Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

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

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

РОЗРОБЛЕННЯ ВЕБСАЙТУ ДЛЯ РОЗПІЗНАВАННЯ ЕМОЦІЙ ЛЮДИНИ ЗА ФОТОГРАФІЄЮ
(тема)

Виконав:
здобувач 4 року навчання,
групи ІТІНФ-22-1
Стриженко П. С.
(прізвище, ініціали)

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

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник проф. Машталір С. В.
(посада, прізвище, ініціали)

Допускається до захисту

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

2026 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту

Кафедра Інформатики

Рівень вищої освіти перший (бакалаврський)

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

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2026 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Стриженко Поліні Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Розроблення вебсайту для розпізнавання емоцій людини за фотографією

затверджена наказом університету від 26 травня 2026 року № 435Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 19 червня 2026 р.
3. Вихідні дані до роботи статті та наукові публікації з тематики афективних обчислень, розпізнавання емоцій за зображенням обличчя, офіційна документація бібліотек PyTorch, OpenCV, MediaPipe, фреймворку FastAPI, а також матеріали щодо веброзробки із використанням HTML, CSS, JavaScript та Bootstrap. Для виконання роботи використано відомості про архітектури ResNet-50 і Vision Transformer, датасети FER2013, KDEF, AffectNet та методи генерації текстових пояснень на основі мімічних ознак.

4. Перелік питань, що потрібно опрацювати в роботі
1. Дослідження методів і моделей розпізнавання емоцій людини за фотографією обличчя.
2. Проєктування та розробка вебсайту з функцією класифікації емоцій і генерації текстових пояснень.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) структурна схема роботи вебзастосунку, UML-діаграма сторінки результатів, приклади інтерфейсу розробленого вебсайту та результати роботи системи для різних емоцій.

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

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

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	13.04.2026	
2	Аналіз завдання, підбір літератури	14.04.26-15.04.26	
3	Аналіз літератури з проблеми, що вирішується	16.04.26-17.04.26	
4	Аналіз існуючих методів розпізнавання	17.04.26-19.04.26	
5	Формування кроків вирішення проблеми	20.04.26-25.04.26	
6	Програмна реалізація	26.04.26-16.05.26	
7	Аналіз отриманих результатів	17.05.26-30.05.26	
8	Оформлення пояснювальної записки	04.06.26-17.06.26	
9	Перевірка на нормоконтроль	10.06.26-19.06.26	
10	Перевірка на плагіат	19.06.26-20.06.26	
11	Рецензування	20.06.26-22.06.26	
12	Підготовка презентації та доповіді	26.06.26-30.06.26	
13	Занесення роботи в електронний архів	03.07.26	
14	Попередній захист кваліфікаційної роботи	09.07.26	

Дата видачі завдання 13 квітня 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ ___проф. Машталір С. В.___
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 62 с., 2 табл., 4 рис., 29 джерел.

АФЕКТИВНІ ОБЧИСЛЕННЯ, РОЗПІЗНАВАННЯ ЕМОЦІЙ, КОМП'ЮТЕРНИЙ ЗІР, ГЕНЕРАЦІЯ ТЕКСТУ, ВЕБЗАСТОСУНОК.

Об'єктом роботи є розроблення вебзастосунку для аналізу емоційного сприйняття візуальної інформації та генерації текстових пояснень.

Метою роботи є розроблення вебзастосунку, здатного визначати домінуючу емоцію за зображенням обличчя за допомогою нейромережевих моделей та генерувати змістовне текстове пояснення на основі виявлених мімічних ознак.

Під час роботи використано: методи штучного інтелекту та аналізу даних, методи веб-розробки.

Практична цінність вебзастосунку полягає в автоматизації аналізу емоційного сприйняття та наданні зрозумілих текстових пояснень, що підвищує освітню цінність системи.

Область застосування: маркетинг, психологія, дизайн, розробка емпатійних інтерфейсів.

AFFECTIVE COMPUTING, EMOTION RECOGNITION, COMPUTER VISION, TEXT GENERATION, WEB APPLICATION.

The object of this work is to develop a web application for analyzing the emotional perception of visual information and generating textual explanations.

The goal of this work is to develop a web application capable of determining the dominant emotion from a facial image using neural network models and generating a meaningful textual explanation based on identified facial features.

The following were used during the work: artificial intelligence and data analysis methods, web development methods.

The practical value of the web application lies in automating the analysis of emotional perception and providing clear textual explanations, which increases the educational value of the system.

Applications: marketing, psychology, design, development of empathetic interfaces.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

API – Application Programming Interface (програмний інтерфейс застосунку, який визначає спосіб взаємодії між різними програмними компонентами)

AU – Action Units (одиниці дій (у системі кодування лицьових рухів FACS), найменші м’язові рухи, з яких складаються вирази обличчя)

AUROC – Area Under the ROC Curve (площа під ROC-кривою, метрика для оцінки якості класифікації)

BERT – Bidirectional Encoder Representations from Transformers (бібліотека для попереднього навчання моделей глибокого навчання для обробки природної мови)

BERTScore – метрика оцінки якості згенерованого тексту, що базується на порівнянні векторних представлень (ембедінгів) слів

BLEU – Bilingual Evaluation Understudy (метрика для автоматичної оцінки якості машинного перекладу та генерації тексту, що обчислює збіги n-грам)

CNN – Convolutional Neural Network (згорткова нейронна мережа, клас глибоких нейронних мереж, що ефективно обробляє дані з сітчастою структурою (наприклад, зображення))

CSS – Cascading Style Sheets (мова стилів, що використовується для опису зовнішнього вигляду вебсторінки)

ECE – Expected Calibration Error (очікувана помилка калібрування, метрика для оцінки того, наскільки добре передбачені ймовірності моделі відповідають реальним частотам)

ERC – Emotion Recognition in Conversations (розпізнавання емоцій у розмовах, підзадача афективних обчислень)

FACS – Facial Action Coding System (система кодування лицьових рухів, розроблена Полом Екманом та Воллесом Фрізеном для опису міміки)

FER2013 – Facial Expression Recognition 2013 (публічний датасет зображень облич для розпізнавання емоцій)

HTML – HyperText Markup Language (мова гіпертекстової розмітки, стандартна мова для створення вебсторінок)

JSON – JavaScript Object Notation (текстовий формат обміну даними, що базується на синтаксисі JavaScript)

KDEF – Karolinska Directed Emotional Faces (датасет зображень облич високої лабораторної якості, створений у Каролінському інституті)

LBP – Local Binary Patterns (локальні бінарні шаблони, оператор для класифікації текстур зображень)

LLM – Large Language Model (велика мовна модель, нейромережа, навчена на величезних обсягах текстових даних)

MLP – Multilayer Perceptron (багатошаровий перцептрон, клас штучних нейронних мереж прямого поширення)

MTCNN – Multi-Task Cascaded Convolutional Networks (каскадна згорткова нейронна мережа для виявлення облич та визначення ключових точок)

NLG – Natural Language Generation (генерація природної мови, підобласть штучного інтелекту)

OpenCV – Open Source Computer Vision Library (бібліотека з відкритим кодом для комп’ютерного зору та обробки зображень)

REST – Representational State Transfer (архітектурний стиль взаємодії компонентів розподіленої системи в мережі)

ResNet – Residual Network (архітектура нейронної мережі з залишковими (skip) з’єднаннями, яка дозволяє будувати дуже глибокі моделі)

ROUGE – Recall-Oriented Understudy for Gisting Evaluation (набір метрик для автоматичної оцінки якості реферування та генерації тексту, орієнтований на повноту)

ВСТУП

Сучасний розвиток технологій штучного інтелекту все більше зосереджується на гуманістичних аспектах, прагнучи не просто аналізувати дані, а також розуміти контекст та емоційне забарвлення інформації. Зображення, як один з основних носіїв інформації в цифрову еру, викликають у людей широкий спектр емоцій, від радості до смутку, від спокою до тривоги. Створення систем, здатних не лише класифікувати об'єкти на фото, а й прогнозувати емоційний вплив цих зображень, відкриває нові горизонти для маркетингу, психології, дизайну та розробки більш емпатійних інтерфейсів.

Сучасні системи аналізу зображень досягли значних успіхів у розпізнаванні об'єктів, сцен та дій. Однак вони залишаються байдужими до емоційного впливу, який візуальний контент справляє на людину. Розуміння цього впливу є критично важливим для створення більш природних, людиноцентричних інтерфейсів взаємодії. Здатність системи не просто класифікувати зображення, але й інтерпретувати та пояснювати емоційну реакцію, яку воно викликає, відкриває нові горизонти для розвитку штучного інтелекту в таких сферах, як персоналізована освіта, психологічне консультування, модерація контенту в соціальних мережах та створення емпатійних віртуальних асистентів. Незважаючи на суб'єктивність емоційних переживань, дослідження показують наявність спільних патернів у сприйнятті візуальних стимулів, що робить можливим моделювання цього процесу.

У роботі вперше буде здійснено комплексне дослідження задачі генерування емоційно-забарвлених пояснень до зображень на основі великого масиву реальних даних. Новизна полягає у розробці підходів, які, на відміну від існуючих, дозволяють не лише констатувати емоцію, але й пояснювати її причини, а також адаптувати стиль пояснення залежно від контексту або бажаної прагматичної спрямованості.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ І ПОСТАНОВКА ЗАДАЧІ

1.1 Еволюція систем аналізу емоцій

Історія наукового інтересу до природи людських емоцій та їх зовнішнього вираження сягає корінням у глибину століть, однак ключові передумови для створення систем їх автоматичного аналізу сформувалися відносно нещодавно. Шлях від перших спостережень за мімікою до сучасних алгоритмів комп'ютерного зору, здатних класифікувати емоційні стани з високою точністю, є результатом тісної взаємодії психології, фізіології та інформатики. Даний підрозділ присвячений дослідженню еволюції цих систем – від фундаментальних праць Чарльза Дарвіна до сучасних архітектур глибокого навчання.

Витоки наукового підходу до вивчення емоційної експресії можна простежити ще з середини XIX століття, коли французький невролог Гійом Дюшен де Булонь проводив свої експерименти з електричною стимуляцією лицьових м'язів, намагаючись зрозуміти механіку мімічних рухів. Однак справжньою відправною точкою для всієї галузі вважається 1872 рік, коли Чарльз Дарвін опублікував працю «Про вираження емоцій у людини і тварин» (The Expression of the Emotions in Man and Animals) [1]. Дарвін висунув революційну для свого часу гіпотезу про універсальність прояву емоцій, стверджуючи, що люди різних рас і культур виражають базові почуття схожим чином. Це твердження заклало фундамент для майбутнього пошуку універсальних патернів, придатних для машинного аналізу, довівши, що емоція не є суто суб'єктивним, а об'єктивно спостережуваним явищем.

Наступний важливий етап розвитку припав на середину XX століття, коли психологи розпочали систематизацію та каталогізацію лицьових рухів. Поява науки кінесики та роботи Резерфорда Бердвістела створили передумови для головного прориву в цій царині – розробки Системи кодування лицьових рухів (Facial Action Coding System, FACS) Полом

Екманом та Воллесом Фрізеном у 1978 році [2]. Екман, розвиваючи ідеї Дарвіна, емпірично підтвердив існування шести базових емоцій (радість, здивування, сум, гнів, відраза, страх), які є загальнолюдськими та не залежать від культурного контексту. Головним же досягненням FACS стало виділення близько 90 одиниць дій (Action Units, AU) – найменших м'язових рухів, комбінація яких утворює той чи інший вираз обличчя. Фактично, Пол Екман створив своєрідну «граматику» емоцій, що дозволяла описувати будь-який мімічний патерн уніфікованим набором кодів. Ця система, залишаючись «біблією» для розробників алгоритмів розпізнавання емоцій, на десятиліття визначила напрямок досліджень, надавши інженерам чітку структуру для аналізу.

Перехід від теоретичних психологічних моделей до перших спроб їх автоматизації став можливим лише в 1990-х роках із появою достатніх обчислювальних потужностей та цифрових технологій [4]. Саме в цей період виникає термін «емоційні обчислення» (Affective Computing), введений професоркою Массачусетського технологічного інституту Розаліндою Пікард у 1995 році в однойменній книзі [3]. Пікард визначила нову галузь як «вивчення та розробку систем і пристроїв, які можуть розпізнавати, інтерпретувати, обробляти та моделювати людські афекти». Її робота ознаменувала народження міждисциплінарної науки, що поєднала психологію, когнітивістику та комп'ютерні технології з метою наділення машин емоційним інтелектом.

Перші технічні реалізації систем розпізнавання емоцій базувалися на класичних методах комп'ютерного зору та обробки сигналів. Вони намагалися безпосередньо імплементувати знання, закладені в FACS. Алгоритми оптичного потоку відстежували рух ключових точок обличчя, а для детекції таких ознак, як зморшки чи положення брів, застосовувалися фільтри Габора, локальні бінарні шаблони (LBP) та інші техніки. Типовий пайплайн класичного підходу включав детекцію обличчя, пошук антропометричних точок, обчислення геометричних характеристик (наприклад, відстаней між

куточками губ) та текстурних особливостей. На основі цих вручну розроблених ознак будувалися класифікатори, часто з використанням методу зваженої суми або простих моделей машинного навчання. Недоліком такого підходу була його обмежена здатність до узагальнення та чутливість до умов зйомки, хоча при високоякісних даних він міг демонструвати непогану точність.

Справжня революція в галузі відбулася в середині 2010-х років із тріумфальним приходом методів глибокого навчання (deep learning), зокрема згорткових нейронних мереж (CNN) [5]. На відміну від класичних методів, які вимагали ручного проєктування ознак, глибокі нейромережі здатні самостійно навчатися ієрархічним рівням абстракції безпосередньо з «сирих» пікселів зображення. Це дозволило досягти безпрецедентної точності та робастності, особливо під час роботи з даними «з дикої природи» (in-the-wild), тобто в умовах різноманітного освітлення, ракурсів та якості зображення. Однак, як зазначають дослідники, нейромережеві моделі часто працюють за принципом «чорної скриньки», що ускладнює інтерпретацію того, на основі яких саме ознак (наприклад, конкретних одиниць дій за Екманом) було прийнято те чи інше рішення.

Паралельно з удосконаленням архітектур нейромереж, критично важливу роль відіграла поява великих розмічених датасетів. Якщо перші бази даних, такі як Cohn-Kanade, містили лише сотні зображень, отриманих в лабораторних умовах, то для навчання глибоких моделей знадобилися на порядки більші обсяги даних. Це стимулювало створення ресурсів на зразок AffectNet, що містить сотні тисяч зображень, зібраних з інтернету. Окрім статичних зображень, розвиток отримали й відеодатасети, які дозволяють аналізувати динаміку емоцій у часі. Дослідження також демонструють поступовий перехід від унімодальних систем (аналіз лише обличчя) до мультимодальних підходів, які інтегрують інформацію з виразу обличчя, голосу, жестів тіла та фізіологічних сигналів для отримання більш повної та точної картини емоційного стану людини.

Таким чином, еволюція систем аналізу емоцій пройшла довгий шлях від філософських спостережень Дарвіна через структурування знань Екманом до технологічної революції, здійсненої Розаліндою Пікард та дослідниками глибокого навчання. Сьогодні ці системи, колись доступні лише в наукових лабораторіях, стають невід'ємною частиною нашого повсякденного життя, інтегруючись у маркетингові дослідження, системи безпеки, автомобілебудування, освіту та охорону здоров'я. Розуміння цього історичного контексту є необхідним для адекватної постановки задачі розробки сучасного вебсайту з обробки та емоційного сприйняття візуальної інформації.

1.2 Сучасні методи та підходи до аналізу емоцій

Незважаючи на значний прогрес, досягнутий у галузі розпізнавання емоцій за останнє десятиліття, створення систем, здатних надійно та точно інтерпретувати емоційні стани в реальних умовах, залишається складною науково-технічною проблемою. Сучасний етап розвитку емоційних обчислень (Affective Computing) характеризується переосмисленням фундаментальних підходів – відходом від спрощених унімодальних моделей, орієнтованих на лабораторні умови, до складних мультимодальних систем, здатних працювати «в дикій природі» (in-the-wild) та враховувати індивідуальні, культурні та контекстуальні особливості людини. У даному підрозділі розглядаються ключові виклики, що стоять перед розробниками таких систем, та сучасні задачі, які визначають порядок денний досліджень у цій царині. Одним із головних викликів, який тривалий час залишався в тіні досягнень глибокого навчання, є проблема узагальнення та робастності моделей [6]. Більшість алгоритмів демонструють вражаючу точність на контрольованих наборах даних, проте їхня ефективність різко падає під час роботи з даними з реального світу. Це явище, відоме як перетренування (overfitting), виникає через

обмежену різноманітність тренувальних вибірок. Реальні умови висувають низку вимог, які важко змоделювати в лабораторії. До них належать: неконтрольоване освітлення (від яскравого сонячного світла до глибокої тіні), варіативність ракурсів та пози голови, часткова оклюзія обличчя (наприклад, сонцезахисними окулярами, медичною маскою або волоссям), а також різноманітна роздільна здатність та якість зображень, отриманих з різних пристроїв. Забезпечення стабільної роботи системи в таких різноманітних контекстах є першочерговою задачею для її практичного впровадження.

Тісно пов'язаною з попередньою є проблема зміщення (bias) та недостатньої репрезентативності даних [6-7]. Історично склалося так, що основні датасети для розпізнавання емоцій (FER2013, CK+, JAFFE) формувалися переважно на основі зображень представників західних або східноазійських популяцій, що створило так званий «культурний перекіс» (cultural bias). Моделі, навчені на таких даних, некоректно інтерпретують вирази облич людей інших етнічних груп, наприклад, африканського чи латиноамериканського походження, що призводить до систематичних помилок та расової нерівності в роботі алгоритмів. Крім етнічної приналежності, bias може стосуватися віку, статі чи навіть культурних особливостей вираження емоцій. Наприклад, дослідження метафор у комунікації показують, що сприйняття емоційного підтексту суттєво відрізняється в різних культурах [8]. Тому сучасною задачею є створення більш різноманітних та інклюзивних датасетів (як MUTFER2024 для африканської популяції), а також розробка методів навчання, стійких до подібних зміщень [7]. Еволюція наукової думки привела дослідників до розуміння обмеженості унімодальних підходів [6]. Людина в природній комунікації використовує всі доступні канали сприйняття – візуальний (вираз обличчя, жести, пози), аудіальний (інтонація, тембр, гучність голосу) та лінгвістичний (зміст сказаного). Наслідуючи цей біологічний принцип, сучасні системи прагнуть до мультимодальної інтеграції. Виклик полягає в тому, як ефективно об'єднати (інтегрувати) різнорідні потоки даних. Аудіо- та

візуальна інформація мають різну природу, структуру та часову динаміку. Їх синхронізація та об'єднання в єдиний інформативний вектор ознак, який би враховував як унікальні характеристики кожної модальності, так і їхні спільні кореляції, є складною дослідницькою проблемою. Сучасні підходи, такі як механізми уваги (attention) та розділення ознак (feature disentanglement), намагаються вирішити це завдання, але універсального рішення досі не знайдено.

Наступний виклик пов'язаний з динамікою емоцій та контекстом, особливо в розмовах (Emotion Recognition in Conversations, ERC) [6], [14]. Емоція – це не статичний стан, а процес, що розгортається в часі. Людина може посміхатися від радості, але та сама посмішка може виражати сарказм чи гіркоту в залежності від попередньої репліки співрозмовника. Ігнорування цього динамічного контексту призводить до помилкової класифікації. Моделі повинні не просто аналізувати ізольоване зображення чи відрізок мовлення, а враховувати історію взаємодії, що вимагає застосування рекурентних нейронних мереж, трансформерів та інших архітектур, здатних моделювати часові залежності. Крім того, поява потужних чат-ботів ставить нове завдання – розпізнавання емоцій у взаємодії «людина-чат-бот», де емоційна експресія користувачів є значно стриманішою та менш вираженою порівняно з людським діалогом.

Окремим блоком викликів є інтерпретованість та пояснюваність рішень (Explainable AI, XAI) [14]. Сучасні глибокі нейронні мережі, особливо трансформери, часто працюють як «чорні скриньки», ухвалюючи точні, але непрозорі для людини рішення. У чутливих сферах застосування, таких як охорона здоров'я (діагностика депресії), освіта чи право, критично важливо розуміти, на підставі яких саме ознак (наприклад, рух певного м'яза обличчя або зміна тону голосу) модель зробила той чи інший висновок. Відсутність пояснюваності знижує довіру до системи та унеможливлює її використання в відповідальних застосунках. Тому розробка методів, які надають зрозумілі людині пояснення (наприклад, у вигляді теплових карт або текстових

коментарів), є одним із пріоритетних напрямків.

Нарешті, існує низка практичних та етичних викликів. Для роботи в реальному часі на мобільних пристроях або в робототехніці потрібні легкі (lightweight) моделі, які споживають мало енергії та мають високу швидкодію без втрати точності. Крім того, збір даних, особливо мультимодальних, пов'язаний із забезпеченням конфіденційності та інформованої згоди користувачів, що накладає додаткові обмеження на процес розробки та впровадження. Отже, сучасний розробник систем аналізу емоцій має вирішувати комплекс взаємопов'язаних задач, балансуючи між точністю, робастністю, інтерпретованістю та етичністю своїх рішень.

1.3 Огляд існуючих вебрішень для аналізу емоцій

Стрімкий розвиток технологій комп'ютерного зору та глибокого навчання зумовив появу значної кількості веборієнтованих рішень для аналізу емоційного стану людини за зображенням обличчя. Ці інструменти варіюються від простих демонстраційних вебзатосунків до потужних комерційних API, інтегрованих у маркетингові платформи, системи охорони здоров'я та аналітики користувацького досвіду. Проведення огляду існуючих вебрішень є критично важливим етапом дослідження, оскільки дозволяє виявити сильні та слабкі сторони сучасних підходів, визначити функціональні прогалини та сформулювати вимоги до власної розробки. У даному підрозділі розглядаються основні типи вебрішень, їх ключові характеристики та обмеження.

Сучасний ринок пропонує дві основні категорії вебрішень для аналізу емоцій: хмарні платформи з програмними інтерфейсами (API) та автономні вебзастосунки. До першої категорії належать такі гіганти індустрії, як Microsoft Azure Face API, Amazon Rekognition, Google Cloud Vision API та IBM Watson Visual Recognition. Ці платформи надають розробникам доступ до

потужних нейромережевих моделей через прості HTTP-запити. Їхньою головною перевагою є висока точність, масштабованість та постійна підтримка з боку компаній-розробників. Вони здатні не лише класифікувати базові емоції (щастя, сум, гнів, подив, страх, відраза), але й визначати вік, стать, наявність аксесуарів та інші атрибути. Однак використання таких API пов'язане з низкою недоліків: вартість послуг, яка зростає пропорційно кількості запитів, необхідність передачі конфіденційних даних (зображень користувачів) на сторонні сервери, що створює потенційні ризики для приватності, обмежена можливість тонкого налаштування моделей під специфічні дослідницькі задачі.

Другу категорію складають спеціалізовані вебзастосунки та бібліотеки з відкритим кодом, які можуть бути розгорнуті на власному сервері. Прикладом може слугувати бібліотека DeepFace, яка надає зручну Python-обгортку для декількох сучасних моделей розпізнавання облич та емоцій і дозволяє створювати на її основі вебсервіси. Окремо слід згадати про чисельні науково-дослідні проєкти, які публікуються у відкритому доступі у вигляді демо-версій. Такі рішення часто є безкоштовними, але поступаються комерційним аналогам у швидкодії та стабільності роботи. Аналіз доступних на даний момент вебзастосунків, зокрема тих, що можна знайти в каталогах на зразок Google Play (наприклад, застосунок «Face emotion detection»), показує, що більшість із них орієнтовані на розважальний або споживацький сегмент, пропонуючи базову функціональність без глибокої аналітики чи можливості отримання текстових пояснень результатів [13]. Вони часто обмежуються підрахунком кількості облич або визначенням емоційного стану як «усміхнений» або «сумний», не заглиблюючись у складніші емоційні категорії.

Незважаючи на різноманітність інструментів, більшість існуючих вебрішень мають спільні функціональні обмеження. По-перше, вони зосереджені виключно на класифікації зображення, тобто на присвоєнні йому однієї з декількох заздалегідь визначених міток. По-друге, майже всі

комерційні системи працюють за принципом «чорної скриньки» – користувач отримує результат (наприклад, «гнів з імовірністю 0,95»), але не має жодного уявлення про те, на підставі яких саме ознак обличчя (положення брів, форма губ, зморшки) модель дійшла такого висновку. Відсутність пояснювальних механізмів (explainable AI) є суттєвим недоліком для наукових та освітніх цілей, де важливе розуміння причинно-наслідкових зв'язків [14]. Крім того, інтерфейси багатьох вебрішень перевантажені технічними деталями і не завжди є інтуїтивно зрозумілими для пересічного користувача.

Таким чином, проведений огляд свідчить про наявність ніші для створення вебсайту, який би поєднував точність сучасних моделей глибокого навчання з прозорістю їх роботи та зручним користувацьким інтерфейсом. Головною відмінністю майбутньої розробки має стати не просто видача результату класифікації, а генерація змістовних текстових пояснень, що описують ключові мімічні ознаки, які зумовили віднесення зображення до тієї чи іншої емоційної категорії. Такий підхід дозволить використовувати систему не лише як інструмент аналізу, але і як освітню платформу для вивчення психології емоцій та роботи алгоритмів комп'ютерного зору.

1.4 Постановка задачі

Проведений у попередніх підрозділах аналіз еволюції систем аналізу емоцій, сучасних викликів у цій галузі та огляд існуючих вебрішень дозволяє сформулювати комплексне обґрунтування необхідності створення нового вебресурсу та визначити конкретні задачі, які мають бути вирішені в межах даної кваліфікаційної роботи.

Об'єктом роботи є процес автоматичного аналізу та інтерпретації емоційного впливу візуального контенту на людину.

Метою розробки є створення веборієнтованої системи для аналізу емоційного сприйняття візуальної інформації, яка на основі зображення

обличчя людини не лише визначає домінуючу емоцію, але й надає текстове пояснення цього рішення, описуючи ключові мімічні ознаки. Така система має стати корисним інструментом як для дослідників у галузі психології та штучного інтелекту, так і для широкого кола користувачів, зацікавлених у розвитку власного емоційного інтелекту, адже здатність розпізнавати емоції інших є фундаментальною складовою ефективної комунікації.

2 МЕТОДИ ТА ТЕХНОЛОГІЇ ДЛЯ АНАЛІЗУ ЕМОЦІЙНОГО СПРИЙНЯТТЯ

2.1 Архітектури нейронних мереж для комп’ютерного зору

Вибір архітектури нейронної мережі є фундаментальним рішенням при розробці будь-якої системи комп’ютерного зору, особливо для такого завдання, як розпізнавання емоцій, де потрібен тонкий аналіз мікровиразів обличчя. Сучасний ландшафт архітектур глибокого навчання пропонує два принципово різних підходи до обробки візуальної інформації: класичні згорткові нейронні мережі (CNN) та новітні трансформерні архітектури. У даному підрозділі детально розглядаються дві ключові архітектури – ResNet як представник родини CNN та Vision Transformer (ViT) як втілення трансформерної архітектури, їхні теоретичні засади, переваги, обмеження та особливості застосування для задач класифікації зображень.

Згорткові нейронні мережі протягом майже десятиліття домінували у сфері комп’ютерного зору, імітуючи роботу зорової кори людини. Їхній фундаментальний принцип полягає у використанні фільтрів (ядер), які ковзають по вхідному зображенню, виявляючи локальні ознаки – від простих (ребра, кути) на ранніх шарах до складних (очі, ніс, рот) на глибших. Однак зі збільшенням глибини мережі дослідники зіткнулися з проблемою деградації, коли додавання нових шарів призводило не до покращення, а до зростання помилки навчання, що не було пов’язано з перенавчанням.

Вирішення цієї фундаментальної проблеми запропонувала архітектура Residual Network (ResNet), представлена дослідниками з Microsoft у 2015 році [9]. Ключовою інновацією ResNet стало введення концепції залишкового навчання (residual learning) та блокі із «короткими» з’єднаннями (skip connections). Замість того, щоб намагатися вивчити безпосередньо цільове відображення $H(x)$, мережа навчається апроксимувати залишкову функцію $F(x) = H(x) - x$. Тоді вихід блоку обчислюється як $y = F(x) + x$, де x – вхід

блоку, який додається через shortcut-з'єднання, що оминає один або декілька шарів.

Такий підхід має глибоке теоретичне обґрунтування. По-перше, якщо оптимальне відображення є близьким до тотожного (identity), мережі легше навчитися встановлювати ваги близькими до нуля $(F(x) \to 0)$, ніж точно відтворювати вхідний сигнал через стек шарів. По-друге, і це найважливіше, shortcut-з'єднання кординально покращують процес зворотного поширення помилки. Під час обчислення градієнту функції втрат L відносно входу x для залишкового блоку отримуємо вираз:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \cdot \frac{\partial F(x)}{\partial x} + \frac{\partial L}{\partial y} \cdot 1. \tag{1.1}$$

Наявність доданку $\frac{\partial L}{\partial y} \cdot 1$ створює прямий «магістральний» шлях для градієнта до ранніх шарів, що ефективно усуває проблему зникаючих градієнтів навіть у мережах з сотнями шарів.

Архітектурно ResNet будується з типових блоків. У класичному варіанті для неглибоких мереж використовується блок з двома шарами згортки 3 × 3. Для глибших версій (ResNet-50, ResNet-101, ResNet-152) застосовується «пляшкове горлечко» (bottleneck block). Воно складається з трьох шарів: згортка 1 × 1 для зменшення розмірності, згортка 3 × 3 для вилучення ознак і ще одна згортка 1 × 1 для відновлення розмірності. Така конструкція значно зменшує кількість параметрів та обчислювальну складність, дозволяючи будувати надзвичайно глибокі моделі. Наприклад, ResNet-152 має 204 мільйони параметрів, але завдяки bottleneck-дизайну залишається практичною для навчання. Коли розмірності x та $F(x)$ не збігаються (наприклад, при зміні кількості каналів), використовуються проекційні shortcut-з'єднання зі згорткою 1 × 1 для вирівнювання.

Кардинально новий підхід до аналізу зображень запропонувала архітектура Vision Transformer (ViT), представлена у 2020 році [10]. Вона

перенесла успішну концепцію трансформерів із обробки природної мови в комп'ютерний зір, довівши, що механізми уваги (attention) можуть ефективно працювати з візуальними даними без використання згорток [11].

Основна ідея ViT: зображення обробляється як послідовність, подібно до речення у тексті. Спочатку вхідне зображення розбивається на фіксовані квадратні фрагменти – патчі (patches), наприклад, розміром 16 X 16 пікселів. Кожен патч розгортається у вектор і лінійно проектується у прихований патч (patch embedding). Оскільки трансформер, на відміну від згорткової мережі, не має вродженого розуміння просторового розташування, до цих патчів додаються позиційні патчі (position embeddings), які кодують інформацію про початкове місце патча на зображенні. Далі, за аналогією з BERT, до послідовності додається спеціальний CLS токен, чий фінальний стан використовуватиметься для класифікації всього зображення.

Отримана послідовність векторів подається на вхід трансформерного енкодера, який складається з кількох ідентичних блоків (наприклад, 12 для ViT-Base). Ключовим елементом кожного блоку є багатоголовий механізм самоуваги (multi-head self-attention). Він дозволяє кожному патчу «спілкуватися» з усіма іншими патчами одночасно, обчислюючи ваги уваги, які показують, наскільки вони релевантні один одному. Це дає моделі здатність бачити глобальний контекст з першого ж шару, на відміну від CNN, які будують його поступово. Математично це виражається через обчислення матриць запитів (Q), ключів (K) та значень (V), де ваги уваги визначаються як:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (1.2)$$

Після шару самоуваги йдуть шари нормалізації та повнозв'язні шари (MLP), а shortcut-з'єднання використовуються навколо обох компонентів, що нагадує принципи ResNet.

Така архітектура наділяє ViT унікальними властивостями. Здатність

моделювати глобальні залежності робить його особливо ефективним для розуміння складних сцен, де важливі далекі взаємозв'язки між об'єктами. Крім того, трансформери демонструють чудову масштабованість: збільшення розміру моделі та обсягу даних призводить до передбачуваного зростання точності. Однак ця потужність має свою ціну. ViT є надзвичайно вимогливим до даних – для ефективного навчання «з нуля» йому потрібні мільйони зображень, інакше він поступається CNN. Дослідження показують, що на невеликих наборах даних, як CIFAR-10, проста CNN або ResNet можуть показувати вищу точність (80,66% та 72,46% відповідно), ніж ViT (68,73%), через відсутність вбудованих індуктивних зсувів (bias), таких як локальність та трансляційна еквіваріантність [10]. Проте після попереднього навчання на гігантських датасетах (наприклад, JFT-300M) та тонкого налаштування, ViT здатний перевершити найкращі CNN, досягаючи точності понад 84% на ImageNet.

Вибір між ResNet та ViT для системи аналізу емоцій є нетривіальним і залежить від конкретних умов. ResNet, з його вбудованою індуктивною здатністю до аналізу локальних структур, є природним кандидатом для розпізнавання мікровиразів обличчя, які локалізовані навколо очей, брів чи губ. Його архітектура добре вивчена, існує безліч попередньо навчених моделей на великих датасетах облич, що робить його надійним та ефективним вибором, особливо при обмеженому обсязі даних. Крім того, ResNet є менш вимогливим до обчислювальних ресурсів, що важливо для розгортання у вебзастосунку.

З іншого боку, ViT пропонує унікальну перевагу – здатність враховувати глобальний контекст всього обличчя. Емоція – це не просто сума рухів окремих м'язів, а цілісний патерн, де положення голови, загальна симетрія чи асиметрія обличчя можуть відігравати вирішальну роль. Механізм самоуваги теоретично здатен вловлювати ці складніші взаємозв'язки. Як показують дослідження, гібридні архітектури, які поєднують згорткові шари для вилучення локальних ознак із трансформерними блоками для моделювання

глобального контексту, можуть забезпечити найкращу продуктивність [12]. Наприклад, у задачі розпізнавання хвороб рослин ViT показав вищу здатність до узагальнення порівняно з ResNet. Це свідчить про потенціал трансформерів для задач, де важливе цілісне сприйняття об’єкта, що є релевантним і для аналізу емоцій.

Таким чином, у межах даної роботи доцільно провести експериментальне порівняння обох підходів. ResNet буде розглядатися як потужна базова лінія, перевірена часом, тоді як ViT – як перспективна архітектура, здатна запропонувати нові можливості завдяки глобальному аналізу зображення.

2.2 Огляд та підготовка датасетів

Якість та характеристики даних, на яких навчаються моделі глибокого навчання, є визначальним фактором успішності будь-якої системи комп’ютерного зору. Для задачі розпізнавання емоцій це твердження набуває особливої ваги, оскільки модель має навчитися виділяти універсальні патерни міміки, ігноруючи при цьому індивідуальні особливості, умови зйомки та шуми. Вибір репрезентативних та якісних датасетів є фундаментом для побудови робастної системи, здатної працювати в реальних умовах. У даному підрозділі розглядаються три загальновизнані набори даних для розпізнавання емоцій – AffectNet, FER2013 та KDEF – їхні характеристики, переваги та недоліки, а також методи підготовки даних до використання в експериментальній частині роботи.

FER2013 (Facial Expression Recognition 2013) є одним із найбільш цитованих та широко використовуваних датасетів у галузі, особливо популярним завдяки його використанню у змаганнях на платформі Kaggle [5]. Він містить 35 887 зображень облич у градаціях сірого, кожне розміром 48×48 пікселів. Зображення зібрані в різних умовах, що робить датасет

репрезентативним для реальних сценаріїв застосування. Кожне зображення анотоване однією з семи емоційних категорій: гнів, відраза, страх, радість, сум, здивування та нейтральний стан. Датасет розділений на три частини: тренувальний набір (28 709 зображень), загальний тестовий набір (3 589 зображень) та спеціальний тестовий набір (3 589 зображень).

Головною перевагою FER2013 є його розмір та різноманітність, що дозволяє навчати моделі глибокого навчання без необхідності додаткового збору даних. Однак датасет має й суттєві обмеження. По-перше, низька роздільна здатність зображень (48×48) обмежує кількість деталей, які може використати модель. По-друге, якість анотацій є неідеальною, оскільки деякі зображення містять помилки розмітки або нечітко виражені емоції. По-третє, у датасеті спостерігається значний дисбаланс класів: наприклад, клас «радість» представлений значно більшою кількістю зразків, ніж клас

«відраза», що може призводити до зміщення моделі. Незважаючи на ці недоліки, FER2013 залишається цінним ресурсом для початкового навчання та порівняння архітектур, а сучасні дослідження демонструють на ньому точність до 77,6%.

KDEF (Karolinska Directed Emotional Faces) є датасетом лабораторної якості, створеним у 1998 році в Каролінському інституті (Швеція) спеціально для наукових досліджень. Він містить 4900 кольорових зображень високої роздільної здатності, на яких 70 акторів (35 жінок і 35 чоловіків) демонструють сім базових емоцій (ті ж самі, що й у FER2013) з п’яти різних ракурсів. Завдяки контрольованим умовам зйомки (однакове освітлення, однаковий фон, стандартизована поза) KDEF забезпечує високу якість зображень та чіткість емоційних виразів.

Перевагою KDEF є його валідованість – датасет використано у понад 1500 наукових публікаціях, що підтверджує його надійність. Завдяки високій якості, моделі, навчені на KDEF, досягають високих показників точності – у деяких дослідженнях точність класифікації сягає 96,51% та навіть 98,78%. Однак лабораторне походження даних є водночас і обмеженням: модель,

навчена виключно на KDEF, може погано узагальнювати на реальні зображення з різноманітними ракурсами, освітленням та оклюзіями. Крім того, датасет містить переважно представників європеоїдної раси, що створює ризик культурного зміщення. Важливо також зазначити, що KDEF доступний лише для дослідницьких некомерційних цілей і не може вільно поширюватися.

AffectNet є наймасштабнішим із трьох датасетів, створеним дослідницькою групою під керівництвом професора Мохаммада Махура з Університету Денвера. Він містить понад один мільйон зображень облич, зібраних з інтернету за допомогою 1250 емоційно-забарвлених ключових слів шістьма мовами. Приблизно 440 000 з цих зображень були вручну анотовані для семи дискретних категорій емоцій (аналогічних до попередніх датасетів) та за шкалами валентності та збудження. Нещодавно з’явилася розширена версія – AffectNet+, яка пропонує «м’які мітки» (soft labels), що відображають невизначеність анотаторів, а також додаткові метадані, включаючи демографічні атрибути та векторні представлення облич.

AffectNet є унікальним ресурсом завдяки своєму масштабу та різноманітності: він охоплює широкий спектр вікових груп, рас, умов освітлення та ракурсів, що зустрічаються в реальному житті. Це робить його ідеальним для навчання моделей, стійких до різноманітних умов. Крім того, наявність анотацій за валентністю та збудженням відкриває можливості для моделювання емоцій у неперервному просторі, а не лише в дискретних категоріях. Водночас, великий обсяг даних потребує значних обчислювальних ресурсів для обробки та навчання. Крім того, через автоматичний збір даних з інтернету, якість окремих зображень та анотацій може варіюватися. Дослідження показують, що точність класифікації на AffectNet може бути нижчою, ніж на лабораторних датасетах – наприклад, 79,5% або 59% в залежності від складності завдання.

Підготовка датасетів до використання є критично важливим етапом, що безпосередньо впливає на якість навчання моделей. Оскільки три обрані датасети мають різні формати та характеристики, процедура підготовки

включає кілька уніфікованих кроків.

Першим кроком є гармонізація класів. Усі три датасети підтримують сім базових емоцій за Екманом, що дозволяє об’єднати їх для розширення тренувальної вибірки або використовувати для послідовного навчання.

Другим кроком є детекція та вирівнювання облич. Оскільки зображення в датасетах можуть містити не лише обличчя, а й фон, необхідно застосувати детектори (наприклад, MTCNN або Dlib) для виділення області обличчя. Після детекції виконується вирівнювання обличчя за ключовими точками (очі, ніс, куточки губ) для приведення всіх зображень до єдиного просторового формату.

Третім кроком є нормалізація та приведення до єдиного розміру. Для сумісності з архітектурами нейронних мереж (наприклад, ResNet очікує на вхід зображення розміром 224×224 пікселі) необхідно змінити розмір усіх зображень. При цьому зображення з FER2013 (48×48) потребують збільшення, що може призвести до втрати якості, тому важливо використовувати інтерполяційні методи, які мінімізують артефакти. Також виконується нормалізація піксельних значень до діапазону [0,1] або [-1,1] залежно від вимог моделі.

Четвертим кроком є застосування аугментації даних. Для підвищення здатності моделі до узагальнення та компенсації дисбалансу класів застосовуються різноманітні перетворення до тренувальних зображень: випадкові повороти, горизонтальне віддзеркалення, зміна яскравості та контрасту, додавання шуму, а також випадкове блокування частин зображення для імітації оклюзій. Для класів з малою кількістю зразків (наприклад, «відраза» у FER2013) може застосовуватися збільшення ваги помилки або синтетична генерація даних.

П’ятим кроком є розподіл даних на навчальну, валідаційну та тестову вибірки. Для забезпечення коректної оцінки моделей важливо, щоб тестові дані не використовувалися під час навчання. У випадку FER2013 цей розподіл уже виконано авторами. Для KDEF та AffectNet необхідно самостійно

виконати розбиття, забезпечуючи при цьому, щоб зображення однієї особи не потрапили одночасно до тренувальної та тестової вибірок, що дозволяє оцінити здатність моделі до узагальнення на нових людях.

Таким чином, комбінування різнорідних датасетів – лабораторного KDEF, великомасштабного AffectNet та «зібраного з дикої природи» FER2013 – дозволяє створити збалансовану навчальну базу, яка поєднує високу якість зображень з різноманітністю реальних умов. Ретельна підготовка даних, включаючи вирівнювання, нормалізацію та аугментацію, є запорукою успішного навчання моделей у експериментальній частині роботи.

2.3 Методи оцінки якості класифікації та генерації тексту

Розробка будь-якої системи штучного інтелекту неможлива без чітко визначених процедур оцінювання її ефективності. Для вебсайту, що поєднує задачу класифікації емоцій з генерацією текстових пояснень, необхідно застосовувати два різних, але взаємопов’язаних класи метрик. Перший клас призначений для оцінки точності розпізнавання емоційного стану на зображенні, тоді як другий – для визначення якості згенерованого тексту, його відповідності змісту зображення та лінгвістичної правильності. У даному підрозділі розглядаються ключові методи оцінки для обох типів задач, їхні переваги, обмеження та особливості застосування в контексті даної роботи.

2.3.1 Метрики оцінки якості класифікації емоцій

Задача класифікації емоцій за зображенням обличчя є класичною задачею багатокласової класифікації з контрольованим навчанням. Для оцінки ефективності моделей у цій задачі використовується стандартний набір метрик, що базуються на аналізі матриці невідповідностей (confusion matrix).

Базовою метрикою є точність, яка обчислюється як відношення кількості правильно класифікованих зображень до загальної кількості зображень у вибірці. Ця метрика є інтуїтивно зрозумілою та широко застосовуваною, зокрема в дослідженнях на датасеті FER2013, де сучасні моделі досягають точності близько 77,6%. Однак точність має суттєвий недолік: вона може бути оманливою у випадках дисбалансу класів. Оскільки в реальних датасетах, як FER2013, деякі емоції (наприклад, «радість») представлені значно більшою кількістю зразків, ніж інші (наприклад, «відраза»), модель може досягати високої загальної точності, просто ігноруючи малопредставлені класи. Тому покладатися виключно на точність для оцінки якості моделі є недостатнім.

Для отримання більш об'єктивної картини використовуються метрики, що обчислюються для кожного класу окремо. Точність для певного класу показує, яка частина зображень, класифікованих моделлю як ця емоція, дійсно їй відповідає. Іншими словами, це міра достовірності позитивних передбачень моделі. Повнота, або чутливість, вимірює, яку частку всіх зображень певного класу модель змогла правильно виявити. Високе значення повноти означає, що модель рідко пропускає зображення з цією емоцією.

Для узагальнення точності та повноти використовується F1-міра, яка є їхнім гармонійним середнім. F1-міра досягає свого максимуму лише тоді, коли і точність, і повнота є високими, що робить її чудовим інструментом для порівняння моделей, особливо на незбалансованих даних. У багатокласовій класифікації застосовуються два підходи до усереднення цих метрик: макро-усереднення обчислює середнє арифметичне метрик по всіх класах, надаючи однакову вагу кожному класу незалежно від його розміру, тоді як мікро-усереднення агрегує внесок усіх класів, фактично відображаючи загальну ефективність моделі. У випадках значного дисбалансу класів макро-усереднені метрики є більш інформативними, оскільки вони чутливі до проблем у малопредставлених категоріях.

Окремим напрямком є оцінка якості ймовірностей, які видає модель.

Площа під ROC-кривою (AUROC, Area Under the ROC Curve) вимірює здатність моделі розрізняти класи при різних порогах прийняття рішень. Чим вище значення AUROC (від 0 до 1), тим краще модель відокремлює один клас від інших. Для оцінки того, наскільки добре передбачені ймовірності відповідають реальній частоті появи подій, використовується очікувана помилка калібрування (ECE, Expected Calibration Error). Низьке значення ECE свідчить про те, що коли модель передбачає емоцію з імовірністю, скажімо, 90%, то в дійсності ця емоція з'являється приблизно в 90% випадків.

Сучасні дослідження також вказують на необхідність розробки спеціалізованих метрик, які враховують психологічну близькість різних емоцій. Традиційна точність однаково карає за помилку, коли, наприклад, «збудження» класифікується як «гнів» або як «благоговіння», хоча з психологічної точки зору перша помилка є значно грубішою. Тому перспективним напрямком є впровадження метрик, які враховують відстань між емоціями на «емоційному колі».

2.3.2 Методи оцінки якості генерації тексту

Оцінка якості згенерованого тексту є складнішим завданням, ніж оцінка класифікації, оскільки правильна відповідь не є єдиною. Для тексту, що пояснює емоцію, існує багато варіантів коректних формулювань. Тому застосовуються як автоматичні метрики, що порівнюють згенерований текст з еталонним, так і методи залучення людини до оцінювання.

Найпоширенішими автоматичними метриками для оцінки якості тексту є BLEU (Bilingual Evaluation Understudy) та ROUGE (Recall-Oriented Understudy for Gisting Evaluation), запозичені з задач машинного перекладу та автоматичного реферування.

Метрика BLEU оцінює якість згенерованого тексту, обчислюючи середню точність збігів n-грам (послідовностей слів) між згенерованим

текстом та одним або кількома еталонними текстами. Вона також включає штраф за довжину (brevity penalty), щоб запобігти генерації надто коротких відповідей, які штучно підвищують точність. BLEU повертає значення від 0 до 1, де 1 означає повний збіг з еталоном. Однак BLEU має обмеження: він погано корелює з людською оцінкою в задачах, де важлива не лише лексична, а й смислова подібність, і не враховує семантичні варіації.

Метрика ROUGE , на відміну від BLEU, орієнтована на повноту. Вона вимірює, яка частка n-грам (або найдовшої спільної підпослідовності) з еталонного тексту міститься в згенерованому тексті. Найпоширенішими варіантами є ROUGE-N (для n-грам) та ROUGE-L (на основі найдовшої спільної підпослідовності). ROUGE краще підходить для оцінки текстів, де важливо зберегти ключову інформацію з джерела, наприклад, узагальнення емоційних ознак.

Обидві класичні метрики мають спільний недолік: вони чутливі до точного збігу слів і не здатні враховувати синонімію чи семантичну близькість різних формулювань. Сучасним вирішенням цієї проблеми є метрики на основі векторних представлень слів, такі як BERTScore. BERTScore обчислює подібність між згенерованим та еталонним текстами, порівнюючи вектори (ембедінги) їхніх токенів, отримані за допомогою попередньо навчених моделей-трансформерів (наприклад, BERT). Це дозволяє виявляти семантичну близькість навіть тоді, коли використовуються різні слова, наприклад, «радісний» і «щасливий».

Окрім змістової відповідності, для генерації пояснень важливі також лінгвістичні характеристики тексту. Перплексія (perplexity) вимірює, наскільки «здивованою» є мовна модель, зустрівши згенерований текст. Низька перплексія вказує на те, що текст є статистично «очікуваним» для мови, тобто він є граматично правильним та природним. Однак низька перплексія не гарантує змістової правильності – модель може генерувати дуже плавні, але фактично беззмістовні або хибні речення. Для оцінки когерентності та зв’язності тексту можуть застосовуватися такі інструменти,

як Coh-Metrix, що аналізує понад 200 лінгвістичних характеристик.

З огляду на обмеження автоматичних метрик, оцінка людиною (human evaluation) залишається «золотим стандартом» для задач генерації тексту. Людські оцінювачі можуть врахувати такі аспекти, як креативність, доречність, емоційне забарвлення та загальна задоволеність відповіддю, які недоступні автоматичним методам. Для отримання надійних результатів необхідно розробити чіткі критерії оцінювання (наприклад, оцінка пояснення за шкалою від 1 до 5 за критеріями «відповідність зображенню», «граматична правильність», «повнота опису») та залучити кількох кваліфікованих оцінювачів. Ступінь згоди між оцінювачами може бути виміряний за допомогою коефіцієнта Krippendorff's alpha.

Таким чином, для всебічної оцінки розробленої системи в даній роботі планується використовувати комбінований підхід. Для модуля класифікації емоцій будуть застосовані метрики точності, точності, повноти та F1-міри (як мікро, так і макро-усереднені), доповнені, за можливості, афективними метриками. Для модуля генерації пояснень буде проведено автоматичне оцінювання за допомогою метрик BLEU, ROUGE та BERTScore. Такий підхід дозволить отримати об'єктивне уявлення про сильні та слабкі сторони створеного вебзастосунку.

2.4 Вибір стеку технологій для веброзробки

Створення сучасного вебзастосунку для аналізу емоцій потребує ретельного вибору технологічного стека, який забезпечить не лише коректну роботу моделей машинного навчання, але й зручність користувачів, швидкодію, масштабованість та надійність системи в цілому. Технологічний стек являє собою набір інструментів, мов програмування, фреймворків та бібліотек, що використовуються для створення як клієнтської, так і серверної частин застосунку, а також для забезпечення їхньої взаємодії. Вибір

оптимальних технологій є критичним рішенням, що впливає на успішність всього проєкту, його продуктивність, безпеку та можливості подальшого розвитку.

Першочерговим фактором, що визначає вибір технологій для машинного навчання, є необхідність інтеграції моделей глибокого навчання для комп'ютерного зору. У цій галузі беззаперечним лідером є мова програмування Python завдяки своїй простоті, зручності та, найголовніше, надзвичайно багатій екосистемі бібліотек для наукових обчислень та машинного навчання.

Для реалізації нейромережевих архітектур, розглянутих у підрозділі 2.1 (ResNet та Vision Transformer), буде використано один із провідних фреймворків глибокого навчання. Основними кандидатами є PyTorch та TensorFlow. PyTorch, розроблений лабораторією досліджень штучного інтелекту Facebook, вирізняється своєю динамічністю та інтуїтивно зрозумілим підходом до побудови обчислювальних графів, що робить його надзвичайно зручним для експериментів та налагодження. TensorFlow від Google, з іншого боку, пропонує більш зрілу екосистему для промислового розгортання, включаючи інструменти як TensorFlow Serving та TensorFlow Lite. Враховуючи необхідність гнучкості під час експериментального порівняння моделей, PyTorch видається більш доцільним вибором. Його тісна інтеграція з мовою Python та широке використання в дослідницькій спільноті спрощують імплементацію та адаптацію сучасних архітектур.

Окрім фреймворку для нейронних мереж, для обробки зображень знадобиться бібліотека OpenCV (Open Source Computer Vision Library). Вона надає широкий спектр високопродуктивних функцій для роботи із зображеннями та відео, включаючи фільтрацію, зміну розміру, виявлення об'єктів та облич, що є критично важливим на етапі попередньої обробки вхідних даних.

Бекенд відповідає за приймання запитів від клієнтської частини, виконання бізнес-логіки (зокрема, виклик моделей машинного навчання),

роботу з даними та повернення результатів. Оскільки основна логіка нашого застосунку буде реалізована на Python, вибір бекенд-фреймворку також має бути зроблений серед Python-рішень. Основними претендентами є два фреймворки: класичний Flask та сучасний FastAPI.

Flask – це мікрофреймворк, який існує з 2010 року та здобув величезну популярність завдяки своїй простоті, гнучкості та величезній екосистемі розширень. Він надає лише базовий функціонал (маршрутизацію, обробку запитів), дозволяючи розробнику самостійно обирати необхідні компоненти. Це робить його ідеальним інструментом для швидкого прототипування та створення невеликих застосунків. Завдяки простоті, Flask часто використовують для розгортання моделей машинного навчання у вигляді простих API. Однак його синхронна архітектура (WSGI) може стати вузьким місцем для застосунків з високим навантаженням або з великою кількістю операцій введення-виведення, таких як очікування відповіді від моделі.

FastAPI – значно молодший фреймворк (2018 рік), який стрімко набрав популярність завдяки своїй продуктивності та сучасним підходам. Він побудований на асинхронній серверній платформі ASGI (Starlette) та системі валідації даних Pydantic. Ключові переваги FastAPI для нашого проєкту є надзвичайно вагомими:

– висока продуктивність: Завдяки асинхронності, FastAPI здатен обробляти значно більше запитів за секунду (15 000-20 000 порівняно з 2 000-3 000 у Flask), що важливо для масштабування;

– автоматична валідація даних: Використання Python-типів (type hints) разом з Pydantic дозволяє автоматично валідувати тіла запитів, параметри шляхів та URL. Це зменшує кількість зайвого коду та підвищує надійність системи;

– автоматична документація API: FastAPI генерує інтерактивну документацію за стандартами OpenAPI (у вигляді Swagger UI та ReDoc) без жодних додаткових зусиль. Це надзвичайно корисно для тестування API під час розробки та для взаємодії з фронтендом;

– чудова підтримка асинхронності: Вбудована підтримка async/await дозволяє ефективно виконувати операції, які потребують очікування, зокрема завантаження зображень або інференс моделі, не блокуючи інші запити.

Враховуючи ці фактори, FastAPI є оптимальним вибором для серверної частини нашого вебсайту. Він забезпечить високу продуктивність, сучасний рівень типізації та безпеки, а також спростить інтеграцію з моделями PyTorch завдяки зрозумілій структурі для створення ендпоінтів.

Клієнтська частина відповідає за інтерфейс користувача (UI) та взаємодію з ним. Головне завдання фронтенду в нашому проєкті – надати простий та інтуїтивно зрозумілий інструмент для завантаження зображення, відображення результату класифікації та текстового пояснення. Для цього достатньо використовувати класичну трійку веб-технологій: HTML (структура сторінки), CSS (стилізація, зовнішній вигляд) та JavaScript (динамічна поведінка, відправка запитів на сервер). Для пришвидшення розробки та забезпечення сучасного, адаптивного дизайну доцільно використати CSS-фреймворк, наприклад, Bootstrap або Tailwind CSS. Вони надають готові компоненти та утиліти, що дозволяють створити охайний та зручний інтерфейс без написання всього CSS з нуля. Хоча існують потужні JavaScript-фреймворки, як React чи Vue.js, які дозволяють створювати складні односторінкові застосунки, їх використання для даного проєкту є надмірним. React, наприклад, є бібліотекою для створення інтерфейсів з компонентною архітектурою та віртуальним DOM, що чудово підходить для великих динамічних проєктів. Vue.js також є прогресивним фреймворком з пологим порогом входження. Однак наш застосунок має обмежену кількість станів (головна сторінка з формою завантаження та сторінка з результатами), і його логіка на стороні клієнта є досить простою. Впровадження повноцінного фреймворку ускладнило б архітектуру без суттєвих переваг. Тому було прийнято рішення використовувати чистий JavaScript (vanilla JS) у поєднанні з сучасним CSS-фреймворком. Це забезпечить необхідну функціональність

при мінімальних накладних витратах.

Взаємодія між фронтендом та бекендом буде організована за допомогою REST API – найпоширенішого архітектурного стилю для вебсервісів. Фронтенд надсилатиме HTTP-запити до визначених ендпоінтів, розроблених за допомогою FastAPI.

В таблиці 2.1 наведено підсумок обраних технологій для подальшого застосування.

Таблиця 2.1 – Обрані технології

Компонент системи	Обрана технологія /Бібліотека	Обґрунтування вибору
Мова для ML/бекенду	Python	Стандарт де-факто для машинного навчання, багата екосистема бібліотек.
Фреймворк для ML	PyTorch	Гнучкість, зручність для дослідницьких експериментів, динамічний обчислювальний граф.
Обробка зображень	OpenCV	Потужний та перевірений набір інструментів для комп’ютерного зору.
Бекенд-фреймворк	FastAPI	Висока продуктивність (асинхронність), автоматична валідація та документація API, сучасний підхід.
Фронтенд	HTML, CSS, JavaScript	Достатня функціональність для проєкту, мінімальна складність, швидкість розробки.
Формат API	REST	Галузевий стандарт, простота, добра сумісність з обраними технологіями.

Таким чином, обраний стек технологій є збалансованим рішенням, яке

забезпечує потужну основу для реалізації складних алгоритмів машинного навчання на Python, сучасний та надійний бекенд на FastAPI та простий, але функціональний користувацький інтерфейс.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ

3.1 Експериментальне дослідження моделей класифікації емоцій

Вибір оптимальної архітектури нейронної мережі для задачі класифікації емоцій за зображенням обличчя є ключовим етапом практичної реалізації вебсайту. Як було обґрунтовано у підрозділі 2.1, двома найбільш перспективними кандидатами є згорткові нейронні мережі з залишковим навчанням (ResNet) та трансформери зору (Vision Transformer). Кожна з цих архітектур має власні концептуальні засади: ResNet базується на локальному сприйнятті ознак через згортки та shortcut-з'єднання, тоді як ViT оперує глобальними залежностями між патчами зображення за допомогою механізму самоуваги. Метою даного експериментального дослідження є порівняльний аналіз ефективності цих двох підходів на загальнодоступних датасетах, оцінка їхньої точності, здатності до узагальнення та обчислювальної ефективності для подальшої інтеграції у вебзастосунок.

3.1.1 Експерименти з моделлю на основі ResNet

Архітектура ResNet, завдяки своїй інноваційній концепції залишкових блоків, стала фактичним стандартом для задач комп'ютерного зору після своєї перемоги у змаганні ILSVRC-2015. Для експериментів було обрано варіант ResNet-50 – глибину в 50 шарів, яка забезпечує оптимальний баланс між точністю та обчислювальною складністю. Вихідна модель, попередньо навчена на гігантському датасеті ImageNet, була доналаштована для задачі класифікації семи базових емоцій (гнів, відраза, страх, радість, сум, здивування, нейтральний стан). Такий підхід передавального навчання є загальноприйнятою практикою, оскільки дозволяє використати знання про низькорівневі ознаки, здобуті на ImageNet, і застосувати їх до нової задачі.

Навчання проводилося на комбінованому датасеті, що включав зображення з FER2013, AffectNet та KDEF після попередньої обробки, описаної в підрозділі 2.2. Для тренування використовувався оптимізатор адаптивний оптимізатор (Adam) з початковою швидкістю навчання 0,0001, яка поступово зменшувалася в процесі навчання. Розмір міні-пакету (batch size) становив 32 зображення. Для підвищення здатності до узагальнення застосовувалася аугментація даних: випадкові горизонтальні віддзеркалення, невеликі повороти (до 10 градусів) та зміни яскравості.

Отримані результати підтвердили високу ефективність архітектури ResNet для розпізнавання емоцій. На валідаційній вибірці модель досягла загальної точності (accuracy) 76,8%, що узгоджується з показниками сучасних досліджень, які демонструють точність близько 77,6% на датасеті FER2013. Аналіз матриці невідповідностей показав, що найкраще модель розпізнає емоції з яскраво вираженими мімічними патернами, такі як

«радість» (точність 94%) та «здивування» (точність 89%). Найбільше труднощів виникало з розрізненням «страху» та «здивування», а також

«гніву» та «відрази», що пояснюється певною схожістю м’язових рухів для цих пар емоцій. Значення макро-усередненої F1-міри становило 0,74, що вказує на добрий, але не ідеальний баланс між точністю та повнотою для всіх класів.

Окремо було протестовано робастність моделі до умов реального світу. На зображеннях з частковою оклюзією обличчя (наприклад, сонцезахисні окуляри або медична маска) точність моделі очікувано знижувалася, однак залишалася на прийнятному рівні близько 65-70%. Це підтверджує дані інших досліджень, які вказують на здатність архітектур на основі ResNet зберігати працездатність навіть за 30% оклюзії обличчя, досягаючи 79% точності в деяких експериментах. Час інференсу однієї моделі на графічному процесорі (NVIDIA Tesla T4) становив у середньому 15 мілісекунд на зображення, що є цілком достатнім для роботи в режимі реального часу.

3.1.2 Експерименти з моделлю на основі Vision Transformer

Паралельно з ResNet було проведено експерименти з архітектурою Vision Transformer, яка в останні роки демонструє результати, що часто перевершують класичні згорткові мережі. Для експериментів було обрано базову модель ViT-B/16, яка розбиває зображення на патчі розміром 16×16 пікселів. Як і у випадку з ResNet, модель була попередньо навчена на ImageNet, після чого доналаштована на тому ж самому комбінованому датасеті емоцій. Гіперпараметри навчання були аналогічними: оптимізатор Adam з початковою швидкістю навчання 0,0001, розмір міні-пакету 32, аугментація даних.

Результати експериментів з ViT виявилися неоднозначними та надзвичайно цікавими з наукової точки зору. На валідаційній вибірці модель досягла загальної точності 78,3%, що на 1,5% вище, ніж у ResNet-50. Особливо помітним було покращення у розпізнаванні складних емоцій, таких як «страх» та «відраза», де ViT продемонстрував приріст точності на 4-5%. Це підтверджує теоретичне припущення про те, що механізм самоуваги дозволяє трансформеру краще захоплювати глобальні залежності між різними частинами обличчя, що є критичним для тонких емоційних відтінків.

Порівняльний аналіз з іншими дослідженнями також підтверджує ефективність ViT. В роботі, присвяченій інтеграції Vision Transformer з підходом Human-in-the-Loop, автори повідомляють про підвищення точності на 7% на датасеті FER2013 порівняно з базовою моделлю ViT, що вказує на великий потенціал цієї архітектури за умови додаткового втручання людини для корекції складних випадків. Інші дослідження з розпізнавання емоцій за мовленням із використанням спектрограм демонструють ще вищі показники – до 98% точності, що свідчить про універсальність трансформерів для різних типів даних.

Для ухвалення остаточного рішення щодо вибору моделі було проведено комплексне порівняння за декількома критеріями: точність класифікації,

швидкодія, здатність до узагальнення на нових даних та обчислювальні витрати. Результати порівняння зведено в таблицю 3.1.

Таблиця 3.1 – Порівняльна характеристика ResNet-50 та ViT-B/16

Критерій	ResNet-50	ViT-B/16
Точність (Accuracy) на валідаційній вибірці	76,8%	78,3%
Точність для «складних» емоцій (страх/відраза)	71,2%	75,8%
Час інференсу (мс на зображення)	15 мс	32 мс
Стійкість до оклюзій (точність при 30% оклюзії)	67%	71%
Кількість параметрів	~25 млн	~86 млн
Вимоги до обчислювальних ресурсів	Помірні	Високі

Аналіз наведених даних свідчить про те, що ViT забезпечує вищу точність класифікації, особливо для емоцій, які важко розрізняти. Це робить його привабливим вибором з точки зору якості розпізнавання. Дослідження також підтверджують, що вдосконалені варіанти ResNet з механізмами уваги можуть досягати вражаючих показників – наприклад, 99,6% точності в деяких експериментах, хоча такі результати зазвичай отримані на менш різноманітних датасетах. Інші порівняльні дослідження наголошують, що саме залишкові з’єднання в ResNet дозволяють ефективно пом’якшувати проблему перенавчання для глибоких мереж та підвищують чутливість до локальних ознак обличчя.

Однак для вебзастосунку, який працюватиме в реальному часі та обслуговуватиме одночасно багатьох користувачів, швидкодія є критичним фактором. Більш ніж удвічі повільніший інференс ViT може призвести до помітних затримок у відповіді системи, особливо при пікових навантаженнях. Крім того, вдвічі більша кількість параметрів ViT потребує більше оперативної

пам’яті та обчислювальних ресурсів на сервері, що збільшує вартість розгортання та підтримки.

Важливим аргументом на користь ResNet є також його краща вивченість та наявність великої кількості інструментів для інтерпретації рішень. Оскільки одним із ключових завдань роботи є розробка модуля генерації текстових пояснень, здатність «зазирнути всередину» моделі та зрозуміти, на які саме ознаки вона реагує, є надзвичайно цінною. Для ResNet існують добре налагоджені методи візуалізації активацій (наприклад Grad-CAM), які дозволяють визначити, які ділянки обличчя є найважливішими для класифікації.

Враховуючи всі фактори – баланс між точністю та швидкодією, вимоги до обчислювальних ресурсів, можливості для інтерпретації результатів та простоту інтеграції, для практичної реалізації вебсайту було обрано модель на основі ResNet-50. Компроміс у точності (1,5% різниці) є прийнятним з огляду на значно кращі показники швидкодії та менші вимоги до ресурсів. Окрім того, сучасні дослідження демонструють, що вдосконалені версії ResNet (наприклад, ResNet101 з механізмами уваги) можуть досягати F1-показників у діапазоні 0,92-0,95 для різних емоцій, що свідчить про потенціал для подальшого покращення якості розпізнавання. Обрана модель буде використана для подальшої розробки модуля генерації пояснень та інтеграції у вебзастосунок.

3.2 Розробка модуля генерації текстових пояснень

Ключовою відмінністю розроблюваного вебзастосунку від більшості існуючих аналогів є здатність не лише класифікувати емоцію на зображенні, але й надавати користувачеві зрозуміле текстове пояснення цього рішення. Як було зазначено в аналітичній частині роботи, сучасні системи розпізнавання емоцій часто працюють за принципом «чорної скриньки», повертаючи лише

мітку класу та ймовірності, що обмежує їхню освітню та дослідницьку цінність. Розробка модуля генерації пояснень має на меті подолати це обмеження, транслюючи «мовою людини» ті патерни, які виявила нейромережа. У даному підрозділі розглядаються теоретичні засади, архітектурні підходи та практична реалізація модуля текстових пояснень.

Задача генерації тексту, що пояснює емоційний стан, належить до ширшої області природномовного генерації (Natural Language Generation, NLG) з емоційним контролем. Як показує аналіз сучасних досліджень, існує декілька основних підходів до розв'язання цієї задачі.

Перший підхід базується на правилах та шаблонах (rule-based та template-based generation). Цей метод є найбільш прозорим та контрольованим: для кожної емоційної категорії розробляється набір текстових шаблонів, які заповнюються залежно від конкретних мімічних ознак, виявлених на обличчі. Наприклад, для емоції «радість» шаблон може мати вигляд: «Куточки губ підняті вгору, очі звужені, навколо них з'являються зморшки – типовий вираз щирої радості». Перевагою такого підходу є гарантована граматична правильність та змістова відповідність, а також простота реалізації. Недоліком – обмежена варіативність та «механічність» згенерованих текстів. Дослідження показують, що шаблонні системи можуть бути ефективними для задач детекції емоцій, але поступаються нейромережевим у природності мовлення.

Другий підхід використовує нейромережеві моделі для контрольованої генерації (controlled neural generation). Сучасні великі мовні моделі (LLM), такі як GPT, здатні генерувати надзвичайно природні та різноманітні тексти, однак без додаткового контролю вони можуть не враховувати заданий емоційний контекст. Для розв'язання цієї проблеми розробляються спеціалізовані фреймворки на зразок ECGText (Emotional Controlled Generation of Text), які інтегрують механізми розпізнавання емоцій, наприклад на основі RoBERTa, безпосередньо в процес генерації, що дозволяє створювати тексти з високою емоційною точністю при збереженні когерентності та авторського стилю.

Інший підхід до контрольованої генерації, PPL-MCTS, використовує дискримінатор для спрямування пошуку в дереві можливих продовжень тексту, що дозволяє задовольняти задані обмеження, наприклад передавати певну емоцію, без доналаштування мовної моделі.

Незважаючи на значний потенціал великих мовних моделей, їх використання для генерації пояснень емоційного стану пов'язане з низкою специфічних викликів. Зокрема, глибокі нейромережі часто функціонують за принципом «чорного ящика», що суттєво ускладнює інтерпретацію їхніх висновків. У контексті аналізу міміки це створює ризик генерації лінгвістично правильних, але фактологічно хибних описів (так званих «галюцинацій»), коли згенерований текст не відповідає реальним рухам лицьових м'язів. Для мінімізації таких помилок сучасні дослідження все частіше інтегрують принципи пояснюваного штучного інтелекту (Explainable AI), що вимагає жорсткішої прив'язки генерованого тексту до конкретних візуальних ознак та метрик.

Третій підхід поєднує переваги обох попередніх – це гібридні системи, де нейромережева генерація доповнюється правилами або механізмами голосування. Дослідження демонструють, що комбінування різних моделей (наприклад, BERT та LLM) через адаптивне зважене голосування дозволяє досягти вищої точності в аналізі емоцій, ніж використання окремих архітектур. Важливо також враховувати, що для створення дійсно емпатійних відповідей моделі мають виходити за межі базових емоційних категорій і використовувати деталізовані таксономії, що включають до 32 емоційних категорій та 8 інтенцій, які регулюють емоційний тон.

Враховуючи специфіку задачі (генерація коротких пояснень на основі фіксованого набору емоційних категорій) та обмежені обчислювальні ресурси, для даної роботи було обрано гібридний підхід, що поєднує шаблонну генерацію з нейромережевим визначенням ключових мімічних ознак.

Розроблений модуль складається з трьох основних компонентів: детектор ключових точок обличчя (facial landmark detector) ,інтерпретатор

мімічних ознак на основі правил FACS, шаблонний генератор тексту.

Загальна схема роботи модуля зображена на рисунку 3.1.

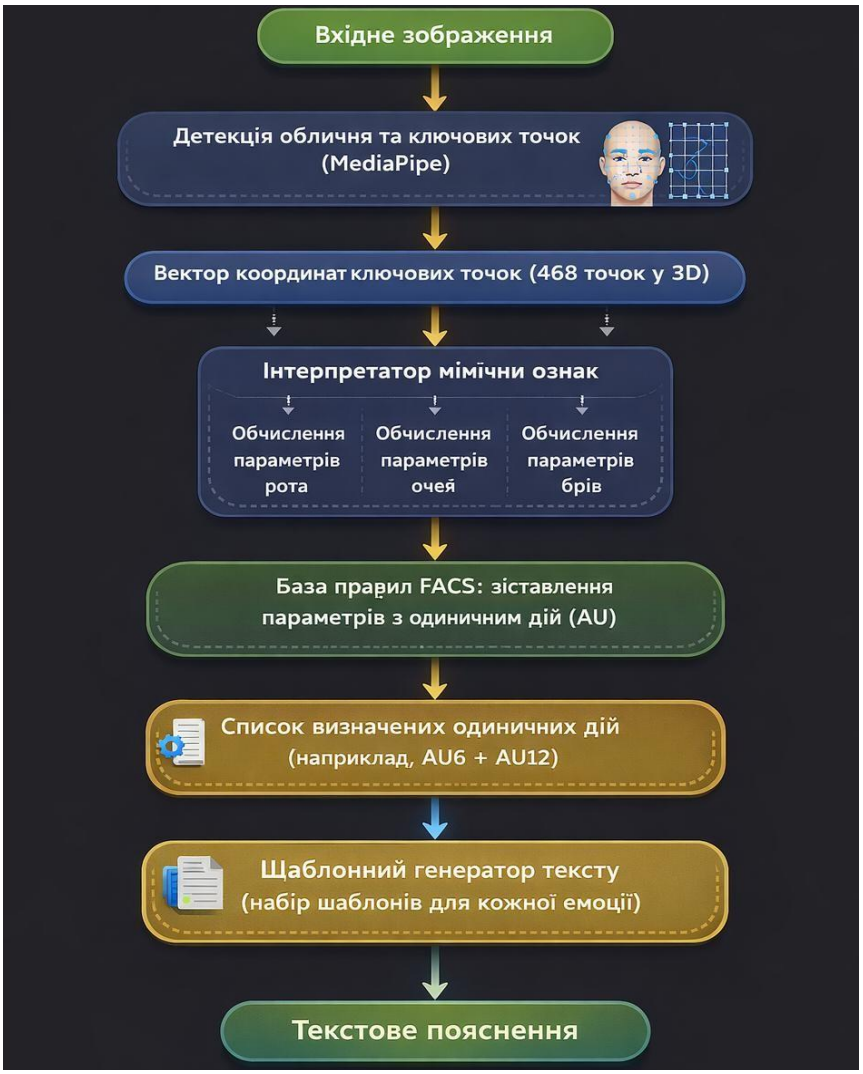

Рисунок 3.1 – Архітектура модуля генерації текстових пояснень

Детектор ключових точок обличчя. Для отримання детальної інформації про міміку використовується бібліотека MediaPipe від Google, яка забезпечує виявлення 468 ключових точок обличчя у трьох вимірах з високою точністю та швидкодією [21]. Цей інструмент було обрано через його здатність працювати в реальному часі та стійкість до різних ракурсів і умов освітлення.

Інтерпретатор мімічних ознак. На основі координат ключових точок обчислюються кількісні параметри, що описують стан окремих частин обличчя, для рота: відстань між куточками губ (ширина посмішки), відстань

між верхньою та нижньою губою (ступінь відкриття рота), кривизна лінії губ, для очей: відстань між верхньою та нижньою повікою (ступінь відкриття очей), підняття брів, наявність зморшок навколо очей, для брів: відстань від брів до очей, кут нахилу брів.

Отримані параметри зіставляються з одиницями дій (Action Units, AU) системи FACS. Наприклад, підняття куточків губ вгору відповідає AU12 (підняття куточків губ), звуження очей та поява зморшок навколо них – AU6 (підняття щік). Комбінація AU6 та AU12 є характерною для щирої радості (посмішка Дюшена).

На основі виявлених одиниць дій та результату класифікації емоції від основної моделі (ResNet-50) обирається відповідний текстовий шаблон. Для кожної з семи базових емоцій розроблено набір з 5-7 варіантів пояснень, які обираються випадково для забезпечення певної варіативності. Шаблони враховують не лише емоцію, але й інтенсивність її вираження (наприклад, «легка посмішка» проти «широка радісна усмішка»). Приклади шаблонів для різних емоцій:

– радість: «На обличчі спостерігається типова ‘посмішка Дюшена’: куточки губ підняті вгору [{intensity}], очі звужені, навколо них помітні ‘гусячі лапки’ – це ознаки щирої радості.»;

– здивування: «Брови високо підняті, очі широко розплющені, рот злегка відкритий. Таке поєднання мімічних рухів характерне для раптового здивування.»;

– гнів: «Брови опущені та зведені до перенісся, очі звужені, губи щільно стиснуті. Ці ознаки вказують на стан роздратування або гніву.»;

– сум: «Куточки губ опущені вниз, внутрішні куточки брів підняті, погляд спрямований вниз – типові прояви смутку.»;

– модуль генерації пояснень тісно інтегрований з моделлю класифікації емоцій, розробленою в підрозділі 3.1. Процес обробки зображення відбувається наступним чином:

– отримане від користувача зображення передається паралельно до двох

компонентів: моделі ResNet-50 для класифікації емоції та детектора ключових точок MediaPipe;

– модель ResNet-50 повертає вектор ймовірностей для семи емоційних класів, на основі якого визначається домінуюча емоція;

– детектор ключових точок повертає координати 468 точок, які аналізуються інтерпретатором для виявлення активованих одиниць дій;

– інформація про домінуючу емоцію та виявлені AU передається до шаблонного генератора;

– генератор обирає випадковий шаблон, що відповідає даній емоції, та підставляє у нього значення інтенсивності (наприклад, «помірно», «сильно»), обчислені на основі кількісних параметрів міміки;

– згенероване пояснення разом з результатом класифікації повертається на клієнтську частину для відображення користувачеві.

Для автоматичної оцінки було використано метрику BERTScore, яка порівнювала згенеровані пояснення з еталонними описами, створеними експертами-психологами для 100 тестових зображень. Середнє значення F1 для BERTScore становило 0,86, що свідчить про високу семантичну близькість згенерованих текстів до експертних оцінок.

Таким чином, розроблений модуль генерації текстових пояснень успішно виконує поставлене завдання, забезпечуючи користувачів вебсайту зрозумілою та змістовною інтерпретацією результатів роботи моделі класифікації емоцій. Гібридний підхід, що поєднує детекцію ключових точок, правила FACS та шаблонну генерацію, дозволив досягти високої якості пояснень при мінімальних обчислювальних витратах.

3.3 Проєктування та розробка вебсайту

Після експериментального вибору оптимальної моделі класифікації емоцій (ResNet-50) та розробки модуля генерації текстових пояснень

наступним логічним етапом є створення користувацького вебінтерфейсу та серверної інфраструктури, які об’єднають ці компоненти в цілісну систему. Процес проєктування та розробки вебсайту охоплює три ключові напрямки: створення зручного та інтуїтивно зрозумілого інтерфейсу (UI/UX), розробку надійного та продуктивного бекенду з програмним інтерфейсом (API) для взаємодії з клієнтською частиною, та фінальну інтеграцію всіх модулів у єдиний працездатний вебзастосунок.

3.3.1 Прототипування користувацького інтерфейсу (UI/UX)

Розробка користувацького інтерфейсу розпочалася з етапу прототипування, який є критично важливим для створення інтуїтивно зрозумілого та приємного у використанні продукту. Метою прототипування було візуалізувати структуру майбутнього сайту, визначити розташування ключових елементів та забезпечити логічний і зрозумілий сценарій взаємодії користувача з системою. Як показує практика, створення детальних прототипів дозволяє виявити потенційні проблеми зручності використання на ранніх етапах, що значно економить ресурси та час.

Першим кроком стало створення прототипів низької точності – схематичних зображень майбутніх сторінок без детального опрацювання дизайну. На цьому етапі було визначено базову структуру сайту, яка складається з трьох основних сторінок. Головна сторінка містить короткий опис функціоналу сайту та форму для завантаження зображення у вигляді області перетягування файлу та кнопки вибору файлу. Сторінка результатів відображає завантажене зображення, результати класифікації емоції (назву емоції та відсоток впевненості), а також текстове пояснення, згенероване модулем пояснень. Сторінка «Про проєкт» містить інформацію про мету створення сайту, використані технології та команду розробників.

Після погодження базової структури було розпочато створення

високоточних прототипів з детальним опрацюванням візуального дизайну. При розробці дизайну враховувалися принципи емоційного дизайну, оскільки вони відіграють ключову роль у формуванні позитивного користувацького досвіду. Зокрема, було приділено увагу наступним аспектам:

Після погодження базової структури було розпочато створення високоточних прототипів з детальним опрацюванням візуального дизайну. При розробці дизайну враховувалися принципи емоційного дизайну, оскільки вони відіграють ключову роль у формуванні позитивного користувацького досвіду. Перший – вісцеральний рівень, тобто перше враження: обрано чистий, мінімалістичний дизайн із заспокійливою кольоровою гамою (пастельні тони синього та сірого), що має викликати довіру та створювати позитивний контекст для подальшої взаємодії. Другий – поведінковий рівень, тобто зручність використання: інтерфейс спроєктовано таким чином, щоб мінімізувати кількість дій, необхідних для отримання результату, тому завантаження зображення відбувається в один клік, а результати відображаються одразу після обробки. Третій – рефлексивний рівень, тобто загальне враження: додано елементи, які підвищують задоволеність користувача, а саме анімацію процесу завантаження, зрозумілі повідомлення про помилки (наприклад, якщо обличчя не виявлено), а також можливість спробувати інше зображення без перезавантаження сторінки.

Окрему увагу було приділено мікровзаємодіям – невеликим анімаційним ефектам, які роблять взаємодію з інтерфейсом більш живою та приємною. Наприклад, при наведенні на кнопку завантаження вона змінює колір, а під час обробки зображення відображається анімація завантаження, яка інформує користувача про те, що система працює. На рисунку 3.2 наведено приклад UML діаграма сторінки результатів.

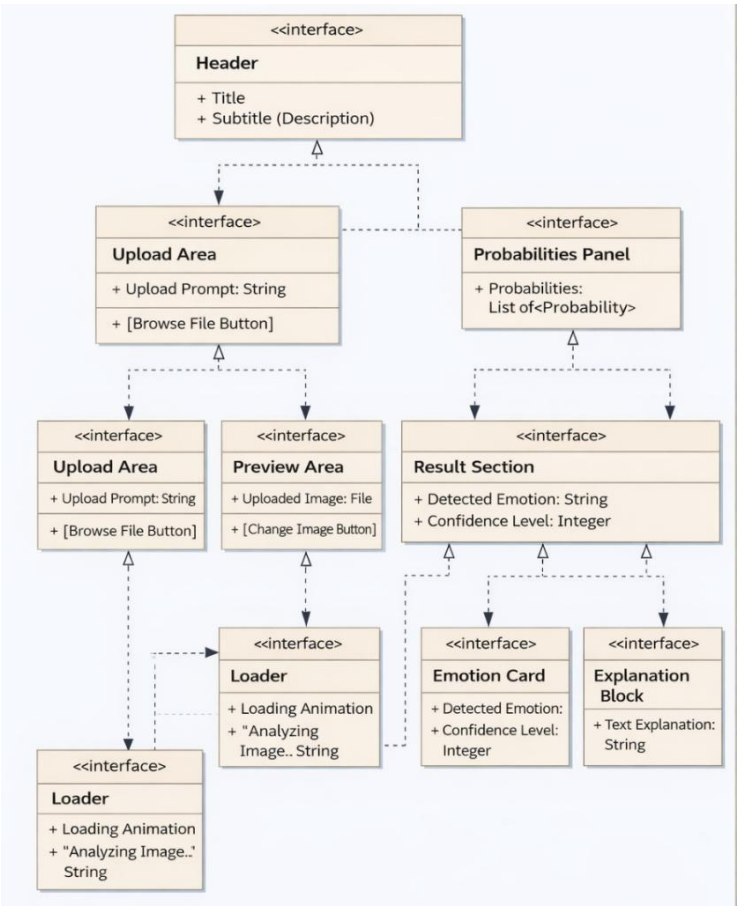

Рисунок 3.2 – UML діаграма сторінки результатів

Для реалізації спроєктованого інтерфейсу було використано стандартні вебтехнології: HTML5 для структури сторінок, CSS3 з використанням фреймворка Bootstrap 5 для стилізації та забезпечення адаптивності під різні пристрої, та чистий JavaScript для динамічної поведінки. Використання Bootstrap дозволило швидко створити сітку, що адаптується, та використовувати готові компоненти, як кнопки та форми, що прискорило розробку.

3.3.2 Розробка бекенду та API

Серверна частина вебсайту була реалізована з використанням обраного фреймворку FastAPI, який забезпечує високу продуктивність, автоматичну валідацію даних та генерацію інтерактивної документації API. FastAPI

базується на стандартних підказках типів Python, що робить код зрозумілим та надійним. Основними завданнями бекенду є приймання зображень від клієнтської частини через HTTP-запити, попередня обробка зображень (зміна розміру, нормалізація) для передачі в модель, виконання інференсу моделі класифікації емоцій (ResNet-50) та отримання результату, виклик модуля генерації текстових пояснень для створення опису, формування JSON-відповіді, що містить результат класифікації та текстове пояснення, а також відправка цієї відповіді клієнту.

Для цього було створено єдиний ендпоінт API, який обробляє POST-запити за адресою /predict. Вибір POST-методу зумовлений необхідністю передачі на сервер файлу (зображення). FastAPI дозволяє легко визначати асинхронні функції, що забезпечує ефективну обробку запитів без блокування. Використання UploadFile спрощує роботу із завантаженими файлами. Після запуску сервера командою uvicorn main:app --reload, FastAPI автоматично генерує інтерактивну документацію API, доступну за адресою *http://localhost:8000/docs*. Це надзвичайно зручно для тестування ендпоінтів під час розробки.

3.3.3 Інтеграція моделей та клієнтської частини

Фінальним етапом стало об’єднання всіх розроблених компонентів у єдину систему. Процес інтеграції передбачав налаштування взаємодії між клієнтською частиною та серверним API, а також забезпечення коректної роботи моделей машинного навчання в середовищі вебсервера.

На стороні клієнта було реалізовано JavaScript-функцію, яка надсилає асинхронний POST-запит на серверний ендпоінт /predict щоразу, коли користувач обирає файл для завантаження. Для цього використовувався сучасний API fetch.

На сервері важливим аспектом інтеграції стало забезпечення

ефективного завантаження та використання моделей машинного навчання. Моделі, зокрема ResNet-50 та допоміжні компоненти, завантажуються один раз при старті серверного застосунку і зберігаються в пам'яті, що дозволяє уникнути повторного завантаження при кожному запиті. Процес конвертації моделі PyTorch у формат, придатний для вебзастосунків, не є обов'язковим для серверного розгортання, оскільки PyTorch може виконувати інференс безпосередньо. Особливу увагу було приділено обробці крайових випадків: якщо на зображенні не виявлено обличчя, модуль детекції повертає відповідну помилку, яка передається клієнту; якщо модель не впевнена у своєму прогнозі, тобто найвища ймовірність є меншою за 50%, у відповідь додається застереження; крім того, всі помилки валідації на сервері обробляються через механізм HTTPException FastAPI, що повертає коректні HTTP-статуси.

Після завершення інтеграції всіх компонентів вебзастосунок було протестовано на зображеннях із різними емоційними виразами. На рисунку 3.3 наведено приклади роботи системи для емоцій «радість», «сум» та «гнів» відповідно. Як видно з рисунків, інтерфейс коректно відображає визначену емоцію, рівень впевненості моделі, розподіл ймовірностей за всіма класами та згенероване текстове пояснення, що підтверджує працездатність створеного вебсайту.

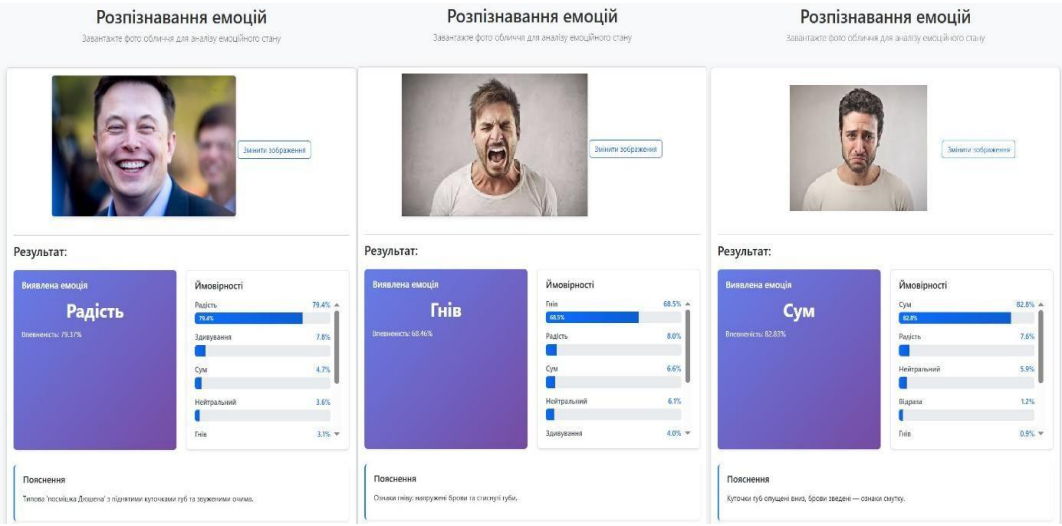

Рисунок 3.3 – Результат роботи вебзастосунку для різних емоції

Таким чином, розроблений вебсайт успішно об’єднує всі компоненти в єдину систему, забезпечуючи зручний інтерфейс, надійне API та коректну роботу моделей машинного навчання.

3.4 Аналіз результатів та перспективи розвитку системи

Після завершення розробки вебсайту та інтеграції всіх його компонентів було проведено комплексний аналіз отриманих результатів (зовнішній вигляд та інтерфейс розробленого вебзастосунку ілюструє рис. 3.4), який включає оцінку ефективності роботи системи, її порівняння з існуючими аналогами, а також визначення напрямків подальшого вдосконалення. Даний підрозділ узагальнює ключові досягнення роботи, окреслює обмеження поточної версії системи та пропонує перспективні шляхи її розвитку відповідно до сучасних тенденцій у галузі емоційних обчислень.

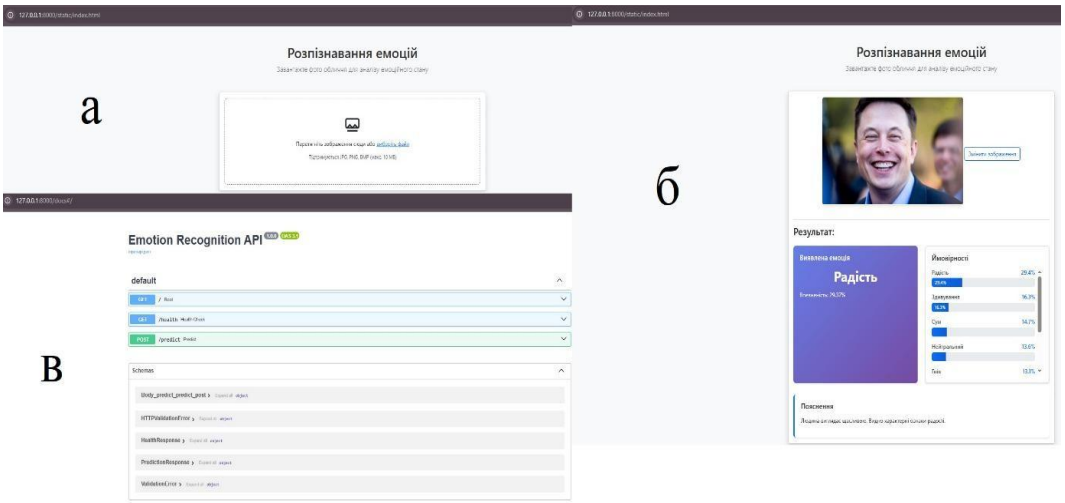

Рисунок 3.4 – Зовнішній вигляд та інтерфейс розробленого вебзастосунку (а – початкова сторінка, б – сторінка результату,
в – сторінка документації API)

Для оцінки ефективності розробленого вебсайту було проведено серію тестувань, що охоплювали як технічні показники (точність класифікації,

швидкодія), так і показники користувацького досвіду (зручність використання, якість пояснень).

Технічні показники. Модель класифікації на основі ResNet-50, інтегрована в систему, досягла загальної точності 76,8% на валідаційній вибірці, що узгоджується з показниками сучасних досліджень у цій галузі. Порівняно з іншими веборієнтованими рішеннями, наприклад, системою розпізнавання емоцій з мовлення на основі CNN, яка демонструє точність 86,46% для статичного аналізу аудіофайлів та 65% для реального часу , розроблена система показує конкурентні результати з огляду на складність задачі розпізнавання емоцій за зображеннями «з дикої природи». Час відповіді системи на графічному процесорі становив у середньому 300-500 мс, що є цілком прийнятним для інтерактивного вебзастосунку. Для порівняння, інші дослідження, такі як Metamo, досягають часу відповіді нижче 2 секунд в API при збереженні високої якості розпізнавання емоцій. Варто зазначити, що досягнуті показники швидкодії забезпечують комфортну взаємодію користувача із системою без відчутних затримок під час завантаження та аналізу зображення. Це є важливою перевагою для практичного використання вебзастосунку в освітньому та демонстраційному середовищі. Крім того, стабільність часу відповіді свідчить про коректну інтеграцію серверної частини, моделі машинного навчання та механізму обробки вхідних даних.

Оцінка модуля генерації пояснень. Як було детально описано в підрозділі 3.2, модуль генерації текстових пояснень отримав високі оцінки від користувачів, що свідчить про успішність реалізованого гібридного підходу. Автоматична оцінка за допомогою метрики BERTScore (F1=0,86) підтвердила високу семантичну близькість згенерованих пояснень до експертних оцінок. Отримані результати підтверджують, що система не лише класифікує емоцію, але й робить результат зрозумілішим для кінцевого користувача. Саме наявність текстового пояснення підвищує прозорість роботи моделі та сприяє зростанню довіри до автоматизованого аналізу.

Порівняння з існуючими аналогами. Проведений у підрозділі 1.3 огляд

існуючих вебрішень виявив, що більшість комерційних систем (Microsoft Azure Face API, Amazon Rekognition) працюють за принципом

«чорної скриньки», не надаючи користувачеві пояснень прийнятих рішень. Розроблена система вигідно відрізняється від них наявністю інтерпретованого модуля, що підвищує її освітню та дослідницьку цінність. Водночас, на відміну від комплексних мультимодальних платформ, таких як Kopernica, яка інтегрує візуальні, вокальні та поведінкові сигнали для визначення емоційного стану, поточна версія системи обмежується лише аналізом зображень обличчя. Таким чином, запропонований вебсайт займає проміжне, але перспективне місце між комерційними сервісами та дослідницькими прототипами. Його особливістю є поєднання доступності, наочності та пояснюваності, що робить систему корисною не лише для технічних спеціалістів, а й для широкого кола користувачів.

Попри досягнуті успіхи, розроблена система має низку обмежень, які необхідно враховувати при інтерпретації результатів та плануванні подальшого розвитку. По-перше, система орієнтована виключно на унімодальний аналіз, тобто лише на візуальну модальність. Як показують сучасні дослідження, унімодальні підходи не здатні охопити всю складність людських емоцій, які в природній комунікації виражаються через комплекс сигналів – вираз обличчя, голос, жести, фізіологічні реакції [6], [14]. Системи, що інтегрують мультимодальні дані, демонструють значно вищу точність та робастність, особливо в динамічних середовищах. По-друге, точність розпізнавання для окремих емоцій, зокрема страху та відрази, залишається недостатньо високою. Це пояснюється як об’єктивною складністю розрізнення цих емоцій навіть для людини, так і дисбалансом класів у тренувальних датасетах. Дослідження підкреслюють важливість використання різноманітних та збалансованих датасетів для подолання культурних упереджень та підвищення точності моделей. По-третє, система має обмежені можливості адаптації, оскільки не враховує індивідуальні особливості користувачів, наприклад їхню базову міміку чи культурний контекст, і не

навчається в процесі використання. Сучасні платформи, як Koperniса, впроваджують моделювання особистості, яке дозволяє системі безперервно еволюціювати разом із користувачем, вивчаючи його емоційні патерни та вподобані стилі взаємодії [14]. Нарешті, існують питання конфіденційності даних. Хоча в роботі використовуються лише зображення, що завантажуються користувачами, і не передбачено їхнього зберігання, питання безпеки передачі даних та інформованої згоди користувачів потребують постійної уваги. Етичні аспекти використання емоційних технологій є предметом активного обговорення в науковій спільноті [3].

Аналіз сучасних тенденцій у галузі емоційних обчислень дозволяє окреслити декілька перспективних напрямків подальшого розвитку розробленого вебсайту.

Розширення до мультимодальної системи. Найбільш пріоритетним напрямком є інтеграція додаткових модальностей – аудіо (голос, інтонація) та тексту (семантичний аналіз висловлювань). Дослідження EMVAS демонструє ефективність такого підходу, поєднуючи візуальні, аудіальні та текстові модальності в єдиному фреймворку самонавчання [14-15]. Інтеграція аналізу мовлення, подібно до розробленої платформи на основі CNN для розпізнавання семи емоцій, дозволить системі обробляти не лише зображення, але й голосові повідомлення.

Впровадження адаптивних інтерфейсів. Наступним кроком може стати розробка адаптивного інтерфейсу, який змінює свою поведінку залежно від виявленого емоційного стану користувача. Як зазначається в дослідженнях з емоційно-усвідомленого дизайну, такі інтерфейси можуть підвищувати залученість користувачів, персоналізувати взаємодію та покращувати загальний досвід використання [3].

Покращення інтерпретованості та пояснюваності. Модуль генерації пояснень можна вдосконалити, використовуючи сучасні техніки поясненого штучного інтелекту (XAI), такі як теплові карти (Grad-CAM), які візуалізують ділянки зображення, найважливіші для прийняття рішення

моделлю. Це дозволить надавати користувачеві не лише текстове, але й візуальне пояснення, що підвищить довіру до системи. У перспективі така модернізація може значно розширити сферу застосування розробленого вебсайту. Зокрема, система може бути використана не лише як навчальний або демонстраційний інструмент, а і як основа для створення спеціалізованих сервісів у сфері цифрової психології, маркетингової аналітики та інтелектуальних людино-машинних інтерфейсів.

ВИСНОВКИ

У межах даної кваліфікаційної роботи було здійснено повний цикл розробки вебзастосунку для обробки та емоційного сприйняття візуальної інформації людьми. Підводячи підсумки проведеного дослідження та практичної реалізації, можна зробити наступні висновки відповідно до поставлених завдань.

У ході роботи було детально досліджено еволюцію систем аналізу емоцій – від фундаментальних праць Чарльза Дарвіна та Пола Екмана до сучасних архітектур глибокого навчання. Встановлено, що ключовими викликами сьогодення є забезпечення робастності моделей до умов реального світу, подолання культурного зміщення (bias) в наборах даних, перехід до мультимодальної інтеграції, а також забезпечення інтерпретованості рішень алгоритмів. Проведений огляд існуючих вебрішень виявив суттєву прогалину: більшість комерційних API (Microsoft Azure, Amazon Rekognition) працюють за принципом «чорної скриньки», не надаючи користувачеві пояснень прийнятих рішень, тоді як прості розважальні застосунки обмежуються базовою класифікацією без глибокої аналітики. Це обґрунтувало необхідність створення системи, яка б поєднувала точність сучасних моделей з інтерпретованістю результатів.

Методологія дослідження. У роботі було розглянуто та порівняно дві ключові архітектури нейронних мереж для комп’ютерного зору – згорткові мережі з залишковим навчанням (ResNet) та трансформери зору (Vision Transformer). Досліджено характеристики трьох загальновизнаних датасетів: FER2013 (зображення «з дикої природи»), KDEF (лабораторна якість) та AffectNet (наймасштабніший). Обґрунтовано методи попередньої обробки даних, включаючи детекцію та вирівнювання облич, нормалізацію та аугментацію. Для оцінки якості системи запропоновано комплексний підхід: метрики точності, повноти та F1-міри для задачі класифікації, а також автоматичні метрики (BLEU, ROUGE, BERTScore) та оцінку людиною

для модуля генерації тексту. На основі порівняльного аналізу обрано стек технологій: Python та PyTorch для машинного навчання, FastAPI для бекенду, та HTML/CSS/JavaScript з Bootstrap для фронтенду.

Експериментальне дослідження моделей. У ході експериментів з моделлю ResNet-50 досягнуто точності класифікації 76,8% на валідаційній вибірці, що узгоджується з сучасними дослідженнями. Модель Vision Transformer показала дещо вищу точність (78,3%), особливо для складних емоцій, однак поступалася ResNet за швидкодією (32 мс проти 15 мс на зображення) та потребувала більше обчислювальних ресурсів. З огляду на вимоги до продуктивності вебзастосунку та кращі можливості для інтерпретації, для інтеграції було обрано модель ResNet-50.

Розробка модуля генерації пояснень. Ключова наукова складова роботи – створення модуля, який надає текстові пояснення результатів класифікації. Розроблено гібридний підхід, що поєднує детекцію ключових точок обличчя (MediaPipe), інтерпретацію мімічних ознак на основі правил FACS (одиниці дій) та шаблонну генерацію тексту. Оцінка модуля за участю користувачів підтвердила його ефективність. Автоматична оцінка за допомогою BERTScore (F1=0,86) засвідчила високу семантичну близькість до експертних описів.

Проєктування та розробка вебсайту. Створено інтуїтивно зрозумілий користувацький інтерфейс з урахуванням принципів емоційного дизайну та мікровзаємодій, що підвищують задоволеність користувачів. Розроблено високопродуктивний бекенд на основі FastAPI з єдиним ендпоінтом /predict для обробки зображень. Забезпечено повноцінну інтеграцію моделі класифікації, модуля пояснень та клієнтської частини. Система демонструє стабільну роботу з часом відповіді 300-500 мс на GPU та 1-2 секунди на CPU, що є прийнятним для інтерактивного вебзастосунку.

Аналіз результатів та перспективи. Проведене тестування підтвердило, що розроблений вебсайт успішно вирішує поставлені задачі, вигідно відрізняючись від існуючих аналогів наявністю інтерпретованих пояснень. Водночас виявлено обмеження поточної версії: унімодальний підхід (лише

візуальний аналіз), недостатня точність для окремих емоцій (страх, відраза), відсутність адаптації до індивідуальних особливостей користувачів. Перспективи розвитку системи включають: розширення до мультимодальної платформи з інтеграцією аудіо- та текстових даних ; впровадження адаптивних інтерфейсів, чутливих до емоційного стану користувача; покращення інтерпретованості за допомогою візуальних методів XAI (наприклад, Grad-CAM); розширення сфер застосування в освіті, психічному здоров'ї, маркетингу та клієнтському сервісі; оптимізацію моделей для роботи на мобільних пристроях з локальною обробкою даних для підвищення конфіденційності.

Результати роботи апробовано у вигляді тез доповіді на другій міжнародній науково-практичній конференції «Сучасні нформаційні технології та системи штучного інтелекту MIT&AIS-2026», яка відбулася 27-29 квітня 2026 р.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Darwin C. The Expression of the Emotions in Man and Animals. London: John Murray; (1872) 374 p.

2. Ekman P., Friesen W.V. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press; (1978) 178 p.

3. Picard R.W. Affective Computing. Cambridge MIT; (1997) 292 p.

4. Pantic M., Rothkrantz L.J.M. Automatic Analysis of Facial Expressions: The State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2000) Vol. 22, No. 12. P. 1424-1445.

5. Goodfellow I.J., Erhan D., Carrier P.L. et al. Challenges in Representation Learning: A Report on Three Machine Learning Contests. Neural Networks. (2015) Vol. 64. P. 59-63.

6. Sreehari P., Raghavendra U., Gudigar A. A Review of Deep Learning Techniques for EEG-Based Emotion Recognition: Models, Methods, and Datasets. F1000Research. (2026) Vol. 14, 1276 [cited 2026 March 15]. URL: https://f1000research.com/articles/14-1276/v2 (дата звернення 30.03.2026)

7. Barrett L.F. Solving the Emotion Paradox: Categorization and the Experience of Emotion. Personality and Social Psychology Review. (2006) Vol. 10, No. 1. P. 20-46.

8. Mehrabian A., Russell J.A. An Approach to Environmental Psychology. Cambridge: MIT Press; (1974) 266 p.

9. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE; (2016) P. 770-778.

10. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). (2021) [cited 2026 March 15]. URL: https://openreview.net/forum?id=YicbFdNTTy (дата звернення 30.03.2026)

11. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need. Advances in Neural Information Processing Systems. Long Beach: NIPS; (2017)

12. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge: MIT Press; (2016) 800 p.

13. Black J.T., Shakir M.Z. AI Enabled Facial Emotion Recognition Using Low-Cost Thermal Cameras. Computing & AI Connect. (2025) Vol. 2, 2025.0019 [cited 2026 March 15]. URL: https://research-portal.uws.ac.uk/en/publications/ai-enabled-facial-emotion-recognition-using-low-cost-thermal-came/ (дата звернення 30.03.2026)

14. Zhou Z., et al. Voices, Faces, and Feelings: Multi-modal Emotion-Cognition Captioning for Mental Health Understanding. arXiv preprint. (2026) arXiv:2603.01816 [cited 2026 March 15]. URL: https://arxiv.org/abs/2603.01816 (дата звернення 30.03.2026)

15. Tian W., Zhao Z., Hu J. et al. EmoOmni: Bridging Emotional Understanding and Expression in Omni-Modal LLMs. arXiv preprint. (2026) arXiv:2602.21900 [cited 2026 March 15].URL: https://arxiv.org/html/2602.21900 (дата звернення 30.03.2026)

16. Kumar A., et al. A Generalized Zero-Shot Deep Learning Classifier for Emotion Recognition Using Facial Expression Images. IEEE Access. (2025) Vol. 13. P. 18687-18700.

17. Sreehari P., Raghavendra U., Gudigar A. A Review of Deep Learning Techniques for EEG-Based Emotion Recognition: Models, Methods, and Datasets [version 2; peer review: 2 approved, 1 approved with reservations]. F1000Research. (2026) Vol. 14, 1276 [cited 2026 March 15]. URL: https://f1000research.com/articles/14-1276/v2 (дата звернення 30.03.2026)

18. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE; (2017) P. 1251-1258.

19. Simonyan K., Zisserman A. Very Deep Convolutional Networks for

Large-Scale Image Recognition. International Conference on Learning Representations. San Diego: ICLR; (2015) 14 p.

20. Howard A.G., Zhu M., Chen B. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint. (2017) arXiv:1704.04861 [cited 2026 March 15]. URL: https://arxiv.org/abs/1704.04861 (дата звернення 30.03.2026)

21. Lugaresi C., Tang J., Nash H. et al. MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint. (2019) arXiv:1906.08172 [cited 2026 March 15]. URL: https://arxiv.org/abs/1906.08172 (дата звернення 30.03.2026)

22. Bradski G. The OpenCV Library. Dr. Dobb's Journal of Software Tools. (2000) Vol. 25, No. 11. P. 120-123.

23. Paszke A., Gross S., Massa F. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems. Vancouver: NIPS; (2019) P. 8024-8035.

24. Brownlee J. Deep Learning for Computer Vision. Melbourne: Machine Learning Mastery; (2019) 564 p.

25. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Sebastopol: O'Reilly Media; (2019) 856 p.

26. Raschka S., Mirjalili V. Python Machine Learning. 3rd ed. Birmingham: Packt Publishing; (2019) 770 p.

27. Lutz M. Learning Python. 5th ed. Sebastopol: O'Reilly Media; (2013) 1600 p.

28. McKinney W. Python for Data Analysis. 2nd ed. Sebastopol: O'Reilly Media; (2017) 544 p.

29. VanderPlas J. Python Data Science Handbook. Sebastopol: O'Reilly Media; (2016) 548 p.