моделі STGT у задачі динамічного розпізнавання емоцій, під час тестування було виявлено низку систематичних помилок класифікації, які заслуговують на окремий аналіз. Вони не лише дозволяють краще зрозуміти обмеження моделі, а й підкреслюють складність самої задачі, яка часто включає психологічно тонкі розмежування між емоціями.

3.4.3.1 Страх – Відраза

Однією з найпоширеніших помилок стала плутанина між емоціями «страх» і «відраза». У багатьох випадках модель помилково класифікувала страх як відразу, особливо коли вираз обличчя не був достатньо вираженим або мав короткочасний характер. Цей тип помилки має як технічне, так і психофізіологічне пояснення.

З технічної точки зору, обидві емоції мають схожі візуальні патерни:

– розширення очей (як реакція на потенційну загрозу або відразу);
– підняття верхньої губи та зморшки на переніссі;
– напруження щелепи або її відкривання.

Для моделі, що аналізує лише координати орієнтирів і не має доступу до глибшого контексту, ці ознаки можуть здаватися еквівалентними. Attention maps у таких випадках показували фокус на областях навколо очей і носа – саме тих, що активуються при обох емоціях.

З психологічного боку, ці емоції мають спільне еволюційне походження як реакції на загрозу: страх – на зовнішню, а відраза – на внутрішню (наприклад, отрута, хвороба). В обох випадках міміка сигналізує

про потребу уникнути чогось. Це ще більше ускладнює завдання моделі, адже кордони між емоціями не завжди чітко виражені навіть для людини.

3.4.3.2 Сум – Нейтральність

Інший поширений випадок – неправильне розпізнавання «суму» як «нейтрального стану». Цей тип помилки виявився особливо частим у фрагментах, де вираз обличчя був м’яким, без яскраво виражених змін.

Причини цього – у природній подібності обох станів:

– обидва можуть супроводжуватися відсутністю різкої міміки;

– м’язи обличчя часто розслаблені;

– погляд направлений вниз, але без активних рухів брів або рота.

З технічної перспективи, модель не завжди може чітко відрізнити пасивне розслаблення (нейтральність) від пригнічення (сум). Особливо це стосується коротких відеофрагментів або низької інтенсивності міміки. Візуалізація Grad-CAM у таких випадках показувала розосереджену увагу, без явного фокусу на ключових точках (наприклад, кути рота чи міжбрів’я).

З психологічної точки зору, сум часто не має яскравих мімічних маркерів, особливо якщо людина намагається стримувати емоцію. Крім того, сум може маскуватися як нейтральність у соціальних контекстах (наприклад, в офіційних або публічних ситуаціях), що також знижує інтенсивність сигналу навіть на рівні орієнтирів.

Узагальнюючи, можна зробити висновок, що помилки класифікації мають як технічну, так і психологічну природу. Деякі емоції є семантично й мімічно близькими, що обмежує здатність моделі до точного розмежування без додаткового контексту. Це підкреслює важливість глибокої інтерпретації attention maps, а також відкриває перспективи для мультимодального підходу, де міміка доповнюється голосом, позою або контекстом ситуації.

3.4.4 Аналіз складних випадків

Реальні умови розпізнавання емоцій значно відрізняються від лабораторного середовища. У повсякденному відео обличчя може бути частково перекритим, освітлення – нерівномірним, а вираз емоцій – стриманим або фрагментарним. Для оцінки стійкості моделі Graph Transformer були проаналізовані кілька типових складних ситуацій, які впливають на якість класифікації.

3.4.4.1 Часткове перекриття обличчя

Одним із поширених сценаріїв у відео з «дикої природи» (in-the-wild) є ситуація, коли обличчя частково закрите – наприклад, рукою, мобільним телефоном, окулярами з великими оправами або нахилом голови за межі кадру. У таких умовах MediaPipe часто втрачає частину орієнтирів, що ускладнює просторове представлення.

Втім, навіть за умов втрати до 20–25% ключових точок, модель зберігала адекватну точність класифікації. Attention maps вказували, що у випадках відсутності інформації з однієї сторони обличчя (наприклад, закрито праву щоку), механізм уваги автоматично переносив фокус на симетричні точки з іншого боку. Це свідчить про здатність моделі до контекстуальної компенсації вхідної інформації – важлива властивість для роботи у реальному часі.

3.4.4.2 Стримані або неповні емоції

Багато емоцій у повсякденному житті проявляються неповністю або стримано. Наприклад, людина може посміхатися лише очима, не задіяючи рот; або навпаки – прояв емоції лише частковий, з мінімальним м'язовим напруженням.

У таких випадках моделі складно зафіксувати класичні патерни, що притаманні яскравим емоційним виразам. Однак виявлено, що модель не завжди потребує повного набору ознак: при наявності навіть локальних сигналів (наприклад, підняття зовнішнього кута ока чи легке опускання брів), вона фіксувала патерни, які достатні для прогнозу. Зокрема, для емоцій «сум», «відраза» або «стримана радість» модель демонструвала прийнятну точність, хоч і знижувалась впевненість (логіти були ближчими до порогових значень).

Ці результати свідчать про гнучкість архітектури: вона не потребує «ідеального» виразу, а навпаки – адаптується до нюансів міміки, що важливо в реальних застосуваннях, де емоції часто пригнічуються або маскуються.

3.4.4.3 Світлові артефакти та фонові завади

Ще одним поширеним ускладненням були тіні на обличчі, контрастні зони освітлення, або яскраве фонове світло, що спричиняли часткове зникнення або зсув орієнтирів. У деяких випадках MediaPipe неправильно локалізував ключові точки, особливо навколо підборіддя, носа або очей.

Попри ці завади, механізм самоуваги у трансформерному блоці дозволяв моделі ігнорувати нерелевантні зони. Attention maps свідчили, що навіть якщо певна область отримала некоректні координати, фокус уваги залишався на стабільних ділянках (наприклад, міжбрів’я або центральна частина рота), що підтверджує внутрішню стійкість моделі до шуму.

Цей ефект можна пояснити глобальним контекстом, який трансформер враховує при прийнятті рішення: окрема аномалія у координатах не домінує, якщо інші ознаки лишаються валідними.

Узагальнюючи, модель Graph Transformer продемонструвала стійкість до низки складних сценаріїв, що типові для роботи в реальних умовах. Навіть у ситуаціях часткового обмеження видимості, стриманих емоцій або

шуму в даних, вона здатна адаптувати вагові фокуси, компенсувати втрати і підтримувати прийнятну точність. Ці результати свідчать про практичну придатність моделі для інтеграції у застосунки, де точність класифікації має бути збережена навіть за непередбачуваних зовнішніх факторів.

3.4.5 Аналіз attention maps у часовій динаміці

Однією з ключових переваг архітектури на основі трансформера є можливість відстежувати динаміку зміщення уваги моделі у просторі та часі. Завдяки механізму self-attention модель не лише приймає рішення, але й одночасно сигналізує, на які саме зони обличчя вона фокусувалася на кожному етапі відео. Така властивість має високу цінність у задачах пояснюваного ШІ (XAI), особливо в контексті емоційної інтерпретації міміки.

У процесі експериментального аналізу було проведено порівняння attention maps для різних емоцій, зокрема для «радості» та «суму». Обидві емоції мають різні патерни прояву, що дозволяє оцінити, наскільки модель вміє адаптувати свою увагу залежно від змісту відео.

3.4.5.1 Динаміка уваги при емоції «радість»

На початку відеофрагмента, коли емоція ще не повністю сформована, модель демонструє розсіяний розподіл уваги – фокус розміщується одночасно на кількох частинах обличчя, включно з чолом, щоками та центральною частиною. Такий розподіл є природним, оскільки початкові зміни у міміці ще не виражені й система намагається визначити ключові області.

У середній фазі, коли емоція «радість» починає проявлятися активніше, увага концентрується переважно на кутах рота, які піднімаються внаслідок активації великої виличної м'язи (musculus zygomaticus major).

Також активізується зона навколо очей, що узгоджується з ознакою справжньої (духеннівської) усмішки, яка включає скорочення м’яза навколо очей. У фінальній частині відео увага моделі часто повертається до центральної осі обличчя – ймовірно, як спосіб фіксації на стабільних орієнтирах при переході до наступної емоції або нейтрального стану.

3.4.5.2 Динаміка уваги при емоції «сум»

На відміну від «радості», емоція «сум» розвивається повільніше та характеризується менш яскраво вираженою мімікою. На початкових кадрах модель проявляє високу невизначеність: розподіл уваги є нестабільним, із незначним фокусом на бровних дугах та зонах навколо очей.

У фазі посилення емоції увага починає зміщуватися до міжбрів’я, внутрішніх кутів очей та куточків рота, які опускаються в результаті активації м’язів, що відповідають за скорботні вирази. Ці зони стають основними регіонами, на які спирається модель при формуванні остаточного прогнозу. Особливо показовим є те, що в момент максимальної експресії «суму» увага моделі концентрується точково, охоплюючи лише 3–5 ключових вузлів обличчя. Це свідчить про те, що модель навчилася фіксувати мінімалістичні, але інформативні патерни, які мають високу діагностичну цінність.

Узагальнюючи, можна зробити висновок, що модель не лише класифікує емоції, а й динамічно змінює свою стратегію фокусування у залежності від типу виразу, його інтенсивності та фази. Така поведінка відповідає логіці людського сприйняття міміки, де початкова невизначеність поступово трансформується у зосередженість на найбільш виразних ознаках. Ці результати демонструють не лише ефективність attention-механізмів, а й відкривають перспективи для побудови довірчих інтерфейсів, де користувач зможе бачити, як саме модель приймає рішення у реальному часі.

ВИСНОВКИ

У результаті проведеного дослідження у межах магістерської кваліфікаційної роботи було розроблено та експериментально апробовано просторово-часову графову трансформерну модель (STGT) для розпізнавання емоцій людини на основі ключових точок обличчя у відеопослідовності. Модель поєднує переваги графових нейронних мереж (GCN) для просторової обробки структури обличчя та механізмів самоуваги трансформера (self-attention) для захоплення динаміки міміки у часі. Такий підхід дозволив створити ефективну архітектуру, яка адаптована до умов реального світу та має високий потенціал для застосування у практичних системах емоційного аналізу.

Для підготовки вхідних даних було реалізовано попередню обробку відео із використанням бібліотеки MediaPipe Face Mesh [17], яка забезпечила точне й швидке вилучення 68 ключових орієнтирів обличчя в кожному кадрі. На основі цих координат було сформовано послідовності графів, які подаються на вхід моделі у вигляді структурованих багатовимірних тензорів. Було підготовлено три версії датасету з довжиною послідовностей 50, 100 і 150 кадрів, що дозволило дослідити вплив часової глибини на якість класифікації.

Особлива увага в дослідженні була приділена проблемі інтерпретованості моделей глибинного навчання. Для цього було інтегровано та адаптовано сучасні XAI-методи: Grad-CAM, Attention Attribution, Attention Rollout та Feature Ablation. Завдяки цим інструментам вдалося візуалізувати, які вузли графа (орієнтири обличчя) та які кадри послідовності найбільше впливають на рішення моделі. Такий аналіз надав змогу не лише підвищити довіру до системи, а й виявити типові помилки класифікації та їх психологічні причини, зокрема плутанину між близькими емоціями – страхом і відразою, сумом і нейтральністю.

Було також проведено розширений аналіз attention maps у часовій динаміці, що дозволив оцінити, як змінюється стратегія фокусування моделі протягом розвитку емоції. Встановлено, що в моменти пікової експресії увага концентрується на найбільш інформативних ділянках обличчя – очах, роті, міжбрів'ї, – що відповідає даним психологічних досліджень. Навіть у складних умовах – при частковому перекритті обличчя, зміненому освітленні або слабких мімічних проявах – модель демонструвала стійкість і здатність до компенсації втрат інформації через глобальний контекст трансформера.

Результати експериментів показали стійке зростання точності при збільшенні довжини вхідної послідовності, що підтверджує важливість урахування повного емоційного контексту. Було також доведено, що використання зваженої функції втрат дозволяє ефективно боротися з дисбалансом класів у навчальному наборі, покращуючи класифікацію малопредставлених емоцій.

У процесі аналізу було виокремлено сильні сторони побудованої архітектури: її гнучкість, масштабованість, інтерпретованість і стійкість до шумів. Водночас ідентифіковано потенційні напрямки для вдосконалення, зокрема – впровадження ієрархічної просторово-часової обробки та поєднання локального й глобального контексту на основі підходів, запропонованих у роботі «Hierarchical Temporal Transformer for 3D Hand Pose Estimation».

Таким чином, дослідження сприяло поглибленню розуміння застосування гібридних архітектур у відеоаналітиці та продемонструвало, що поєднання GCN і Transformer, у поєднанні з XAI-інструментами, може стати основою для створення етично відповідальних, адаптивних і пояснюваних AI-систем, здатних працювати в реальних умовах та надавати користувачам не лише результат, а й обґрунтування своїх рішень.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ahn Y., Chae J., Shin J. Text-to-Speech With Lip Synchronization Based on Speech-Assisted Text-to-Video Alignment and Masked Unit Prediction. *IEEE Signal Processing Letters*. 2025. P. 1–5. URL: https://doi.org/10.1109/lsp.2025.3537949 (date of access: 26.04.2025).

2. Bhati D., others. Neural Network Interpretability with Layer-Wise Relevance Propagation: Novel Techniques for Neuron Selection and Visualization. *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*. Las Vegas, USA, 2025. P. 00441–00447. URL: https://doi.org/10.1109/ccwc62904.2025.10903721 (date of access: 26.04.2025).

3. Molnar C. Interpretable machine learning: навчальний посібник. 3rd ed. *lulu.com*, 2025. 318 p. URL: https://christophm.github.io/interpretable-ml-book (date of access: 26.04.2025).

4. Chumachenko K., Iosifidis A., Gabbouj M. MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, USA, 2024. P. 4673–4682. URL: https://doi.org/10.1109/cvprw63382.2024.00470 (date of access: 26.04.2025).

5. Comparison of Potential Road Accident Detection Algorithms for Modern Machine Vision System / O. Byzkrovnyi et al. *Environment. Technology. Resources*. 2023. Vol. 3. P. 50–55. URL: https://doi.org/10.17770/etr2023vol3.7299 (date of access: 26.04.2025).

6. Das S., Das D. Natural Language Processing (NLP) Techniques: Usability in Human-Computer Interactions. *2024 6th International Conference on Natural Language Processing (ICNLP)*. Xi'an, China, 2024. P. 783–787. URL: https://doi.org/10.1109/ICNLP60986.2024.10692776 (date of access: 26.04.2025).

7. Effectiveness of modern text recognition solutions and tools for common data sources / K. Smelyakov et al. *CEUR Workshop Proceedings*. 2021. Vol. 2870. P. 154–165. URL: https://ceur-ws.org/Vol-2870/ (date of access: 26.04.2025).

8. Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications / K. Smelyakov et al. *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S\&T)*. Kharkiv, Ukraine, 2020. P. 187–191. URL: https://doi.org/10.1109/PICST51311.2020.9467919 (date of access: 26.04.2025).

9. Francis J., Subha M. An Overview of Natural Language Processing (NLP) in Healthcare: Implications for English Language Teaching. *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. Kirtipur, Nepal, 2024. P. 824–827. URL: https://doi.org/10.1109/I-SMAC61858.2024.10714890 (date of access: 26.04.2025).

10. Fu Z., others. Emotion recognition based on multi-modal physiological signals and transfer learning. *Frontiers in Neuroscience*. 2022. Vol. 16. URL: https://doi.org/10.1109/access.2021.3051171 (date of access: 26.04.2025).

11. Jung Y.-J., Han S.-H., Choi H.-J. Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation. *IEEE Access*. 2021. Vol. 9. P. 18670–18681. URL: https://doi.org/10.1109/access.2021.3051171 (date of access: 26.04.2025).

12. Lei S., others. Watch the Speakers: A Hybrid Continuous Attribution Network for Emotion Recognition in Conversation With Emotion Disentanglement. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. Atlanta, USA, 2023. URL: https://doi.org/10.1109/ictai59109.2023.00133 (date of access: 26.04.2025).

13. Li H., Miao S., Feng R. DG-FPN: Learning Dynamic Feature Fusion Based on Graph Convolution Network For Object Detection. *2020 IEEE International Conference on Multimedia and Expo (ICME)*. London, United

Kingdom, 2020. P. 6–10. URL: https://doi.org/10.1109/icme46284.2020.9102838 (date of access: 26.04.2025).

14. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis / Z. Ye et al. 2025. URL: https://doi.org/10.48550/arXiv.2502.04128 (date of access: 26.04.2025).

15. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild / Y. Liu et al. *Proceedings of the 30th ACM International Conference on Multimedia*. Lisbon, Portugal, 2022. P. 9. URL: https://doi.org/10.1145/3503161.3548190 (date of access: 26.04.2025).

16. Oveis A., others. Explainability In Hyperspectral Image Classification: A Study of Xai Through the Shap Algorithm. *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. Athens, Greece, 2023. URL: https://doi.org/10.1109/whispers61460.2023.10430776 (date of access: 26.04.2025).

17. Sánchez-Brizuela G., others. Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. *Virtual Reality*. 2023. URL: https://doi.org/10.1007/s10055-023-00858-06 (date of access: 26.04.2025).

18. Smart Object Detection Using ESP32-CAM Based on YOLO Algorithm / R. P. Narwaria et al. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. Coimbatore, India, 2024. P. 817–820. URL: https://doi.org/10.1109/ICoICI62503.2024.10696374 (date of access: 26.04.2025).

19. Song T., others. MPED: A Multi-Modal Physiological Emotion Database for Discrete Emotion Recognition. *IEEE Access*. 2019. Vol. 7. P. 12177–12191. URL: https://doi.org/10.1109/access.2019.2891579 (date of access: 26.04.2025).

20. Sun G., Lian Z. Deepfake Video Detection Based on the Decomposition of Spatial-Temporal Attention Mechanism in ViViT. *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*. Kaifeng, China, 2024. P. 1629–1634. URL: https://doi.org/10.1109/ispa63168.2024.00221 (date of access: 26.04.2025).

21. Tian Y., Li M., Wang D. DFER-Net: Recognizing Facial Expression In The Wild. *IEEE International Conference on Image Processing (ICIP)*. Anchorage, USA, 2021. URL: https://doi.org/10.1109/icip42928.2021.9506770 (date of access: 26.04.2025).

22. Vaswani A., others. Attention is All you Need. *Advances in Neural Information Processing Systems*. 2017. URL: https://doi.org/10.48550/arXiv.1706.03762 (date of access: 26.04.2025).

23. Velychko D., others. Image Preprocessing and YOLO Architectures for Enhanced Small and Slow-Moving Object Detection. *2024 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. Rochester, USA, 2024. P. 1–4. URL: https://doi.org/10.1109/wnyispw63690.2024.10786503 (date of access: 26.04.2025).

24. Wang S., Zhang Y. Grad-CAM: Understanding AI Models. *Computers, Materials \& Continua*. 2023. Vol. 76, no. 2. P. 1321–1324. URL: https://doi.org/10.32604/cmc.2023.041419 (date of access: 26.04.2025).

25. Wang Z., Liu Y. STAA: Spatio-Temporal Attention Attribution for Real-Time Interpreting Transformer-based Video Models. 2024. URL: https://www.semanticscholar.org/reader/6bfa663955410c4f59d5b9bbdd29f8c36670c463 (date of access: 26.04.2025).

26. Zhao Q., others. Density Division Face Clustering Based on Graph Convolutional Networks. *26th International Conference on Pattern Recognition (ICPR)*. Montreal, Canada, 2022. URL: https://doi.org/10.1109/icpr56361.2022.9956670 (date of access: 26.04.2025).

27. Zheng H., others. A separable spatial-temporal graph learning approach for skeleton-based action recognition. *IEEE Sensors Letters*. 2024. P. 1–4. URL: https://doi.org/10.1109/lsens.2024.3475515 (date of access: 26.04.2025).