Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження архітектур CNN, RNN та ViT для автоматичного розпізнавання емоцій за мімікою людини з метою створення адаптивних інтелектуальних систем
(тема)

Виконав:
здобувач 2 року навчання
групи ІПЗм-23-1

Артур ЮЩЕНКО
(Власне ім'я, ПРІЗВИЩЕ)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник проф. Кирило СМЕЛЯКОВ
(посада, Власне ім'я, ПРІЗВИЩЕ)

Допускається до захисту
Зав. кафедри _____ Кирило СМЕЛЯКОВ
(підпис) (Власне ім'я, ПРІЗВИЩЕ)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

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

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Ющенку Артуру Сергійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи «Дослідження архітектур CNN, RNN та ViT для автоматичного розпізнавання емоцій за мімікою людини з метою створення адаптивних інтелектуальних систем»

Затверджена наказом по університету від 15.04. 2025р. № 290 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 18.06.2025

3. Вихідні дані до роботи розпізнавання емоцій людини, OS MacOS, мова програмування TypeScript, Python, React, TensorFlow, середовище розробки PyCharm

4. Перелік питань, що потрібно опрацювати в роботі

мета роботи, аналіз предметної галузі і постановка задачі, проведення дослідження різних видів архітектур для розпізнавання емоцій людини, надання рекомендацій щодо їх використання, розробка програмного рішення

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Отримання завдання	16.04.2025	*виконано*
2	Аналіз предметної галузі і постановка задачі	18.04.2025	*виконано*
3	Назва за розділами теоретичного і практичного дослідження	24.04.2025	*виконано*
4	Назва за розділами теоретичного і практичного дослідження	24.04.2025	*виконано*
5	Підготовка до апробації результатів дослідження. Публікація матеріалів	29.04.2025	*виконано*
6	Назва за розділами теоретичного і практичного дослідження	8.05.2025	*виконано*
7	Підготовка пояснювальної записки	14.05.2025	*виконано*
8	Підготовка презентації та доповіді	25.05.2025	*виконано*
9	Перевірка на плагіат	11.06.2025	*виконано*
10	Нормоконтроль	12.06.2025	*виконано*
11	Рецензування	13.06.2025	*виконано*
12	Попередній захист	17.06.2025	*виконано*
13	Занесення диплома в електронний архів	17.06.2025	*виконано*
14	Допуск до захисту у зав. кафедри	17.06.2025	*виконано*

Дата видачі завдання 16 квітень 2025р.

Студент (ка) _____ Артур ЮЩЕНКО
(підпис)

Керівник роботи _____ проф. Кирило СМЕЛЯКОВ
(підпис) (посада, Власне ім’я, ПРІЗВИЩЕ)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 78 с., 12 рис., 15 табл., 16 джерел.

ВЕБ-ЗАСТОСУНОК, ЕМОЦІЙНІ СТАНИ, КОМП'ЮТЕРНИЙ ЗІР, МАШИННЕ НАВЧАННЯ, МІМІКА, РОЗПІЗНАВАННЯ ЕМОЦІЙ, ШТУЧНИЙ ІНТЕЛЕКТ, PYTHON, REACT, TENSORFLOW.

Об'єктом дослідження є архітектури Convolutional Neural Networks, Recurrent Neural Networks, Vision Transformer для розпізнавання емоцій людини.

Метою роботи є порівняння та аналіз різних архітектур нейронних мереж для задач розпізнавання емоцій та для подальшої їх інтеграції у веб-додатки з використанням технологій Flask (бекенд) та React (фронтенд).

Проведено аналіз існуючих досліджень у сфері емоційного аналізу за мімікою, виконано огляд найпоширеніших архітектур нейронних мереж (Convolutional Neural Networks, Recurrent Neural Networks, Vision Transformer), розроблено веб-додаток для інтеграції аналізу мімічних даних у реальному часі, проведено експериментальне порівняння продуктивності нейронних мереж за ключовими метриками: точність, площа під кривою (AUC), показник F1, затримка, пропускна здатність і простота інтеграції.

WEB APPLICATION, EMOTIONAL STATES, COMPUTER VISION, MACHINE LEARNING, FACIAL EXPRESSIONS, EMOTION RECOGNITION, ARTIFICIAL INTELLIGENCE, PYTHON, REACT, TENSORFLOW.

The object of the research is the architectures of Convolutional Neural Networks, Recurrent Neural Networks, and Vision Transformer for human emotion recognition.

The aim of the work is to compare and analyze different neural network architectures for emotion recognition tasks and their subsequent integration into web applications using Flask (backend) and React (frontend) technologies.

An analysis of existing research in the field of emotion recognition based on facial expressions has been conducted. A review of the most common neural network architectures (Convolutional Neural Networks, Recurrent Neural Networks, Vision Transformer) has been carried out. A web application for real-time integration of facial expression analysis has been developed. An experimental comparison of neural network performance has been conducted based on key metrics: accuracy, area under the curve (AUC), F1-score, latency, throughput and ease of integration.

Завідувачу кафедри
ПІ_____
(скорочена назва кафедри)
проф. Кирилу СМЕЛЯКОВУ_
(вчене звання, сласне ім'я, прізвище)

ЗАЯВА
щодо самостійності виконання кваліфікаційної роботи та можливості її публікації (та/або публікації анотації кваліфікаційної роботи) в електронному архіві відкритого доступу ElAr KhNURE

Я, Ющенко Артур Сергійович, студент(ка) гр. ІПЗм-23-1, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження архітектур CNN, RNN та ViT для автоматичного розпізнавання емоцій за мімікою людини з метою створення адаптивних інтелектуальних систем», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений(на) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

Дата Підпис

ЗМІСТ

Перелік скорочень

CNN – Convolutional Neural Networks

RNN – Recurrent Neural Networks

ViT – Vision Transformer

ШІ – штучний інтелект

ВСТУП

Сучасний розвиток штучного інтелекту (ШІ) та комп'ютерного зору відкриває широкі можливості для вирішення складних завдань, зокрема аналізу емоційного стану людини за її мімікою. Ця задача є надзвичайно актуальною через її міждисциплінарний характер і зростаючий попит у таких сферах, як психологія, маркетинг, медицина та наука. Автоматизоване розпізнавання емоцій сприяє кращому розумінню людської поведінки, людських емоцій, підвищенню ефективності взаємодії між людьми та машинами, а також створенню інноваційних рішень для вирішення соціальних проблем.

Актуальність теми дослідження зумовлена зростаючою потребою в аналізі великих обсягів даних у реальному часі, що дозволяє оптимізувати процес прийняття рішень. Інтеграція технологій розпізнавання емоцій у веб-додатки відкриває нові можливості для створення сервісів, здатних працювати у режимі реального часу, що є важливим для IT-індустрії. Крім того, розробка таких систем відповідає актуальним тенденціям у сфері персоналізації та автоматизації, які є ключовими драйверами сучасних технологій.

Метою роботи є аналіз та порівняння популярних архітектур нейронних мереж для розпізнавання емоційного стану людини за мімікою обличчя із подальшою інтеграцією в веб-додаток. Це рішення покликане забезпечити високу точність, швидкодію та зручність використання, що дозволить впроваджувати його в різноманітних галузях, таких як маркетинг, освіта чи медицина.

Для досягнення мети роботи поставлено такі завдання:

- провести аналіз сучасних досліджень у галузі розпізнавання емоцій та огляд сучасних архітектур нейронних мереж;
- розробити архітектуру веб-додатка із використанням React для фронтенду та Flask для бекенду для інтеграції нейронних мереж для розпізнавання емоцій;
- налаштувати та навчити нейронні мережі з архітектурами CNN, RNN та ViT, для розпізнавання емоцій задля подальшого аналізу та дослідження;

- протестувати різні архітектури нейронних мереж на різних дата сетах, оцінити їх продуктивність за ключовими метриками;
- сформувати та надати рекомендації щодо застосування подібних рішень у різних прикладних галузях.

Об’єктом дослідження є процес розпізнавання емоційного стану людини за мімікою обличчя. Предметом дослідження є сучасні архітектури нейронних мереж які використовуються для аналізу емоційного стану людини в тому числі які залучаються у інтеграцію веб-додатків.

У ході роботи було розроблено прототип веб-додатка, що здатний здійснювати аналіз емоцій за мімікою обличчя на основі завантаженого фото або відео. Веб додаток надає змогу одразу побачити результати по всім варіантам залучених архітектур щоб побачити в яких ситуаціях який тип архітектури є найкращим.

Отримані результати мають значний потенціал для практичного застосування: у маркетингу вони допомагають аналізувати реакції споживачів на рекламу чи продукт, у психології – спостерігати зміни емоційного стану пацієнтів на протязі сеансу, у науці для аналізу більш детального аналізу відео. Результати дослідження є вагомим внеском у розвиток технологій автоматизованого розпізнавання емоцій і мають перспективи для подальшого розвитку та впровадження у різних галузях.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Огляд існуючих підходів та їх ефективності

Розпізнавання емоцій за мімікою є одним із ключових напрямків у сфері комп’ютерного зору та штучного інтелекту, яке активно використовується в таких галузях як медицина, маркетинг, наука та розваги. Сучасні підходи до цього завдання [1] базуються на нейронних мережах, що забезпечують високу автоматизацію аналізу емоційних виразів обличчя та значно покращують точність класифікації. Основними архітектурами які наданий час застосовуються для вирішення цієї задачі є наступні типи: Convolutional Neural Networks (згорткові нейронні мережі), Recurrent Neural Networks (рекурентні нейронні мережі) та Vision Transformer (трансформатори бачення).

Згорткові нейронні мережі є найбільш поширеними для аналізу статичних зображень обличчя. Їхня ефективність обумовлена здатністю виділяти локальні ознаки, такі як положення брів чи форма рота, які є важливими для розпізнавання емоцій. CNN демонструють високу точність у контрольованих умовах, але їх продуктивність може суттєво знижуватися при зміні ракурсів, освітлення або в разі наявності перешкод на обличчі.

Рекурентні нейронні мережі, найчастіше застосовують для аналізу часових залежностей і використовуються у сценаріях, де необхідно враховувати динаміку мімічних змін, наприклад у відеозаписах. Ці мережі дозволяють розпізнавати емоції у реальному часі, однак їх використання вимагає значних обчислювальних ресурсів, що може обмежувати застосування у системах із низькою затримкою.

Трансформатори бачення представляють новий і перспективний підхід, здатний аналізувати глобальні контексти у зображеннях завдяки своїй архітектурі, яка забезпечує паралельну обробку даних. Вони показують високу ефективність при роботі з великими наборами даних та складними зображеннями. Проте ці моделі також є ресурсномісткими і потребують значної апаратної підтримки, що може ускладнювати їх використання у реальних умовах.

Порівняння цих підходів у різних умовах є важливим завданням, яке дозволить визначити їхні сильні сторони та недоліки. У контексті цього

дослідження така оцінка сприятиме обґрунтованому вибору моделей для інтеграції у веб-додаток, який забезпечить зручність порівняння продуктивності нейронних мереж. Загалом, незважаючи на високу ефективність існуючих підходів, їх продуктивність залежить від якості вхідних даних і зовнішніх факторів. Це підкреслює необхідність рекомендацій для подальшого вдосконалення алгоритмів та адаптації до складних реальних умов.

1.2 Визначення основних обмежень у застосуванні методів

Попри значні успіхи в розпізнаванні емоцій людини за допомогою нейронних мереж [2], існує низка обмежень, яка впливають на застосування на практиці цих методів. По перше, повинен бути великий якісний набір даних для тренування моделі. По друге, на практиці результат залежить від якості вхідних даних, складнощі у врахуванні різноманітності мімічних проявів у різних людей, високі вимоги до обчислювальних ресурсів та етичні виклики, пов’язані зі збором та обробкою персональних даних.

Одним із ключових обмежень є чутливість моделей до зовнішніх умов. Наприклад, змінне освітлення, низька якість камер або наявність перешкод (таких як окуляри, маски чи бороди) можуть суттєво знижувати точність розпізнавання. CNN, які є стандартом для аналізу зображень, демонструють високу ефективність у контрольованих лабораторних умовах, але стають менш точними у реальному середовищі, де такі фактори, як ракурси чи освітлення, є нестабільними, що дуже важливо врахувати при виборі архітектури.

Іншою проблемою раніше загаданою є обмеженість навчальних наборів даних. Більшість доступних датасетів для навчання моделей розпізнавання емоцій не враховують усього спектру демографічних, культурних та вікових особливостей. Наприклад, вираз емоцій може суттєво відрізнятися у представників різних культур або вікових груп, і моделі, навчені на одному типі даних, можуть показувати значно гірші результати на іншому.

Ресурсомісткість є ще одним обмеженням, особливо для моделей, що працюють у реальному часі. RNN, які використовуються для аналізу

послідовностей кадрів у відео, потребують значних обчислювальних ресурсів, що ускладнює їх використання на мобільних або вбудованих пристроях. Схожа проблема стосується і ViT, які хоч і є більш ефективними на великих наборах даних, вимагають значно більшої потужності, ніж традиційні методи.

Не менш важливим є питання конфіденційності даних. Збирання та аналіз мімічних виразів вимагають дотримання суворих етичних стандартів, таких як GDPR, оскільки ці дані можуть містити чутливу інформацію про емоційний стан і психічне здоров'я користувачів. Саме тому іноді необхідно розгорнути власне рішення. Наразі, використання хмарних сервісів для обробки даних підвищує ризик витоку інформації, що створює додаткові виклики для впровадження технологій.

Загалом, обмеження існуючих методів вказують на потребу в розробці адаптивних алгоритмів, які будуть стійкими до зовнішніх факторів і враховуватимуть якщо треба необхідну швидкодію. Також важливими напрямками для вирішення цих проблем є створення рішення яке враховує роботу на пристроях з обмеженими ресурсами або ж з врахуванням обмежених даних для тренування моделі.

1.3 Масштаб проблеми та поточні виклики

Розпізнавання емоцій за мімікою людини є складною і багатогранною проблемою, яка охоплює як технічні, так і етичні аспекти. Масштаб проблеми визначається широким спектром її застосування – від медицини та освіти до маркетингу й безпеки. Попри прогрес у цій сфері, дослідники та розробники стикаються з низкою викликів, які необхідно подолати для створення універсальних, точних і практичних систем.

Однією з основних проблем є висока варіативність мімічних проявів серед людей. Емоції можуть виражатися по-різному залежно від віку, статі, культурного контексту, фізичних особливостей та психічного стану. Наприклад, одна й та сама емоція, така як радість або сум, може мати різні прояви у представників різних культур або навіть у однієї людини за різних обставин. Це ускладнює створення

моделей, які могли б із високою точністю розпізнавати емоції в широкому діапазоні сценаріїв.

Ще одним викликом є адаптація моделей до реальних умов. Хоча алгоритми, такі як CNN або ViT, демонструють високу точність у лабораторних умовах, їхня ефективність значно знижується за межами контрольованих середовищ. Фактори, такі як погане освітлення, зміна ракурсів, низька якість камер або перешкоди (маски або окуляри), значно впливають на результати. Ця проблема є особливо актуальною для додатків, які працюють у реальному часі, таких як системи моніторингу водіїв або відеоспостереження.

Технічні обмеження також створюють серйозні виклики. Багато сучасних моделей є ресурсомісткими, що ускладнює їх використання на мобільних пристроях або вбудованих системах. Наприклад, RNN і ViT потребують значної обчислювальної потужності, що обмежує їх впровадження у проектах із низькими бюджетами чи апаратними обмеженнями. Це особливо критично для задач реального часу, де затримка навіть у кілька секунд може бути неприйнятною.

Окремо варто зазначити етичні виклики, пов’язані з конфіденційністю даних. Обробка персональних даних, таких як мімічні вирази, підпадає під суворі правові регулювання, такі як GDPR, і викликає обґрунтовані побоювання щодо захисту приватності. Додатково, алгоритми часто діють як “чорний ящик”, і пояснення того, як модель прийшла до певного висновку, є недостатньо прозорим. Це знижує рівень довіри користувачів до таких систем, особливо в медичних і освітніх сферах.

Таким чином, масштаб проблеми виходить далеко за межі технічних аспектів і охоплює соціальні, етичні та культурні питання. Для подолання цих викликів необхідно розробляти адаптивні алгоритми, які враховують варіативність міміки, забезпечують високу точність у реальних умовах і мінімізують обробку персональних даних. Додатково потрібно щоб рішення було ресурсоємним для мобільних пристроїв. Подолання цих викликів відкриє нові можливості для впровадження технологій у широкий спектр галузей, зробивши їх більш доступними та ефективними.

2 ОГЛЯД Й АНАЛІЗ ЛІТЕРАТУРНИХ, НАУКОВИХ ДЖЕРЕЛ

2.1 Огляд основних джерел

Для дослідження теми розпізнавання емоційного стану за мімікою людини з використанням нейронних мереж було розглянуто різноманітні літературні та наукові джерела. Вибрані джерела охоплюють фундаментальні теоретичні основи, сучасні методи комп'ютерного зору, результати прикладних досліджень і огляди тенденцій розвитку технологій, також було зроблено акцент на актуальність джерел. Джерела систематизовано за наступними категоріями: оглядові статті, наукові праці, книги, технічна документація та прикладні посібники.

Оглядові статті та наукові публікації:

- "Facial Emotion Detection Using Deep Learning" [3]: у статті представлено систему штучного інтелекту для розпізнавання емоцій за зображенням обличчя. Описано три основні етапи: виявлення обличчя, виділення ознак та класифікація емоцій. Запропоновано архітектуру на основі CNN, яку протестовано на наборах даних FERC-2013 та JAFFE;
- "Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion" [4]: було запропоновано нову ефективну систему розпізнавання мімічних виразів обличчя в реальному часі, яка демонструє високу точність навіть в умовах зашумленості, змін освітлення та часткового перекриття обличчя. Метод фокусується на виділенні ознак із найінформативніших ділянок обличчя та поєднує їх з текстурними й орієнтаційними ознаками;
- "Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions" [5]: розглядає використання Vision Transformers (ViT) для класифікації емоцій. Стаття показує, що трансформери здатні ефективно працювати з великими наборами даних і у складних сценаріях завдяки своїй здатності враховувати глобальний контекст.

Книги та монографії:

- “Computer Vision: Algorithms and Applications” [6]: у книзі розглядаються алгоритми та прикладні аспекти комп’ютерного зору, зокрема аналіз і інтерпретація зображень для вирішення реальних завдань в тому числі розпізнавання емоцій. Автор пропонує науковий підхід до формулювання задач, детально аналізує класичні та глибокі нейронні моделі, доповнюючи їх інженерними методами;
- “Deep Learning with Python” [7]: книжка доступно пояснює основи глибокого навчання за допомогою мови Python. Посібник поєднує інтуїтивні пояснення з практичними прикладами, охоплюючи застосування в комп’ютерному зорі, обробці природної мови та генеративному моделюванні.

Технічна документація та платформи:

- “TensorFlow Documentation” [8]: у документації представлено один із найпотужніших фреймворків для розробки та навчання моделей машинного та глибокого навчання. Вона охоплює повний цикл розробки від побудови нейронних мереж до їх розгортання на різних платформах, зокрема мобільних, серверних і веб-середовищах;
- “Flask Documentation” [9]: у документації описано легкий та гнучкий веб-фреймворк на мові Python, який дозволяє швидко створювати веб-додатки від простих API до повноцінних вебсервісів.

Критеріями відбору джерел була по перше авторитетність, а саме використано джерела, опубліковані у високорейтингових журналах, монографіях відомих авторів і технічній документації від провідних розробників. По друге актуальність, джерела переважно публікувались у період 2019 – 2024 років, що забезпечує сучасність даних і відповідність трендам. По третє об’єктивність, джерела містять аналітичний підхід до проблеми, відсутність упередженості та різносторонній аналіз. Ну і на кінець застосування запропонованих методів підтверджено результатами тестувань і прикладними кейсами.

2.2 Аналіз літератури

На основі оглянутих джерел було здійснено аналіз ключових теорій, моделей та підходів до розпізнавання емоцій за мімікою людини з використанням нейронних мереж. Основну увагу приділено сучасним методам, викликам та перспективам їх подолання в умовах реальних сценаріїв.

Основні теорії та концепції:

- “Facial Emotion Detection Using Deep Learning”: у статті досліджується підхід до розпізнавання емоцій за виразами обличчя із застосуванням глибокого навчання. Основний акцент зроблено на використанні згорткових нейронних мереж (CNN) для автоматичного вилучення візуальних ознак та класифікації емоцій. Ефективність моделі перевірено на двох відкритих датасетах, що демонструє її потенціал для задач соціальної взаємодії людини і машини;
- “Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion”: у статті розглянуто підхід до розпізнавання емоцій за виразами обличчя в умовах реального світу, де якість зображень може бути низькою, а обличчя — частково закритим чи слабо освітленим. Автори пропонують поєднання геометричних ознак, отриманих із розмітки ключових точок обличчя, з текстурними ознаками, які фіксують тонкі зміни міміки;
- “Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions”: у статті досліджується автоматичне розпізнавання емоцій у соціальній взаємодії на основі мімічних сигналів. Основну увагу приділено використанню архітектури ViT для побудови моделі, здатної точно інтерпретувати емоції, зокрема у представників азійських народів, що є недостатньо вивченим напрямом.

Огляд підтверджує, що нейронні мережі лежать в основі сучасних методів розпізнавання емоцій. CNN добре справляються з виділенням ключових ознак на статичних зображеннях обличчя, однак їх ефективність може знижуватися за умов

змінного освітлення, різних ракурсів або наявності перешкод. RNN дають змогу враховувати часову динаміку, що є особливо корисним для аналізу відео, хоча їх використання обмежується високою обчислювальною складністю. Натомість ViT, здатні аналізувати глобальні ознаки, показують високу результативність при роботі з великими наборами даних, що робить їх перспективними для розв'язання задач емоційного аналізу в складних умовах.

Як зазначено у "Recognition of Facial Expressions under Varying Conditions", складні умови, такі як зміни освітлення або культурні відмінності, знижують точність моделей. Інтеграція мультимодальних даних, включаючи текст та голос, дозволяє підвищити ефективність. Висока ресурсомісткість моделей залишається ключовою проблемою. У джерелах пропонуються lightweight-архітектури та хмарні обчислення як засоби для зменшення навантаження. У статті "Facial Emotion Detection Using Deep Learning" підкреслюється важливість забезпечення різноманітності навчальних наборів даних для подолання упередженості моделей щодо певних демографічних груп.

Аналіз літератури демонструє, що поєднання традиційних методів з глибокими нейронними мережами дозволяє підвищити точність у складних умовах. Використання Vision Transformers відкриває нові можливості для класифікації емоцій у великомасштабних проектах. Інтеграція мультимодальних даних є перспективним напрямом для покращення роботи моделей у реальному часі, а оптимізація обчислювальних ресурсів сприяє впровадженню алгоритмів у пристроях з обмеженими можливостями.

2.3 Оцінка актуальності та новизни

Актуальність розглянутих джерел обумовлена швидким розвитком технологій аналізу міміки за допомогою нейронних мереж, які знаходять застосування у міждисциплінарних сферах: медицині, освіті, комерції та розвагах. Ця тема стає дедалі важливішою в умовах зростаючої потреби в автоматизації, персоналізації та адаптивності систем взаємодії з користувачами. Розглянуті

джерела охоплюють як фундаментальні аспекти теорії, так і прикладні інновації, що відображають сучасний стан і перспективи розвитку галузі.

Основною науковою новизною є інтеграція традиційних підходів до комп'ютерного зору з сучасними методами глибокого навчання, такими як ViT. Наприклад, стаття "Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions" демонструє ефективність ViT у врахуванні глобального контексту зображень, що є критичним для задач розпізнавання емоцій у складних сценаріях. Цей підхід поступово витісняє традиційні CNN, завдяки кращій масштабованості та точності.

Інший важливий внесок це акцент на виділенні ознак із найінформативніших ділянок обличчя та поєднування їх з текстурними й орієнтаційними ознаками які розглянуто у "Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion", методика демонструє покращення ефективності в умовах шумності, масок або окулярів.

Технічна документація, як-от "TensorFlow Documentation", пропонує детальний опис як створити ці самі масштабовані рішення використовуючи їх бібліотеку для навчання моделей. В документації все гарно та детально описано що дозволяє спробувати створити власне рішення будь-кому.

Новизна джерел також проявляється в їхньому міждисциплінарному підході. У медицині такі технології можуть використовуватись для моніторингу психічного здоров'я, у тому числі в умовах віддаленого спостереження, а в науці – для більш детального аналізу в наукових роботах. Розробка систем, які враховують демографічну різноманітність і забезпечують стійкість до упередженості, що підкреслюється у "Facial Emotion Detection Using Deep Learning", відкриває нові можливості для масового застосування цих технологій.

Загалом, розглянуті джерела пропонують інноваційні рішення для сучасних викликів у галузі розпізнавання емоцій. Вони демонструють значний прогрес у застосуванні нейронних мереж та відкривають перспективи для створення ефективних, масштабованих і універсальних систем. Актуальність обраних

джерел підтверджується як їхньою відповідністю сучасним тенденціям (2020–2024 роки), так і значущістю для практичного застосування у реальних умовах.

2.4 Висновок з огляду

Проведений огляд літератури дозволяє зробити низку ключових висновків щодо сучасного стану досліджень і розробок у галузі розпізнавання емоцій за мімікою людини з використанням нейронних мереж. Аналіз наукових праць, статей і експериментальних досліджень свідчить про значний прогрес у застосуванні штучного інтелекту у повсякденному житті або ж у спецефічних сферах діяльності таких як психологія, наука або ж маркетинг.

Сучасні методи, такі як згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN) і Vision Transformers (ViT), демонструють високу ефективність у задачах класифікації емоцій. CNN ефективно виділяють локальні ознаки на статичних зображеннях обличчя, тоді як RNN дозволяють враховувати часову динаміку, що робить їх придатними для аналізу відеоданих. Vision Transformers, які відносно недавно були адаптовані для обробки зображень, відкривають нові перспективи завдяки здатності ефективно враховувати глобальні залежності у візуальній інформації, а також працювати з великими обсягами даних, що підвищує точність класифікації навіть у складних умовах.

Водночас оглянуті джерела вказують на наявність низки викликів, які залишаються нерозв’язаними. Однією з ключових проблем є зниження точності існуючих моделей у реальних умовах, зокрема через змінне освітлення, різні ракурси обличчя чи наявність перешкод, таких як окуляри чи маски. Це вказує на потребу у створенні моделей, здатних адаптуватися до широкого спектра умов.

Ще одним викликом є обмежена універсальність моделей через їхню упередженість до певних демографічних груп, що знижує ефективність у глобальному масштабі. Інтеграція більш різноманітних навчальних наборів даних та підходів, орієнтованих на забезпечення інклюзивності, є важливим завданням для майбутніх досліджень.

Інтеграція мультимодальних підходів, які враховують не лише міміку, але й інші аспекти, такі як голос, текст чи жести, є перспективним напрямком для підвищення точності й адаптивності систем. Наприклад, мультимодальні методи, представлені в сучасній літературі, демонструють ефективність у складних сценаріях, включаючи взаємодію з користувачами різного культурного чи демографічного походження.

Також залишається актуальною проблема високої ресурсомісткості моделей, яка обмежує їх застосування на пристроях із низькою обчислювальною потужністю. Використання lightweight архітектур та впровадження хмарних обчислень пропонуються як перспективні рішення для підвищення ефективності систем у реальному часі.

Крім того, питання етичності, конфіденційності даних і забезпечення безпеки залишаються недостатньо висвітленими в сучасній літературі. Це особливо важливо для впровадження технологій у таких чутливих сферах, як медицина, освіта та безпека.

Загалом, розглянуті джерела закладають потужну основу для розвитку технологій розпізнавання емоцій за мімікою. Вони вказують на значний прогрес у застосуванні глибокого навчання та надають інструменти для подолання сучасних викликів. Подальші дослідження мають бути спрямовані на розробку більш стійких, адаптивних і енергоефективних рішень, що відкриє нові можливості для їхнього впровадження у різних галузях.

3 ПОСТАНОВКА ЗАДАЧІ

3.1 Мета і кінцеві результати дослідження

Метою дослідження є глибоке порівняння різних варіантів архітектур нейронних мереж для задачі розпізнавання емоцій за мімікою людини, що дозволить не лише оцінити їхню ефективність, але й визначити найкращі підходи для конкретних умов. Ця мета включає детальний аналіз таких популярних архітектур, як згорткові нейронні мережі (CNN), рекурентні мережі (RNN) та Vision Transformers (ViT), які є провідними у сфері комп'ютерного зору. Порівняння моделей буде здійснюватися за кількома критеріями, серед яких точність класифікації емоцій, швидкість обробки даних, вимоги до ресурсів і здатність працювати у реальних умовах зі змінними факторами, такими як освітлення, ракурси або перешкоди (окуляри, маски).

Особливий акцент зроблено на розробці інтерактивного веб-інтерфейсу, який дозволить користувачам зручно тестувати, порівнювати та аналізувати результати роботи різних моделей. Цей інтерфейс матиме функціонал для завантаження зображень або відео, запуску розпізнавання та надання результатів у вигляді зрозумілих графіків, таблиць та ключових метрик. Такий підхід забезпечить не лише візуалізацію продуктивності моделей, але й дозволить проводити аналіз у динаміці, що особливо важливо для задач реального часу.

Кінцевим результатом стане не лише порівняльний аналіз різних архітектур нейронних мереж, але й практичний інструмент, який зможуть використовувати розробники, дослідники та інженери для вибору оптимальної моделі залежно від специфіки їхніх завдань. Наприклад, у медичних застосуваннях, де точність є критично важливою, можна буде обрати модель з максимальними показниками точності. У маркетинговій сфері, де важлива швидкість, можна буде використовувати менш ресурсомісткі архітектури.

Крім того, дослідження передбачає розробку практичних рекомендацій щодо інтеграції найкращих моделей у реальні сценарії. Це стосується таких галузей, як медицина (моніторинг емоцій пацієнтів), маркетинг (аналіз реакцій

споживачів) та розваги (персоналізація контенту). Таким чином, мета дослідження охоплює як теоретичний внесок у розвиток технологій розпізнавання емоцій, так і практичний інструмент для їх використання. Цей підхід сприятиме підвищенню точності, зручності та ефективності рішень у широкому спектрі застосувань.

3.2 Обґрунтування вибору методів та інструментів

Для досягнення мети дослідження, яка полягає в порівнянні різних варіантів нейронних мереж для розпізнавання емоцій за мімікою людини, обрано сучасні методи і програмні інструменти, які забезпечують ефективність, гнучкість і точність. Основними архітектурами, які будуть аналізуватись, є CNN, RNN та ViT. Вибір цих архітектур зумовлений їхньою перевіреною ефективністю у задачах комп'ютерного зору, а також здатністю справлятися з різними типами вхідних даних, такими як статичні зображення або відеопослідовності

Для реалізації моделей використовуватиметься такий фреймворк, як TensorFlow, який надає широкі можливості для розробки, навчання та тестування алгоритмів. Він підтримує інтеграцію попередньо навчених моделей, що дозволяє пришвидшити процес розробки та порівняння. Вибір цього інструмента також обумовлений його сумісністю з хмарними сервісами, такими як Google Colab або AWS, що забезпечує доступ до обчислювальних ресурсів із GPU для навчання моделей.

Для оцінювання ефективності кожної архітектури планується використовувати метрики, такі як точність (accuracy), F1-міра, затримка обробки (latency), площа під кривою (AUC), пропускна здатність. Дані для навчання та тестування беруться з кількох відкритих наборів, які є загальновизнаними у галузі розпізнавання емоцій за мімікою. Один із найбільш популярних — це FER2013 [10], який містить понад 35 тисяч зображень облич різних людей, зібраних у різних умовах освітлення і ракурсах. У цьому датасеті представлено сім базових класів емоцій, таких як щастя, сум, злість, страх, здивування, відраза і

нейтральний вираз. Завдяки своїй доступності і великій кількості даних FER2013 часто використовується як базовий набір для тренування та порівняння моделей.

Ще один важливий набір — RAF-DB [11], який відрізняється тим, що зображення у ньому більш якісні і різноманітні за емоційними виразами. Цей датасет містить близько 30 тисяч фотографій, і у ньому розглядається 7 основних категорій емоцій, подібних до FER2013, але також є розширена версія, яка включає 12 складніших класів, таких як захоплення або спокій. RAF-DB містить фотографії з різних вікових груп і етнічних належностей, що робить його корисним для навчання моделей, які мають бути більш універсальними.

Найбільший із цих наборів — AffectNet [12], який включає приблизно 450 тисяч зображень, отриманих із відкритих джерел в інтернеті. AffectNet відзначається не тільки великою кількістю даних, але й тим, що включає 8 класів емоцій, а також можливість роботи з непрямими виразами, які складніше класифікувати. У цьому датасеті представлені зображення з дуже різноманітними умовами зйомки — від професійних фотографій до повсякденних знімків із веб-камер, що дозволяє розробляти моделі, більш стійкі до шуму і змін навколишнього середовища.

Також важливою частиною дослідження є розробка інтерактивного веб-інтерфейсу для візуалізації та порівняння продуктивності моделей. Для цієї мети буде використовуватись сучасний веб-стек, який може включати Python (Flask) для бекенд-розробки та TypeScript (React) для створення інтуїтивно зрозумілого фронтенду.

Обрані методи та інструменти забезпечують досягнення поставлених цілей, дозволяючи отримати науково обґрунтовані висновки щодо порівняння моделей, а також створити практичний інструмент для їх подальшого використання.

3.3 Обмеження у процесі дослідження

У процесі виконання дослідження існують певні обмеження, які можуть впливати на його результати та реалізацію. Одним із ключових обмежень є залежність від доступних датасетів. Набори даних, такі як FER2013, AffectNet та

RAF-DB, хоч і є широко використовуваними у науковій спільноті, мають обмеження щодо репрезентації демографічного та культурного різноманіття. Це може призвести до упередженості моделей і зниження їхньої ефективності при тестуванні на нових групах користувачів або у реальних умовах.

Ще одним обмеженням є висока ресурсомісткість процесу навчання сучасних архітектур нейронних мереж, таких як ViT чи RNN. Навчання та тестування моделей потребують значних обчислювальних потужностей, що може бути проблемою у разі обмеженого доступу до GPU чи хмарних сервісів. Це обмеження особливо важливе для задач реального часу, де продуктивність і швидкість обробки даних відіграють ключову роль.

Додатково, чутливість моделей до зовнішніх факторів, таких як змінне освітлення, наявність перешкод (окуляри, маски) або варіативність ракурсів, може впливати на результати. Попри те, що обрані моделі CNN, ViT є ефективними у стандартних умовах, реальні сценарії використання можуть виявити їхні слабкі сторони, що потребує додаткової адаптації або використання комбінованих підходів.

Етичні аспекти також створюють обмеження, особливо у контексті використання персональних даних. Мімічні вирази можуть розкривати чутливу інформацію про емоційний стан або психічне здоров’я користувачів. Для дотримання конфіденційності необхідно враховувати правові стандарти, такі як GDPR. Це ускладнює реалізацію проєкту, оскільки потребує додаткових зусиль для збереження даних і забезпечення етичності.

Нарешті, часові обмеження також є важливим фактором, оскільки розробка, навчання моделей, тестування та створення веб-інтерфейсу є трудомісткими процесами, які потребують ретельного планування. Усі ці обмеження вказують на необхідність оптимізації підходів і використання адаптивних стратегій для досягнення цілей дослідження в умовах обмежених ресурсів.

4 ТЕОРЕТИЧНЕ ДОСЛІДЖЕНЯ

4.1 Аналіз підходів для дослідження

Для порівняння різних нейронних мереж, що застосовуються для розпізнавання емоцій за мімікою людини, розглянуто три основні підходи: згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN) і трансформатори бачення (ViT). Кожна архітектура має свої сильні сторони й обмеження, які впливають на її ефективність у різних сценаріях.

CNN, широко використовуються для аналізу статичних зображень обличчя. Вони спеціалізуються на виділенні локальних ознак, таких як форма рота, положення очей і брів, що є ключовими для розпізнавання емоцій. CNN забезпечують високу точність у контрольованих умовах, але їх ефективність знижується в складних реальних сценаріях, де можуть бути присутніми такі фактори, як змінне освітлення, ракурси або перешкоди (окуляри, маски). Крім того, CNN відносно швидко виконують обробку зображень, що робить їх зручними для систем із низькими затримками. Проте цей підхід може бути обмеженим у задачах, де важлива динаміка змін міміки.

RNN, орієнтовані на обробку часових послідовностей, що дозволяє їм враховувати динамічні зміни міміки у відеопослідовностях. Їх використання є виправданим у сценаріях, де важливо визначати емоційні переходи в реальному часі. Модифікації на основі LSTM забезпечують стійкість до зникнення градієнта під час навчання й дозволяють моделювати довгострокові залежності. Однак дуже цікаво як вони себе поведуть у випадку датасету який складається просто з фото, бо RNN є ресурсомісткими й мають більш тривалий час обробки, що може обмежувати їх використання в додатках із вимогами до низької затримки.

ViT, є інноваційним підходом, який працює з глобальним контекстом зображень завдяки використанню механізму самопідсилення. Вони демонструють високу ефективність на великих наборах даних і складних сценаріях, де необхідно обробляти зображення з великою кількістю деталей. ViT забезпечують гнучкість і адаптивність, але вимагають значних обчислювальних потужностей, що може

стати перешкодою для їхнього використання в реальних додатках із обмеженими ресурсами.

На архітектурному рівні CNN забезпечують швидку обробку даних і високу точність у статичних умовах, що робить їх оптимальними для простих завдань або проектів зі статичними даними. RNN пропонує більше можливостей для аналізу динамічних змін, але є більш вимогливими до ресурсів. ViT є перспективним вибором для завдань із великими наборами даних або складними умовами, але потребують значних технічних ресурсів для навчання та роботи.

Порівняння цих підходів у рамках інтерактивного веб-інтерфейсу дозволяє наочно оцінити їхню ефективність за такими ключовими параметрами, як точність, швидкодія та стійкість до зовнішніх факторів. Це сприяє вибору оптимальної архітектури залежно від конкретних вимог і умов використання.

4.2 Архітектура та проектування ПЗ

Для реалізації програмного забезпечення для тестування було обрано тришарову клієнт-серверну архітектуру на основі React, Flask і PostgreSQL (див. рис. 4.1).

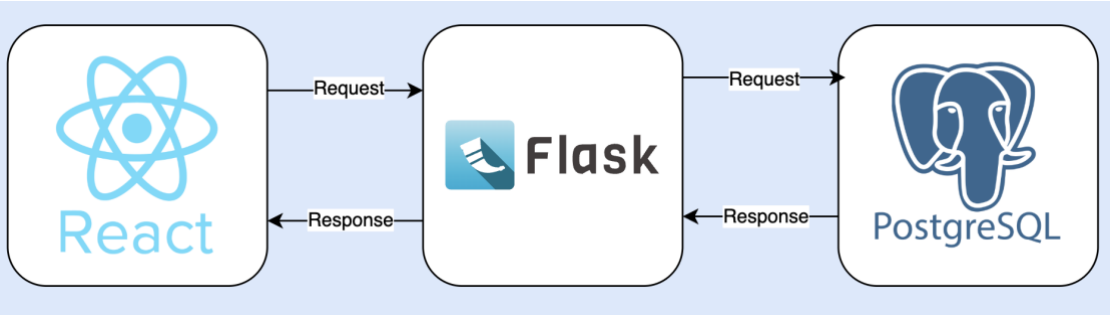

Рисунок 4.1 – Архітектура додатку (рисунок створено самостійно)

Ця архітектура забезпечує чітке розділення обов’язків між клієнтською частиною, сервером і базою даних, що сприяє масштабованості, зручності підтримки та гнучкості в розширенні функціональності.

На клієнтському рівні використовується React, що дозволяє створювати інтерактивний користувацький інтерфейс для порівняння продуктивності нейронних мереж. Основними компонентами є:

– Завантаження даних: Інтерфейс для завантаження зображень або відео для аналізу;
– Візуалізація результатів: Графічні компоненти для відображення метрик продуктивності, таких як точність і швидкість обробки.

На серверному рівні використовується Flask, що виконує роль обробника запитів від клієнта до навчених моделей. Сервер реалізує REST API, який дозволяє взаємодіяти з базою даних і запускати моделі нейронних мереж. Основні ендпоїнти включають:

– POST /analyze: Виконання аналізу даних нейронною мережею одразу після завантаження зображення чи відео;
– GET /analysis-history: Отримання історії аналізів які були проведенні;
– GET /analyze-detail: Отримання детальних метрики конкретного аналізу який був проведений раніше.

Рівень бази даних реалізовано на основі PostgreSQL, що забезпечує зберігання результатів тестування, для подальшого їх перегляду.

Основними таблицями є:

– models: Інформація про використовувані моделі (назва, тип, версія);
– tests: Табличка тестування до якого прив’язанні завантаження, тип, дата;
– uploads: Лінки на зображення або відео до них прив’язані результати;
– results: Результат де є процент впевненості, тип емоції, тип моделі.

Для кращого розуміння структури даних і зв’язків між ними буде створено схему БД (див. рис. 4.2):

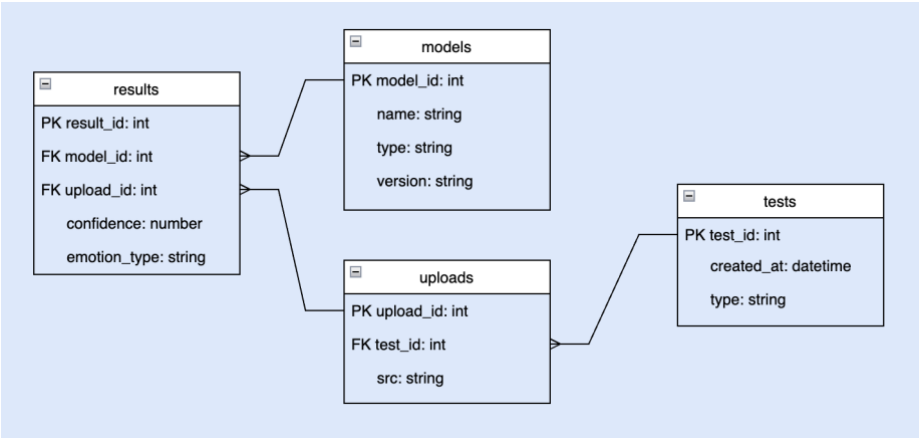

Рисунок 4.2 – Схема БД (рисунок створено самостійно)

Вона дозволяє наочно відобразити логічну модель бази даних, демонструючи основні сутності, їх атрибути та зв’язки. Схема бази даних слугуватиме важливим інструментом для аналізу й проектування бази даних, забезпечуючи структуроване уявлення про взаємодію між компонентами системи та сприяючи подальшій реалізації проекту.

Для кращого розуміння функціональних можливостей системи та сценаріїв взаємодії користувача з додатком була розроблена Smart Use Case діаграма (див. рис. 4.3):

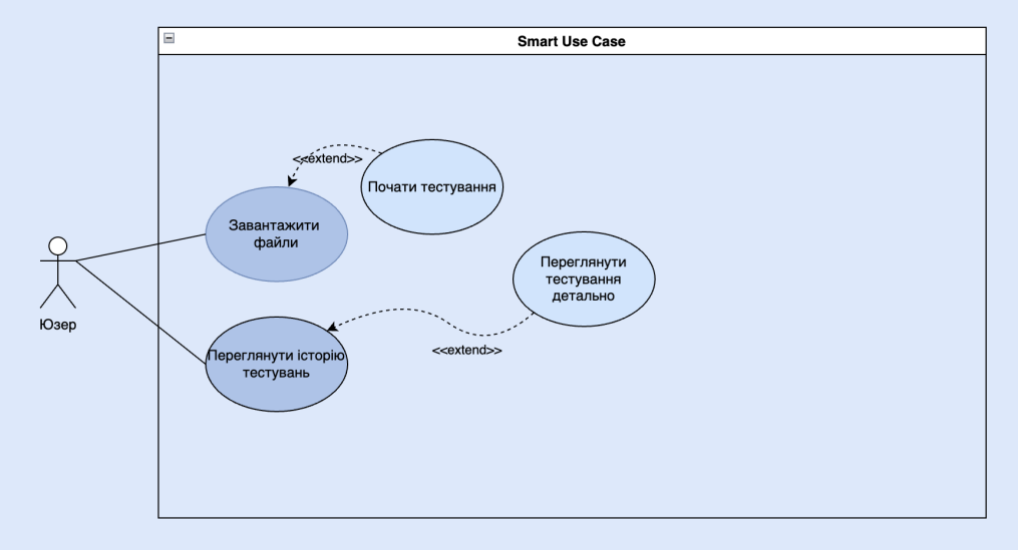

Рисунок 4.3 – Діаграма Smart Use Case (рисунок створено самостійно)

Ця діаграма наочно демонструє основні дії, які можуть виконувати користувачі, а також зв’язки між різними сценаріями використання.

Далі трохи детальніше розглянемо архітектури нейронних мереж які ми будемо застосувати для навчання наших моделей.

Першою архітектурою яку ми будемо застосувувати це CNN, вона є потужним інструментом для роботи з візуальними даними, включаючи зображення та відео. Її основна особливість полягає в автоматичному виділенні ознак із вхідних даних, що дозволяє зберігати просторову структуру зображень. Завдяки цьому CNN здатна навчатися ієрархічним ознакам, починаючи з базових, таких як краї та форми, і досягаючи високорівневих, наприклад складних об’єктів.

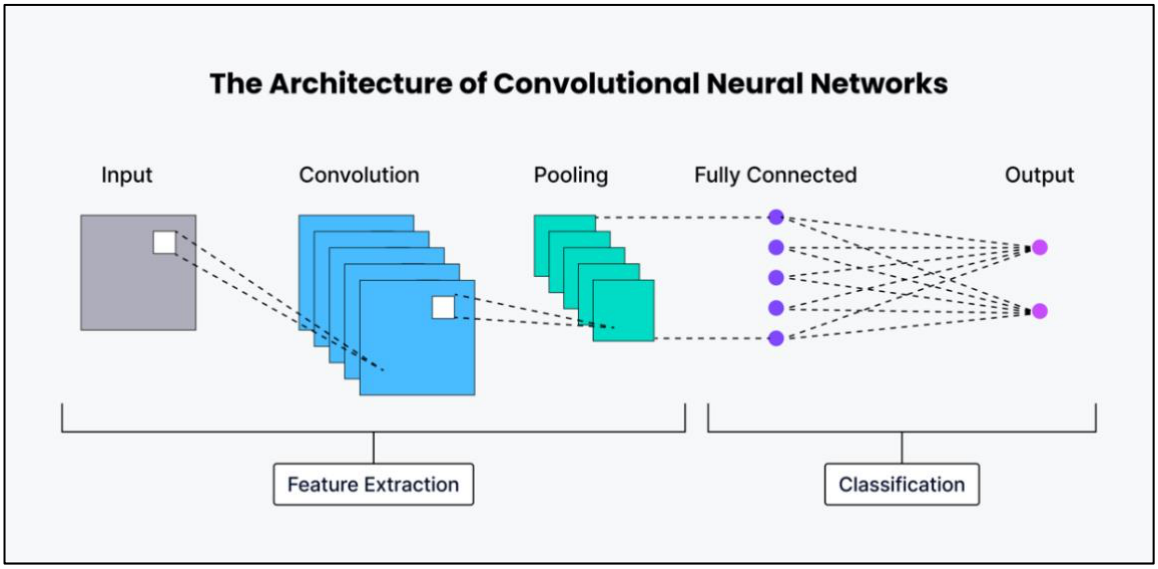

Рисунок 4.4 – Архітектура CNN (за даними [13])

Першим ключовим елементом CNN є згортковий шар, який виділяє основні ознаки на зображеннях. Для цього використовуються фільтри або ядра – невеликі матриці, які проходять через зображення, виділяючи локальні патерни, як-от краї, кути чи текстури. Щоб зберегти розмір зображення після згортки, часто застосовується padding – додавання рамки з нулями навколо зображення. На цьому етапі також використовується активаційна функція ReLU, яка вводить нелінійність, замінюючи всі негативні значення на нулі.

Наступним елементом є пулинговий шар, метою якого є зменшення розміру карти ознак (feature map) без втрати важливих характеристик. Це зменшує обчислювальну складність і водночас знижує ризик перенавчання, видаляючи несуттєві деталі. Найпоширенішими методами пулингу є max pooling, який бере максимальне значення в межах певної області, та average pooling, що розраховує середнє значення.

Багатошарові згорткові шари забезпечують поступовий перехід від базових ознак, таких як контури, до більш складних, наприклад форм об’єктів. Ця ієрархічна структура дозволяє мережі ефективно вивчати й комбінувати різні рівні ознак, створюючи унікальні можливості для розпізнавання складних об’єктів.

Повністю зв’язаний шар (Fully Connected Layer, FC) виступає фінальним етапом обробки. На цьому етапі дані перетворюються на вектор, що використовується для класифікації. Кожен нейрон цього шару пов’язаний із кожним нейроном попереднього шару. Вихід обробляється активаційними функціями, такими як softmax, які допомагають визначити ймовірність приналежності даних до певних класів. Завершує обробку вихідний шар, який генерує остаточний результат, наприклад, класифікацію об’єкта чи визначення його ймовірності.

Таким чином, архітектура CNN є складною та добре оптимізованою системою, що забезпечує автоматичне виділення ознак і класифікацію даних із високою точністю. Це робить її незамінною в завданнях аналізу візуальної інформації.

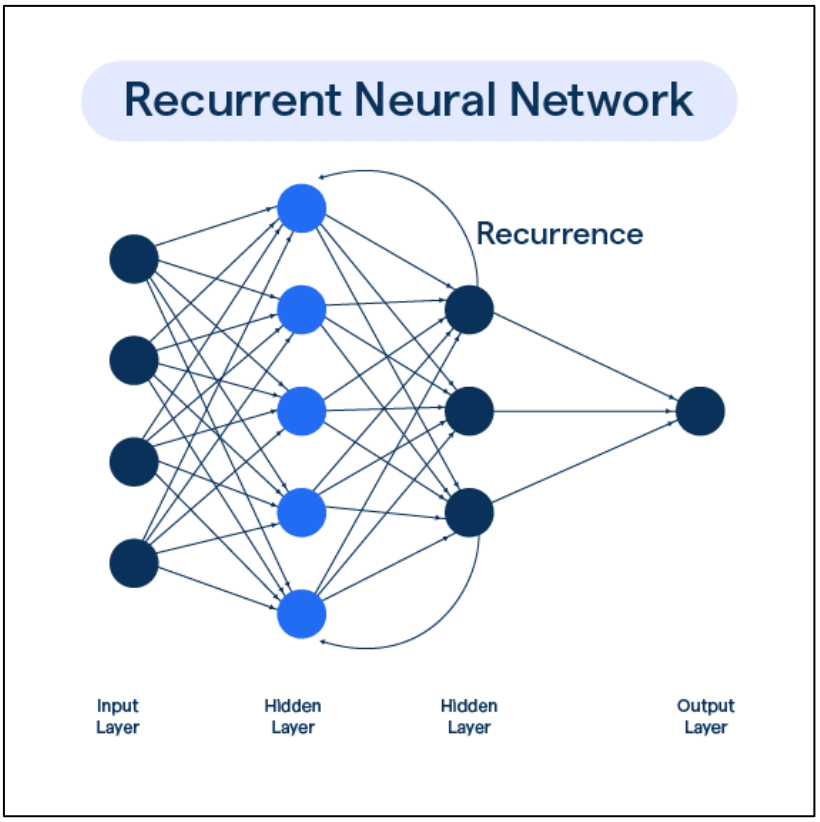

Рисунок 4.5 – Архітектура RNN (за даними [14])

Другою архітектурою яку ми будемо застосувати це рекурентні нейронні мережі (RNN), вони складаються з кількох основних компонентів, що забезпечують їх здатність обробляти послідовні дані. Центральним елементом є

рекурентні зв’язки, які дозволяють передавати інформацію між часовими кроками. Ці зв’язки утворюють внутрішню пам’ять мережі, яка зберігає стан і враховує контекст попередніх даних.

Основу архітектури становлять нейрони, організовані у шари. Кожен нейрон отримує вхідний сигнал як від поточного елемента послідовності, так і від попереднього стану. Для обчислення вихідного сигналу використовується активаційна функція (зазвичай sigmoid або tanh), що додає нелінійність до моделі. Вихід кожного шару оновлює стан пам’яті, що дозволяє враховувати залежності між елементами послідовності.

Процес навчання RNN базується на зворотному поширенні через час (Backpropagation Through Time, BPTT), яке використовується для оптимізації ваг мережі. Під час цього процесу обчислюються впливи попередніх станів на поточний вихід, а градієнти ваг передаються назад через часові кроки. Однак під час навчання можуть виникати проблеми, такі як зникаючий градієнт, коли сигнали стають занадто малими, або вибухаючий градієнт, коли значення ваг неконтрольовано зростають. Для вирішення цих проблем застосовують методи, як-от обрізання градієнтів.

Розширені архітектури, такі як LSTM (Long Short-Term Memory) і GRU (Gated Recurrent Unit), включають спеціальні механізми воріт (input, forget, output), які регулюють потік інформації між станами. Ці механізми дозволяють ефективно працювати з довготривалими залежностями, зберігаючи важливу інформацію протягом усього процесу обробки послідовності.

Таким чином, архітектура RNN забезпечує обробку послідовних даних за допомогою рекурентних зв’язків, внутрішньої пам’яті та спеціалізованих механізмів, що дає змогу вирішувати складні завдання з урахуванням контексту і залежностей між елементами послідовності.

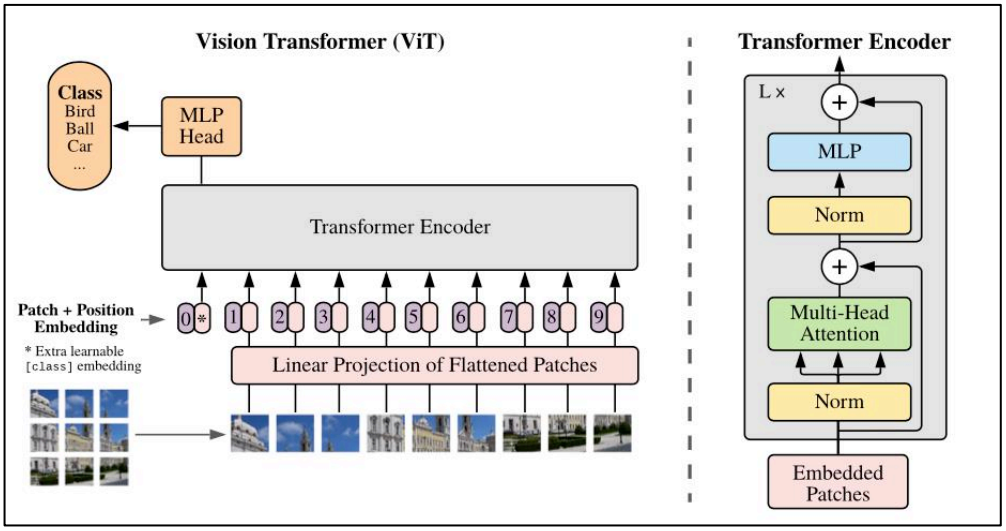

Рисунок 4.6 – Архітектура ViT (за даними [15])

На кінець третя архітектура яку будемо застосувати, це Vision Transformer (ViT) складається з кількох ключових компонентів, які забезпечують її здатність обробляти візуальні дані за допомогою механізмів самоуваги. Основою цієї архітектури є розбиття зображення на попередньо визначені патчі фіксованого розміру. Кожен патч перетворюється на вектор за допомогою лінійного шару для створення ембедингів, які зберігають локальні характеристики патчів.

До ембедингів додаються позиційні ембединги, які допомагають зберегти інформацію про просторове розташування кожного патча в межах зображення. Ця послідовність ембедингів подається на вхід трансформерному енкодеру, який є основним обчислювальним блоком моделі.

Трансформерний енкодер складається з багатоголових шарів самоуваги (Multi-Head Self-Attention), які обчислюють взаємозв'язки між усіма патчами, дозволяючи моделі аналізувати як локальні, так і глобальні залежності. Після кожного шару самоуваги результат обробляється через нелінійний feed-forward шар. Для стабільності навчання та покращення продуктивності в кожному блоці використовуються нормалізація шару (Layer Norm) та залишкові зв’язки (Residual Connections), які забезпечують проходження сигналу без затримок через нелінійні шари.

Вихідні дані останнього трансформерного блоку агрегуються, і спеціальний класифікаційний токен (CLS token) передається через повнозв’язний шар для

визначення кінцевого класу зображення. На етапі навчання модель оптимізується за допомогою великих наборів даних, після чого її адаптують (fine-tuning) до конкретних задач, таких як класифікація зображень, сегментація чи розпізнавання об’єктів. Таким чином, архітектура ViT базується на трансформерах, які опрацьовують послідовності ембедингів, а не традиційні піксельні масиви. Це дозволяє моделі аналізувати візуальні дані з меншою кількістю обчислень порівняно з класичними згортковими нейронними мережами, досягаючи високої точності на масштабних задачах комп’ютерного зору.

4.3 UI/UX дизайн системи

Дизайн проекту пройшов два етапи трансформації — від початкового прототипу до фінальної версії інтерфейсу, яка вдало поєднує функціональність і естетику, враховуючи потреби користувачів. На першому етапі було створено прототип (див. рис. 4.7), що демонстрував базову структуру продукту.

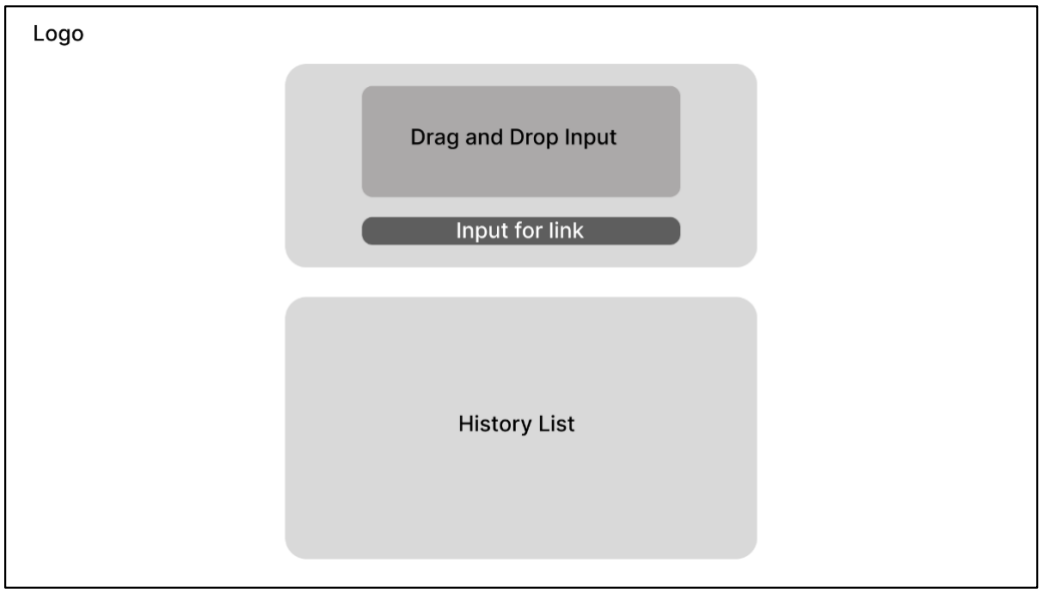

Рисунок 4.7 – Прототип дизайну (рисунок створено самостійно)

Його вигляд був максимально мінімалістичним, зосередженим виключно на ключових функціях. У верхньому лівому куті розташовувався логотип для ідентифікації бренду. Центральна частина містила зону для завантаження файлів, де можна було використати Drag & Drop або вставити посилання в окреме поле. Під цією областю знаходився список історії тестування. Ця початкова версія була

простою, функціональною, але позбавленою деталей, які б робили її привабливою для користувача. Фінальна ж версія інтерфейсу значно змінилася (див. рис. 4.8):

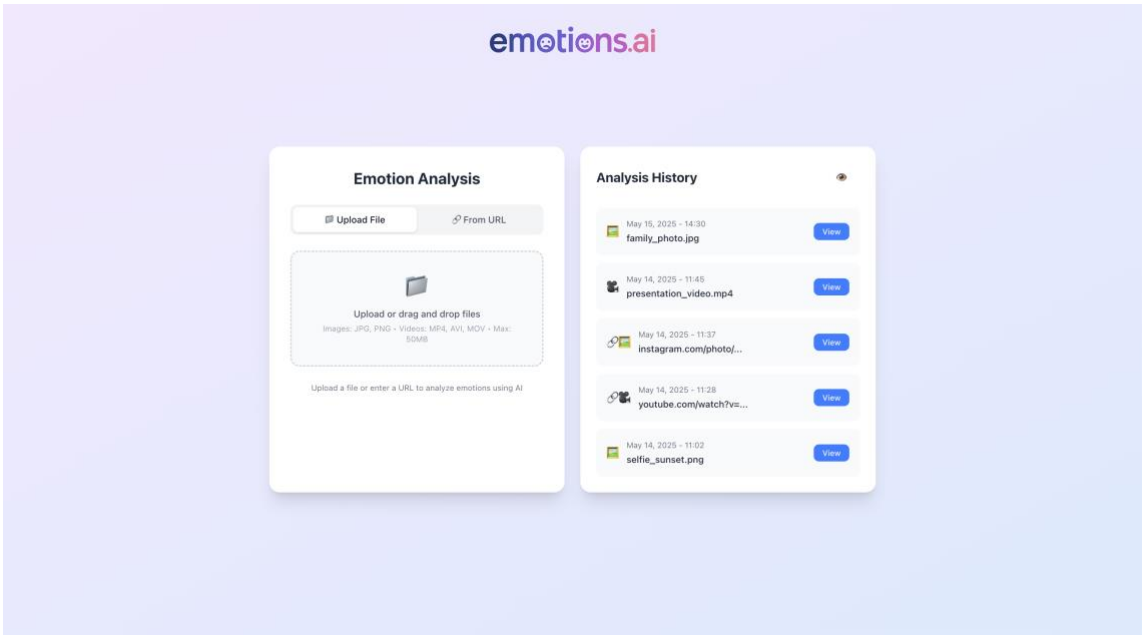

Рисунок 4.8 – Дизайн головної сторінки (рисунок створено самостійно)

Фінальна версія інтерфейсу значно змінилася. Логотип перемістився у центральну частину, але тепер він чіткіше й краще інтегрується в загальний стиль. Дизайн отримав легкий градієнт на фоні, що створює відчуття легкості та сучасності. Поле для файлів та поле для посилань було зроблено через вкладки для зручності.

Список історії аналізів також зазнав змін, він тепер відображається справа від поля завантаження файлів для того щоб все вміщалось на одному екрані. В самому списку ми можемо бачити відображення за допомогою іконок типу аналізованого файлу, назву, дату й час. Також є кнопка котра надає можливість більш детально переглянути аналіз.

Таким чином, еволюція дизайну показала, як із простого функціонального прототипу можна створити естетично привабливий і зручний інтерфейс. Використання сучасних принципів UX/UI дозволило перетворити ідею на завершений продукт, що відповідає очікуванням користувачів і виглядає професійно.

4.4 Методи оцінки продуктивності

Для оцінки ефективності різних архітектур нейронних мереж (CNN, RNN, ViT) у завданні розпізнавання емоцій за мімікою було розроблено детальний план тестування. Основна мета цього етапу – об'єктивно оцінити переваги й обмеження кожної моделі в контексті реальних умов використання, а також визначити оптимальне рішення для інтеграції у веб-додаток. Тестування проводитиметься за різними сценаріями, що враховують типові виклики для задач комп'ютерного зору.

Першим етапом дослідження є чітке визначення ключових метрик продуктивності, що дадуть змогу об'єктивно порівняти ефективність різних моделей. До основних показників належать: точність класифікації, площа під ROC-кривою (AUC), F1-міра як гармонійне середнє між точністю та повнотою, час інференсу, а також продуктивність у реальному часі, вимірювана кількістю зображень, оброблених за секунду. Для забезпечення стандартизованості експериментів та достовірності результатів буде використано набір даних із попередньо розміченими емоційними станами.

Сценарії тестування охоплюватимуть кілька ключових аспектів. Для забезпечення різноманітності вхідних даних і більш повної оцінки моделей будуть використані три різні датасети: FER-2013, RAF-DB, не повний AffectNet та повний AffectNet. Вони суттєво відрізняються за якістю, кількістю зразків, ступенем анотації та умовами зйомки, що дозволяє перевірити здатність моделей працювати в різних умовах. Наприклад, FER-2013 містить низькороздільні чорно-білі зображення з великою кількістю шуму, що імітує реальні обмеження пристроїв, RAF-DB забезпечує вищу якість і баланс класів, а AffectNet представляє наймасштабніший набір із реалістичними зображеннями з Інтернету, які охоплюють широкий спектр емоцій. Такий підхід дозволяє протестувати CNN на задачах класифікації статичних зображень з різним освітленням і ракурсами, оцінити здатність RNN враховувати динаміку міміки при обробці послідовностей, а також перевірити ефективність Vision Transformers на складних, високодеталізованих зображеннях у масштабних сценаріях.

Для збору даних про продуктивність моделей буде використано набір інструментів, побудованих на базі бібліотеки TensorFlow, яка надає широкі можливості для моніторингу та аналізу роботи нейронних мереж.

Експерименти здійснювались у контрольованих умовах, що забезпечують відтворюваність та об’єктивність результатів. Усі тестування проводились на одному пристрої — MacBook Pro 16” (2021) з чіпом Apple M1 Pro та 16 ГБ оперативної пам’яті. Для кожного тестового сценарію було виконано кілька повторних запусків, після чого обчислювались середні значення основних метрик, що дозволяє зменшити вплив випадкових флуктуацій у продуктивності. Отримані результати будуть представлені у вигляді наочних графіків і структурованих таблиць, що сприятиме зручному порівнянню моделей та глибшому аналізу їхньої ефективності.

Також буде розроблено веб застосунок де наочно можна буде завантажити власне зображення чи відео, та протестувати точність всіх навчених моделей. Це дозволить протестувати моделі у реальних умовах. Зберігання даних тестувань які були проведені у веб застосунку організоване у базі даних PostgreSQL, яка містить інформацію про конфігурації моделей, параметри запуску, результати експериментів і отримані метрики.

Структура бази даних дозволяє легко отримувати доступ до історичних даних для порівняння різних підходів. Таким чином, розроблена методика тестування й аналізу продуктивності забезпечить об’єктивну оцінку ефективності CNN, RNN та ViT, а також дозволить визначити їхні сильні сторони й обмеження для впровадження у веб-додатки, що працюють у реальному часі.

5 ПРАКТИЧНЕ ДОСЛІДЖЕНЯ

5.1 Опис проведення дослідження

Для початку я налаштовував CNN для розпізнавання емоцій на основі зображень з набору даних FER2013. З огляду на те, що зображення в цьому датасеті мають роздільну здатність 48 на 48 пікселів і представлені у відтінках сірого, я задав відповідну форму входу моделі, а саме 48 на 48 пікселів з одним каналом.

Перед початком навчання я використав генератори даних з аугментацією. Це дозволило розширити обсяг навчального набору штучно, застосовуючи різноманітні перетворення до зображень: обертання, масштабування, горизонтальне відображення, зсуви, зміну яскравості тощо. Для валідаційного набору я застосував лише нормалізацію значень пікселів.

Архітектуру мережі я побудував з урахуванням сучасних підходів, додавши залишкові блоки (residual blocks) з SE-модулями (Squeeze-and-Excitation), які покликані автоматично переналаштовувати вагу кожного каналу, підвищуючи значущість інформативних ознак. Кожен залишковий блок складався з двох згорткових шарів з регуляризацією та батч-нормалізацією, а також шару Dropout для зниження ризику перенавчання. На виході з кожного блоку застосовувався SE-механізм і підсемплювання через MaxPooling.

Після кількох таких блоків з поступовим збільшенням кількості фільтрів (від 32 до 256) я додав глобальний пулінг, повнозв’язний шар з 512 нейронів, Dropout, і на виході — шар softmax для класифікації на відповідну кількість класів емоцій.

Я компілював модель з функцією втрат CategoricalCrossentropy, в якій було увімкнено label smoothing для зменшення впевненості моделі в помилкових передбаченнях. Оптимізатором виступав Adam зі швидкістю навчання 0.001. Під час тренування використовувалися кілька callback-ів: збереження найкращої моделі, раннє зупинення при відсутності покращення, зменшення швидкості

навчання, логування в TensorBoard, а також власний callback для обчислення F1-міри, ROC-AUC та продуктивності моделі.

Модель навчалася протягом 50 епох, а результати навчання, зокрема історія метрик, зберігалися для подальшого аналізу. У фіналі тренування модель зберігалася у вигляді HDF5-файлу, а також виконувалася оцінка її якості на валідаційному наборі. Результат можна побачити в таблиці (див. табл. 5.1):

Таблиця 5.1 – результати тренування CNN архітектури на FER2013 дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	64.47
AUC	0.9232
F1-score	0.6343
Latency (ms)	1.23
Throughput (images/s)	811.22

Далі я перейшов до дата сету RAF-DB, налаштування були трохи змінені так як RAF-DB складається з картинок з вищою роздільною здатністю, а саме 100 на 100 пікселів, тож був замінений вхідний розмір на 100 пікселів. Інші налаштування були збережені для об’єктивності порівняння. В результаті було отримано наступні результати (див. табл. 5.2):

Таблиця 5.2 – результати тренування CNN архітектури на RAF-DB дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	70.53
AUC	0.9307
F1-score	0.7026
Latency (ms)	1.13
Throughput (images/s)	885.01

Далі я перейшов до дата сету AffectNet, для початку я взяв не повну версію яка налічує 30 тисяч зображень. Налаштування були збережені з минулого дата сету для об’єктивності дослідження. В результаті були отриманні наступні показники (див. табл. 5.3):

Таблиця 5.3 – результати тренування CNN архітектури на AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	54.17
AUC	0.8744
F1-score	0.5399
Latency (ms)	1.22
Throughput (images/s)	819.47

Після цього для більш об’єктивного дослідження, щоб зрозуміти як архітектури ведуть себе на великих дата сетах, я взяв повну версію AffectNet дата сету яка налічує 420 тисяч зображень. Проте так як ми розпізнаємо тільки 7 емоцій які є і в інших дата сетах, то кількість зображень зменшується до 287 тисяч але все одно це набагато більше ніж в дата сетах які ми використовували раніше. Результатами навчання ми отримали наступні значення (див. табл. 5.4):

Таблиця 5.4 – результати тренування CNN архітектури на повному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	56.73
AUC	0.8862
F1-score	0.5602
Latency (ms)	1.23
Throughput (images/s)	817.63

Після навчання CNN архітектури на різних дата сетах, я перейшов до навчання RNN архітектури. Перед початком навчання я використав генератори даних з аугментацією. Для навчального набору застосовувалися різноманітні перетворення, зокрема обертання, масштабування, зсуви, горизонтальне відображення, зміна яскравості та зсуви по ширині/висоті. Це дозволило підвищити узагальнюючу здатність моделі та знизити ризик перенавчання. Для валідаційного набору застосовувалась лише нормалізація значень пікселів у діапазон [0, 1].

Архітектура моделі поєднує згорткові та рекурентні компоненти. На початку я побудував два згорткові блоки з фільтрами розміром 32 та 64. Кожен

блок містив згортковий шар з активацією ReLU, батч-нормалізацію, підсемплювання через MaxPooling2D та Dropout для регуляризації. Така комбінація дозволила витягнути просторові ознаки із зображень розміром 48 на 48 пікселів.

Після згорткових блоків було виконано перетворення тензора у форму, придатну для обробки рекурентною мережею: розмірність (12, 12 на 64). Це дозволило інтерпретувати просторові ознаки як послідовність векторів ознак. Далі модель переходить до рекурентної частини, побудованої на двох шарах LSTM. Обидва шари є двонаправленими (Bidirectional), що дозволяє враховувати контекст як зліва направо, так і справа наліво. Перший LSTM має return_sequences=True, дозволяючи другому шару обробити повну послідовність. Другий шар повертає лише останній стан, що далі подається до повнозв'язаних шарів. Для покращення регуляризації після кожного LSTM-шару також додано Dropout.

Після рекурентної частини йде повнозв'язний шар з 256 нейронів, BatchNormalization та Dropout для уникнення перенавчання. На виході моделі — softmax-шар, що забезпечує багатокласову класифікацію на основі ймовірностей.

Я компілював модель з функцією втрат CategoricalCrossentropy з активованим label smoothing, що допомагає уникати надмірної впевненості моделі в своїх помилкових передбаченнях. В якості оптимізатора використовувався Adam із змінною швидкістю навчання за допомогою експоненційного спадання. Навчання проводилося протягом 50 епох з використанням batch size 32.

Отже після тестування моделі RNN на FER2013 дата сеті, я отримав наступні результати (див. табл. 5.5):

Таблиця 5.5 – результати тренування RNN архітектури на FER2013 дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	45.01
AUC	0.8178
F1-score	0.4225
Latency (ms)	4.11
Throughput (images/s)	243.25

Після цього як і в минулій архітектурі я перейшов до RAF-DB дата сету зберігши налаштування архітектурі для максимальної об’єктивності оцінювання, отримавши наступні результати (див. табл. 5.6):

Таблиця 5.6 – результати тренування RNN архітектури на RAF-DB дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	46.48
AUC	0.8203
F1-score	0.3808
Latency (ms)	4.24
Throughput (images/s)	235.69

Ну і на кінець було проведено навчання на неповному AffectNet (див. табл. 5.7), та на повній версії дата сету AffectNet (див. табл. 5.8) задля того щоб побачити як веде себе архітектура на більш масштабній версії дата сету.

Таблиця 5.7 – результати тренування RNN архітектури на неповному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	33.73
AUC	0.7296
F1-score	0.2961
Latency (ms)	4.09
Throughput (images/s)	244.86

Таблиця 5.8 – результати тренування RNN архітектури на повному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	35.93
AUC	0.7412
F1-score	0.3105
Latency (ms)	4.01
Throughput (images/s)	249.24

Після успішного навчання RNN архітектури на всіх наявних дата сетах, я перейшов до ViT архітектури. На першому етапі модель розбивала вхідне зображення на непересічні патчі розміром 6 на 6 пікселів, кожен з яких

векторизувався та подавався на проєкційний шар. Після цього я додав позиційні ембеддинги, щоб модель могла враховувати просторовий порядок патчів — оскільки Transformer сам по собі є порядконезалежним.

Далі йшла основна частина архітектури — стек з чотирьох блоків Transformer. Кожен блок складався з багатоголової механізму уваги (Multi-Head Attention), шару нормалізації та двошарової MLP-голови з регуляризацією (Dropout і L2). Після кожного шару додавались залишкові зв'язки (residual connections), що допомагає уникати затухання градієнтів і стабілізує навчання. Для глибших шарів я використовував двоступеневу MLP-структуру з активацією ReLU та зменшенням розмірності.

На виході трансформер-блоків виконувався глобальний pooling по часовій осі (GlobalAveragePooling1D), що дозволило зібрати узагальнене представлення з усіх патчів. Далі я додав MLP-голову з одним повнозв'язним шаром і Dropout для остаточної обробки ознак. Кінцевий шар softmax забезпечував класифікацію на відповідну кількість класів емоцій.

Для оптимізації як і в минулих випадках я використав Adam з експоненційним зменшенням швидкості навчання. Першим дата сетом на якому я навчив ViT був FER2013 і в результаті я отримав наступне (див. табл. 5.9):

Таблиця 5.9 – результати тренування ViT архітектури на FER2013 дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	31.60
AUC	0.7156
F1-score	0.2475
Latency (ms)	2.58
Throughput (images/s)	387.89

Далі як і у випадку з минулими архітектурами зберігши налаштування архітектури для об'єктивності я перейшов до RAF-DB дата сету отримавши наступні результати (див. табл. 5.10):

Таблиця 5.10 – результати тренування ViT архітектури на RAF-DB дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	50.03
AUC	0.8394
F1-score	0.4346
Latency (ms)	2.94
Throughput (images/s)	340.44

Ну і на кінець я навчив модель на не повному дата сеті AffectNet (див. табл. 5.11) та повному дата сеті AffectNet (див. табл. 5.12) задля того щоб подивитись як себе веде модель на великому наборі даних.

Таблиця 5.11 – результати тренування ViT архітектури на неповному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	19.90
AUC	0.5762
F1-score	0.1509
Latency (ms)	2.90
Throughput (images/s)	344.82

Таблиця 5.12 – результати тренування ViT архітектури на повному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	23.51
AUC	0.5803
F1-score	0.1117
Latency (ms)	2.91
Throughput (images/s)	343.5

Після тестування всіх моделей на всіх дата сетах було вирішено провести додаткове тестування на предтренованих моделях. В якості дата сету на якому будемо до навчати обрали неповний AffectNet так як на цьому дата сеті були найгірші результати у всіх моделей і цікаво як вони себе поведуть якщо будемо навчати їх не з нуля. Першою для тестування обрали ViT, для до навчання обрали google/vit-base-patch16-224-in21k модель, вона навчена на ImageNet який налічує приблизно 14 мільйонів зображень, при розмірі патчу 16 та вхідному розмірі

картинок 224 та має ViT архітектуру як нам і треба. В результаті я отримав наступне (див. табл. 5.13):

Таблиця 5.13 – результати тренування предтренованої ViT архітектури на неповному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	52.38
AUC	0.8724
F1-score	0.5162
Latency (ms)	34.70
Throughput (images/s)	342.1

Далі я перейшов до CNN архітектури, тут в якості для до навчання була обрана модель ResNet50, ця модель навчена так само на ImageNet, має такий самий розмір патчів та вхідний розмір картинок, проте має іншу архітектуру а саме CNN як нам і треба. Тож в результаті ми отримали такі значення метрик (див. табл. 5.14):

Таблиця 5.14 – результати тренування предтренованої CNN архітектури на неповному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	34.99
AUC	0.7627
F1-score	0.3310
Latency (ms)	11.02
Throughput (images/s)	191.69

Ну і на кінець ми взяли протестувати архітектуру RNN. Так само як і раніше модель побудована як гібридна CNN – RNN, де в якості екстрактора ознак використовується попередньо натренована ResNet50 яка були використана для CNN, а вихідний тензор розглядається як послідовність фіксованої довжини, що подається до двох шарів BiLSTM. Всі налаштування були збережені, задля збереження об’єктивності дослідження, результатами до навчання були наступні значення метрик (див. табл. 5.15):

Таблиця 5.15 – результати тренування предтренованої RNN архітектури на неповному AffectNet дата сеті (таблиця виконана самостійно)

Метрика	Значення
Accuracy (%)	22.56
AUC	0.6275
F1-score	0.1825
Latency (ms)	8.72
Throughput (images/s)	114.67

Після проведених досліджень, моделі були збережі для подальшого їх використання у веб-застосунку.

5.2 Аналіз результатів досліджень

Навчивши CNN для розпізнавання емоцій на дата сеті FER2013, ми можемо побачити, що під час аналізу матриці помилок (див. рис. 5.1) видно, що клас happy розпізнається доволі добре, значна частина зразків потрапляє в діагональ, що відповідає цьому класу, і лише невелика частина зміщується в інші категорії.

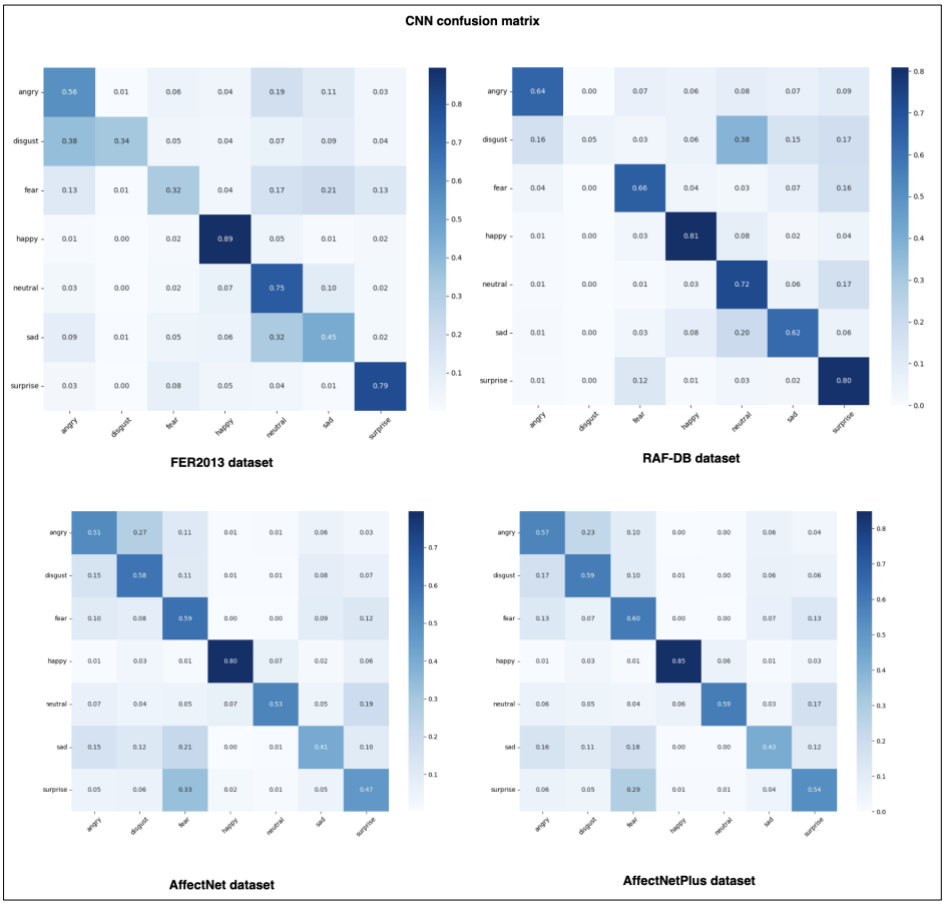

Рисунок 5.1 – Матриця помилок CNN (рисунок створено самостійно)

Це пояснюється тим, що усмішка є часто вираженою й легко виявляється згортковою мережею. Водночас класи fear і sad демонструють помітні змішання, а саме частина зображень зі страхом розпізнається як surprise чи neutral, а вираз суму іноді інтерпретується як нейтральний або навіть як розгнівлений. Ці спостереження є типовими труднощами які має дата сет FER2013, низька роздільна здатність 48 на 48 робить тонкі зміни в міміці менш виразними, тож нейронна мережа іноді не може чітко відокремити fear від surprise чи sad від neutral. У конфузійній матриці помітно також плутанину між angry і disgust, що теж не дивно, оскільки обидва вирази часто супроводжуються напруженими рисами обличчя. Загалом аналіз матриці підтверджує значення точності 64.5% та високий AUC 0.9232, та вказує нам на слабкі місця такі як класи з менш однозначними міміками.

Переходячи до RAF-DB, де зображення мають роздільну здатність 100×100, зберігаючи решту налаштувань CNN, я побачив у таблиці (див. табл. 5.2) зростання точності до 70.53 відсотків і AUC 0.9307. Матриця помилок демонструє більш виразні діагоналі: happy й surprise з високо відсотковим розпізнаванням, neutral також чіткіше відділений. Водночас fear тепер менше плутається з neutral і частіше правильно класифікується, що можна пояснити тим, що більше пікселів дає змогу мережі помітити тонкі деталі в очах або губах. Однак клас disgust усе ще дещо має похибки він іноді змішується з angry чи neutral, хоча вищий розмір зображення дає нейронній мережі більше шансів помічати відмінності, але, мабуть, у RAF-DB цих прикладів недостатньо, або вони менш послідовні за виразом. Матриця помилок показує, що мережа стала впевненіше розпізнавати ключові категорії, але тонкі емоції все ще викликають складнощі.

Не повна версія AffectNet з 30 тисячами зображень показала зниження показників нейронної мережі CNN, accuracy став близько 54.17%, AUC 0.8744. Матриця помилок для цього випадку ілюструє набагато більш розмите розподілення по класах: хоча happy й neutral досі знаходять себе в діагоналі частіше, частка помилок значно більша, а класи angry, disgust, fear, sad мають низьку діагональну частку. Часто бачимо, як нейронна мережа перекидає

складніші емоції в neutral або happy, бо ці класи домінують у тренувальному наборі. Це вказує, що різноманітність умов (освітлення, різні похили голови, варіації виразу) у AffectNet які ускладнюють навчання, а саме мережа іноді пристосовується до загальних ознак, коли тонкі прояви конкретних емоцій можуть губитися.

Коли я збільшив обсяг даних до повної версії дата сету AffectNet, це приблизно 287 тис. для 7 класів емоцій, спостерігалося лише помірне зростання accuracy до 56.73% та AUC до 0.8862. Матриця помилок демонструє трохи кращу роздільність для деяких класів, наприклад fear та sad трохи рідше плутаються з іншими типами емоцій, але загалом проблема з неоднорідністю даних лишається. Навіть великий обсяг не компенсував повністю складність різноманітних прикладів. Швидкість обробки залишилася на рівні близько 1.2 мс latency та 820 зображень на секунду, що підтверджує ефективність архітектури, проте показує: мережа встигає обробити багато даних, але якість ознак обмежена самою природою набору.

Після розгляду CNN я розглянемо результати RNN архітектури на тих самих наборах. На FER2013 accuracy впала до 45.01%, а у випадку AUC до 0.8178. Матриця помилок (див. рис. 5.2) містить менш виразну діагональ, happy розпізнається відносно краще але не так впевнено, як у CNN.

Емоції типу angry, disgust, fear часто змішуються, іноді підпадають під neutral чи happy. Це свідчить, що рекурентна частина, яка обробляє послідовність патчів, не додає цінної інформації для статичних зображень з роздільною здатністю 48 на 48 пікселів. RNN-частина, можливо, додає складнощі інтерпретації просторових відносин, коли просторові зв’язки вже добре виловлюються згортками. Latency близько 4.11 мс і throughput 243.25 img/s підтверджує, що модель значно повільніша, ніж CNN, але при цьому ще й точність набагато нижча.

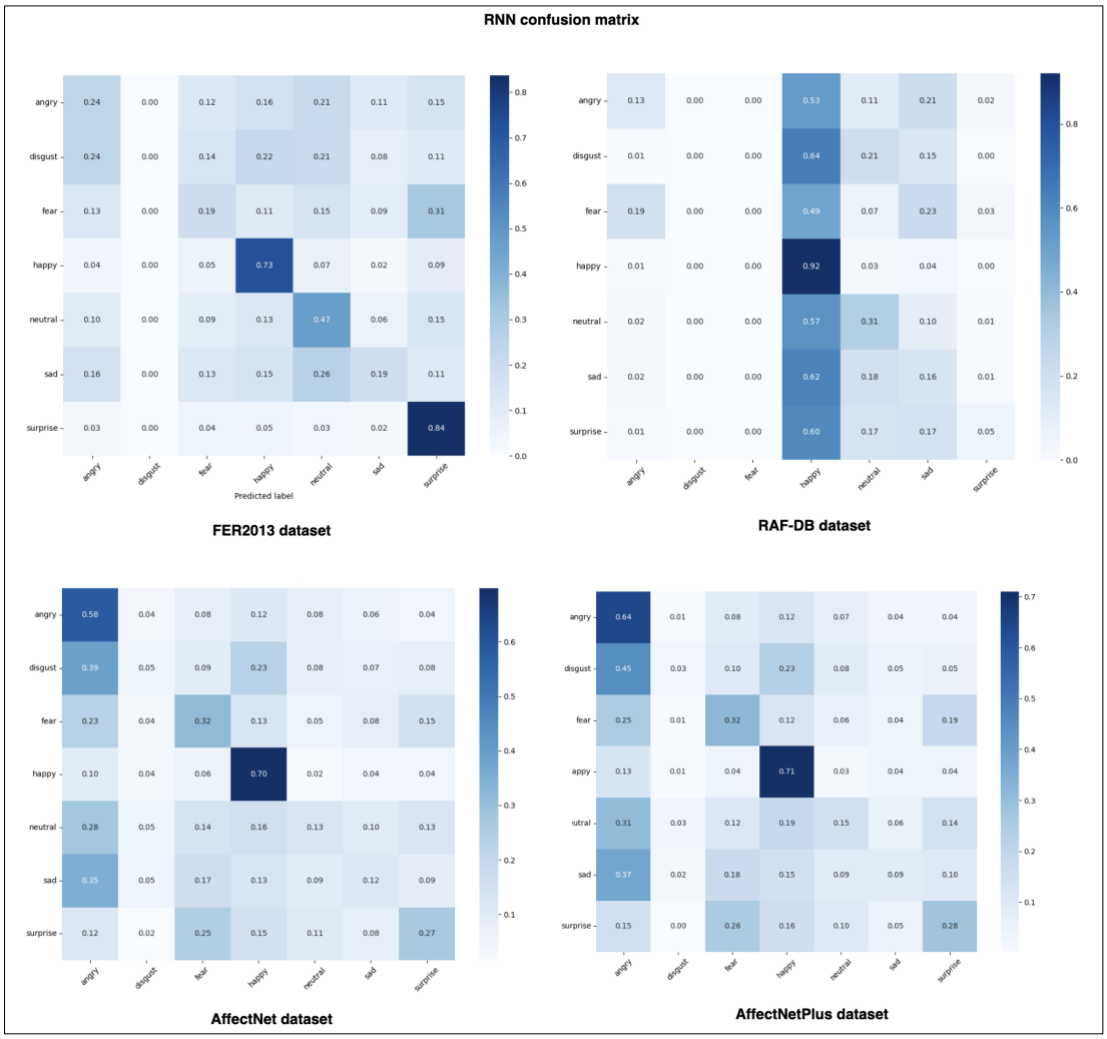

Рисунок 5.2 – Матриця помилок RNN (рисунок створено самостійно)

На дата сеті RAF-DB архітектура RNN дає accuracy 46.48%, у випадку AUC дає 0.8203. Матриця помилок демонструє, що happy та surprise іноді все ще виловлюються краще, але менш впевнено, ніж у CNN, а частина зразків happy може потрапляти в neutral або angry. Емоції disgust та fear плутаються ще частіше, хоча вищі розміри зображень могли би дати більше деталей проте архітектурні особливості заважають архітнеткрі RNN чітко витягати ці ознаки. Latency і throughput підкріплюють тенденцію, ми бачимо що модель значно важча в обчисленні.

Для дата сету AffectNet RNN на неповному наборі дає accuracy 33.73%, AUC 0.7296, а на повному 35.93%, AUC 0.7412. Матриці помилок у цих випадках не вираженні, діагоналі майже непомітні, крім happy та surprise, але навіть вони мають низькі показники порівняно з CNN. Модель часто віддає перевагу

домінантним класам, інколи хаотично переміщує інші емоції туди. Це свідчить, що RNN архітектура не витягує достатньо інформативних характеристик із просторових ознак для складних наборів з великою різноманітністю.

Тепер проаналізуємо останню архітектуру нейронних мереж, а саме ViT. На FER2013 вона показала accuracy 31.60% та AUC 0.7156, F1 у цьому випадку доволі низьке, latency близько 2.6 мс, throughput близько 388 img/s. Матриця помилок (див. рис. 5.3) демонструє домінування happy та neutral як пастки для більшості інших класів, а діагоналі для angry, disgust, fear, sad майже не виразні. Це пояснюється тим, що малий розмір зображень та відсутність попереднього великого предтренування призводять до того, що патчі містять недостатньо корисного сигналу, Transformer просто не може вловити важливих залежностей.

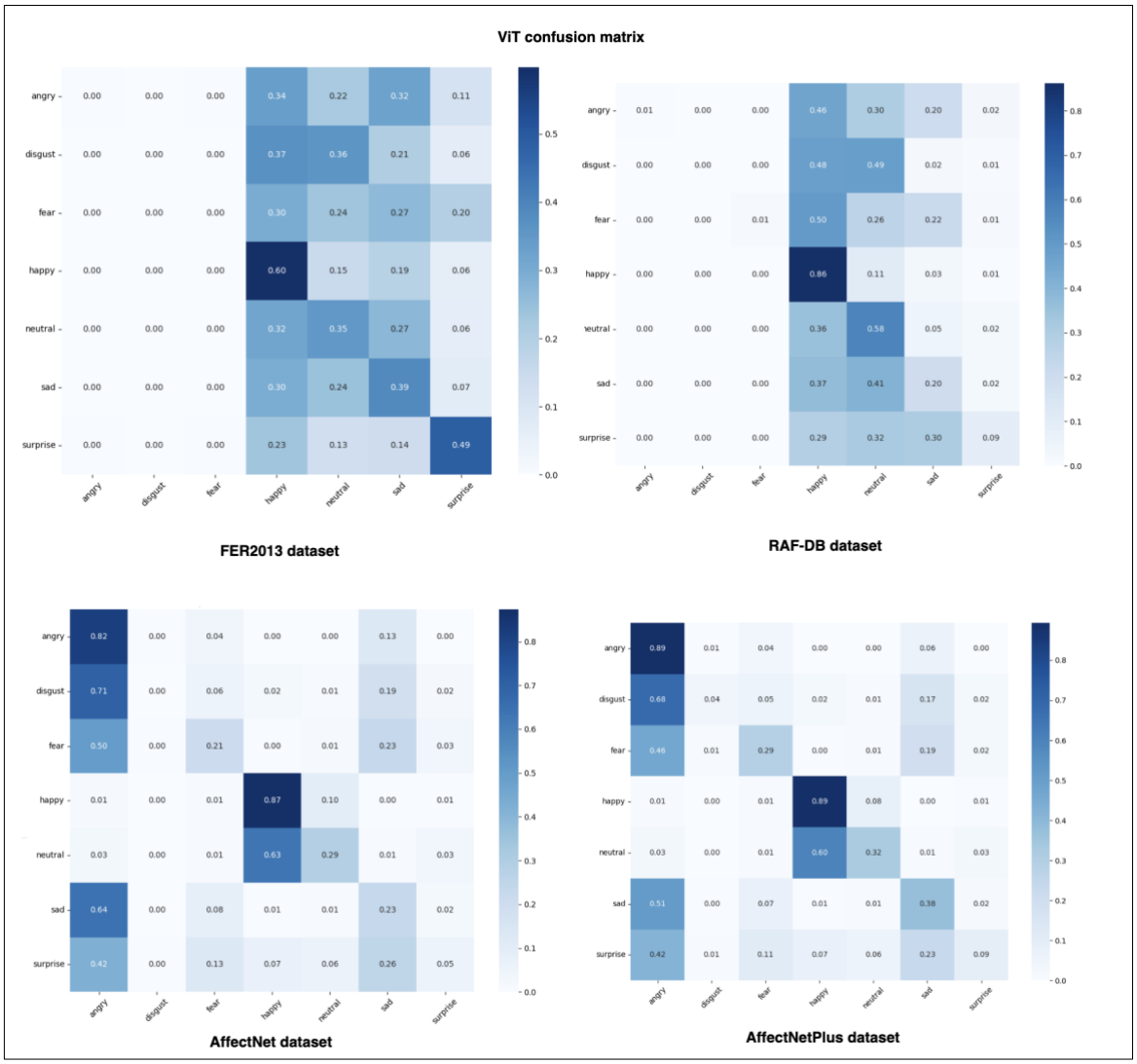

Рисунок 5.3 – Матриця помилок ViT (рисунок створено самостійно)

На RAF-DB результати трохи кращі, а саме accuracy 50.03%, AUC 0.8394. Матриця помилок, демонструє, що happy та surprise розпізнаються краще, деяке поліпшення в емоції neutral, але все ще слабке розпізнавання типів емоцій таких як disgust, fear, sad. Хоч розмір зображень більший, обсяг даних недостатній для повноцінного навчання ViT з нуля, тому модель частково вловлює базові ознаки, але нижче за CNN. На неповному дата сеті AffectNet архітектура ViT показує accuracy 19.90%, AUC 0.5762, а на повному 23.51%, AUC 0.5803. Матриці помилок має не чітку структуру, модель хаотично класифікує багато зразків в кілька домінантних класів, діагоналі майже непомітні. Навіть значний обсяг даних не дав ViT можливості встановити стабільні патерни, оскільки для Transformer потрібні хоча б або більші зображення, або великі перетреновані ваги.

Результати використання предтренованих моделей на дата сеті неповному AffectNet демонструють кардинально іншу картину порівняно з навчанням архітектур з нуля. Найбільш вражаючі покращення спостерігаються у архітектури ViT, де використання google/vit-base-patch16-224-in21k моделі, предтренованої на ImageNet, призвело до великого зростання точності з 19.90% до 52.38%. AUC також значно покращився з 0.5762 до 0.8724, а F1-score зріс з 0.1509 до 0.5162.

Матриця помилок предтренованої ViT моделі (див. рис. 5.4) демонструє кардинально покращену структуру порівняно з навчанням з нуля. Діагональні елементи стали значно більш вираженими, особливо для класів happy, surprise та neutral. Клас happy тепер розпізнається з високою впевненістю, а емоції fear та sad, які раніше майже повністю губилися серед інших класів, почали чіткіше виділятися. Проте складні емоції типу disgust та angry все ще демонструють деяку плутанину між собою, що свідчить про те, що навіть потужне предтренування не може повністю розв'язати проблему тонких відмінностей між близькими за виразом емоціями.

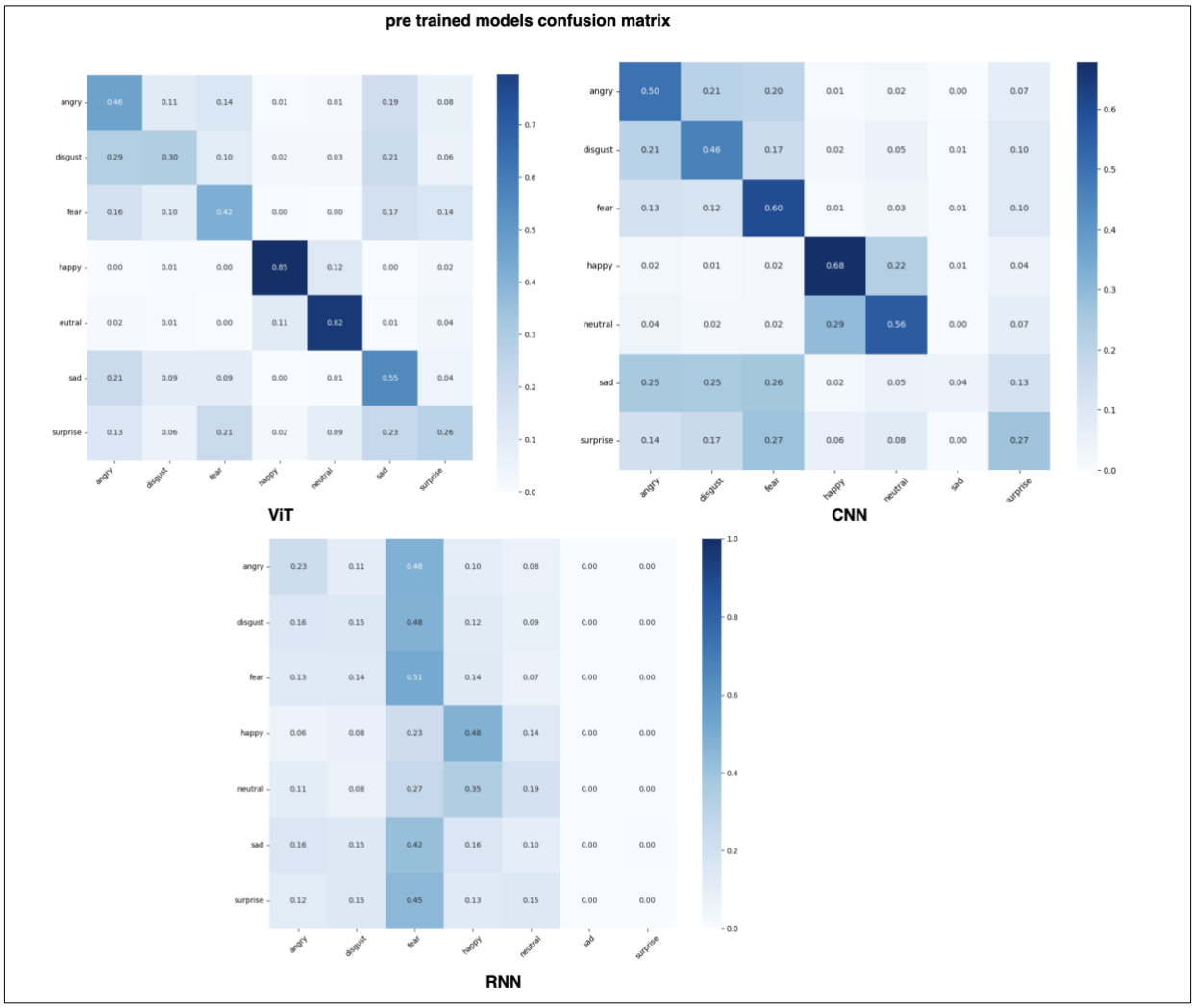

Рисунок 5.4 – Матриця помилок до навчених моделей (рисунок створено самостійно)

Предтренована CNN архітектура на основі ResNet50 показала гірші результати. Точність зменшилась з 54.17% (для CNN архітектури навченої з нуля) до тільки 34.99% для предтренованої ResNet50, що є несподіваним результатом. AUC знизився з 0.8744 до 0.7627, а F1-score з 0.5399 до 0.3310. Ці результати свідчать про те, що архітектура ResNet50, попри своє предтренування на ImageNet, виявилася менш адаптованою до завдання розпізнавання емоцій порівняно зі натренованою CNN архітектурою з нуля. Матриця помилок предтренованої CNN моделі демонструє менш чітку структуру порівняно з власною архітектурою. Хоча деякі класи, такі як happy та neutral, все ще розпізнаються відносно добре, загальна якість класифікації знизилася. Це може

пояснюватися тим, що ResNet50, розроблена для загальних завдань комп'ютерного зору, не має спеціалізованих компонентів для роботи з емоціями обличчя, які були впроваджені у власну архітектуру.

Найгірші результати показала предтренована RNN архітектура, де використання ResNet50 як екстрактора ознак з подальшою обробкою через BiLSTM призвело до accuracy лише 22.56%, AUC 0.6275 та F1-score 0.1825. Ці показники є навіть нижчими за власну RNN архітектуру, навчену з нуля. Матриця помилок демонструює майже повну відсутність структури та більшість зразків класифікуються хаотично, з домінуванням кількох класів.

Порівнюючи результати предтренованих моделей з архітектурами, навченими з нуля, можна зробити кілька важливих висновків. По-перше, предтренування є критично важливим для ViT архітектури, без нього вона показує неприйнятно низькі результати, але з предтренуванням стає конкурентоспроможною альтернативою CNN. По-друге, не всі предтреновані моделі автоматично забезпечують кращі результати успіх залежить від відповідності архітектури специфіці завдання.

Цікавим спостереженням є те, що CNN архітектура навчена з нуля, перевершила стандартну ResNet50, попри відсутність предтренування. Це підкреслює важливість архітектурних рішень, спрямованих на конкретне завдання, та демонструє, що предтренування не завжди компенсує архітектурні недоліки.

Результати також показують різний вплив предтренування на різні архітектури: ViT отримала максимальну користь, CNN та RNN не змогла ефективно використати переваги предтренованого екстрактора ознак. Це свідчить про те, що вибір стратегії предтренування повинен враховувати особливості архітектури та специфіку завдання.

5.3 Висновок та рекомендації

Отже на основі проведеного експериментального дослідження та детального аналізу результатів, можна зробити низку висновків щодо ефективності

архітектур CNN, RNN та ViT для задачі розпізнавання емоцій за зображеннями обличчя.

Найвищі результати за всіма основними метриками (Accuracy, AUC, F1-score), а також за швидкістю обробки (Latency, Throughput) стабільно продемонструвала архітектура CNN. Вона ефективно витягує просторові ознаки, особливо на зображеннях з високою роздільною здатністю, таких як у RAF-DB. Застосування залишкових блоків, SE-модулів та аугментації дозволило моделі досягти високої точності навіть на більш складних та нерівномірних наборах, як AffectNet. Проте на наборах із великою варіативністю умов (освітлення, поза голови), мережа все ще демонструвала труднощі у класифікації менш очевидних емоцій, зокрема fear та disgust.

RNN архітектура, яка поєднувала згорткові шари з двонаправленими LSTM, не показала суттєвих переваг у задачі класифікації статичних зображень. Усі результати (особливо на AffectNet) були гіршими за CNN, з помітно нижчою точністю, повільнішою швидкістю та менш чіткими матрицями помилок. Це свідчить, що рекурентні моделі є неефективними для обробки зображень без часової динаміки, а їх використання варто обмежити лише задачами, що мають послідовну природу, наприклад відео.

Щодо ViT, модель, навчена з нуля, показала найгірші результати серед трьох. Причиною цього є слабка здатність виловлювати ознаки без великого обсягу даних або предтренованих ваг. Лише після до навчання предтренованої ViT-моделі було досягнуто суттєвого покращення: точність зросла майже втричі, з’явилась виразна структура у матриці помилок, зменшилась плутанина між емоціями. Це чітко свідчить, що ViT потребує попереднього навчання на великому датасеті для успішного застосування в задачах розпізнавання емоцій.

У випадку з CNN предтренована ResNet50 не перевершила власноруч створену архітектуру CNN, що вказує на важливість архітектурної адаптації під конкретне завдання. Подібна ситуація і з RNN, предтреновані ознаки не покращили результати, навпаки, виявилися менш ефективними.

Таким чином, рекомендації для задач розпізнавання емоцій на статичних зображеннях найкращим вибором залишається CNN, особливо коли важлива як точність, так і швидкодія.

Що до ViT доцільно застосовувати лише в предтренованому вигляді або з дуже великим дата сетом, без цього вона неконкурентоспроможна і немає сенсу, навіть AffecNet який мав 320 тисяч це зображень замало.

RNN не рекомендується для задач на статичних зображеннях, її доцільніше використовувати лише для відео або послідовностей.

При навчанні слід звертати увагу не лише на обсяг, а й якість та збалансованість датасету. Наявність домінантних класів, як-от happy або neutral, впливає на навчання й потребує врахування (наприклад, через зважування втрат або аугментацію рідкісних класів).

У майбутніх роботах доцільно спробувати гібридні архітектури, наприклад поєднання CNN та ViT, щоб об’єднати сильні сторони обох підходів.

Також варто звернути увагу на мультиканальні підходи, наприклад поєднання зображення з аудіо- або текстовими сигналами, для покращення якості класифікації у реальних умовах.

Тож ці висновки та рекомендації мають практичну цінність для розробників емоційно-інтелектуальних систем і можуть слугувати основою для подальших досліджень та розробок.

ВИСНОВКИ

У результаті виконання комплексного дослідження було проведено ґрунтовний аналіз сучасних підходів до розпізнавання емоцій, вивчено особливості застосування різних архітектур нейронних мереж та спроєктовано клієнт-серверну архітектуру для інтеграції моделей у веб-додаток. Основна мета роботи полягала у визначенні найефективніших методів аналізу міміки людини для автоматизованого розпізнавання емоцій і створенні зручного інструменту для тестування та порівняння результатів.

У результаті виконаного дослідження було здійснено ґрунтовний аналіз сучасних підходів до розпізнавання емоцій за мімікою людини, порівняно ефективність трьох ключових архітектур нейронних мереж (CNN, RNN та Vision Transformer), розроблено та реалізовано клієнт-серверну архітектуру для інтеграції моделей у веб-застосунок, а також проведено детальне експериментальне тестування на різноманітних наборах даних із різним обсягом та якістю (FER2013, RAF-DB, неповний і повний AffectNet).

Відзначено, що CNN стабільно забезпечує найвищу точність класифікації й найменшу затримку обробки у задачі розпізнавання емоцій на статичних зображеннях обличь, навіть за наявності шуму, змінного освітлення чи ракурсів; застосування залишкових блоків, SE-модулів та системної аугментації дозволило досягти конкурентних результатів на різних дата сетах, хоча проблеми з класифікацією тонких емоційних проявів зберігаються через неоднорідність даних і домінування деяких класів.

Архітектури на основі RNN виявилися неефективними для статичних зображень: додаткові рекурентні компоненти не покращували якість класифікації, натомість суттєво збільшували затримку та ресурсоємність моделі, що узгоджується з тим, що RNN доцільно використовувати лише для послідовних даних наприклад відео чи інші часові ряди.

Vision Transformer без великого попереднього навчання чи надзвичайно великого набору даних демонструє низьку якість класифікації на малих і середніх дата сетах, проте після до навчання предтренованої ViT-моделі на великому

наборі ImageNet спостерігається суттєве підвищення точності та покращення структури матриць помилок, що свідчить про критичну потребу у предтренуванні для трансформерних архітектур у задачах розпізнавання емоцій.

На основі аналізу предметної галузі було розроблено клієнт-серверну архітектуру, що базується на використанні React для клієнтської частини, Flask для серверної частини та PostgreSQL для зберігання даних. Така архітектура забезпечує чітке розмежування функціональності, дозволяє зручно інтегрувати різні нейронні мережі та забезпечує масштабованість і надійність системи. Веб-додаток, розроблений у рамках дослідження, дозволяє завантажувати зображення чи відео, та отримувати результати розпізнавання по кожній з навчених до цього моделей.

Отже, у задачі автоматизованого розпізнавання емоцій за статичними зображеннями облич найдоцільніше застосовувати CNN-архітектури, які демонструють найкраще поєднання точності та швидкодії. Vision Transformer виправдовує себе тільки за умови наявності потужного предтренування на великому наборі даних або доступу до надвеликих датасетів зображень, без цього навіть великі дата сети наприклад, приблизно 328 тис. зображень AffectNet залишають її неконкурентоспроможною. RNN архітектури в цій задачі не рекомендовані через надмірну ресурсоємність і відсутність покращення якості.

Водночас при підготовці даних слід приділяти увагу не лише обсягу, а й якості, збалансованості й розмаїттю прикладів: домінування класів потребує врахування через методи зважування втрат або цільової аугментації для рідкісних класів. У майбутніх дослідженнях може бути корисним експериментувати з гібридними підходами (поєднання CNN і ViT для використання локальних і глобальних ознак) та мультиканальними системами (комбінація зображення з аудіо- або текстовими сигналами) для покращення стійкості та точності класифікації емоцій у реальних умовах. Ці висновки й рекомендації мають практичне значення для розробки емоційно-інтелектуальних систем і формують основу для подальших наукових і практичних розробок.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Smelyakov, K., Bohomolov, O., Kizitskyi, M., Chupryna, A., Identification of Modern Facial Emotion Recognition Models, CEUR Workshop Proceedings, 2022.

2. Smelyakov, K., Chupryna, A., Bohomolov, O., Hunko, N., The Neural Network Models Effectiveness for Face Detection and Face Recognition, 2021 IEEE Open Conference of Electrical, Electronic and Information Sciences, eStream 2021 – Proceedings, 2021.

3. Facial Emotion Detection Using Deep Learning [Електронний ресурс]. URL: https://ieeexplore.ieee.org/document/9154121 (дата звернення: 19.05.2025).

4. Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion [Електронний ресурс]. URL: https://onlinelibrary.wiley.com/doi/10.1155/2019/9185481 (дата звернення: 19.05.2025).

5. Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions [Електронний ресурс]. URL: https://ieeexplore.ieee.org/document/10284712 (дата звернення: 19.05.2025).

6. Computer Vision: Algorithms and Applications / Dr. Richard Szeliski, 2022. 947 с. (дата звернення: 19.05.2025).

7. Deep Learning with Python / François Chollet, 2017. 384 с. (дата звернення: 12.12.2024).

8. TensorFlow Documentation [Електронний ресурс]. URL: https://www.tensorflow.org/api_docs (дата звернення: 19.05.2025).

9. Flask Documentation [Електронний ресурс]. URL: https://flask.palletsprojects.com/en/stable (дата звернення: 21.05.2025).

10. FER2013 Dataset [Електронний ресурс]. https://www.kaggle.com/datasets/msambare/fer2013 (дата звернення: 17.05.2025).

11. RAF-DB Dataset [Електронний ресурс]. https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset (дата звернення: 17.05.2025).

12. AffectNet Dataset [Електронний ресурс].

https://www.kaggle.com/datasets/minhtmnguyntrn/affectnet-new/ (дата звернення: 17.05.2025).

13. What is a Convolutional Neural Network? An Engineer's Guide [Електронний ресурс]. https://zilliz.com/glossary/convolutional-neural-network (дата звернення: 19.05.2025).

14. Deep Learning(Part4). Recurrent Neural Network (RNN) [Електронний ресурс]. https://medium.com/@sumbatilinda/deep-learning-part4-recurrent-neural-network-rnn-0714e0852581 (дата звернення: 19.05.2025).

15. Step-by-step guide to replicating a Machine Learning paper — Part 1 of 3 [Електронний ресурс]. https://medium.com/@carissa.cullen/step-by-step-guide-to-replicating-a-machine-learning-paper-part-1-of-3-7921425f6460 (дата звернення: 19.05.2025).

16. GitHub – arturysh / emotion-detection. *GitHub*. URL: https://github.com/arturysh/emotion-detection (дата звернення: 28.05.2025).

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

1. Smelyakov, K., Bohomolov, O., Kizitskyi, M., Chupryna, A., Identification of Modern Facial Emotion Recognition Models, CEUR Workshop Proceedings, 2022.

2. Smelyakov, K., Chupryna, A., Bohomolov, O., Hunko, N., The Neural Network Models Effectiveness for Face Detection and Face Recognition, 2021 IEEE Open Conference of Electrical, Electronic and Information Sciences, eStream 2021 – Proceedings, 2021.