Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

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

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Мультимодальне виявлення зображень, згенерованих штучним інтелектом
(тема)

Виконав:
здобувач другого року навчання,
групи СШІм-24-1
Дамір ГУКОВ
(власне ім'я, прізвище)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник проф. Ваган ТЕРЗІЯН
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Лариса ЧАЛА
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

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

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

Освітня програма Системи штучного інтелекту
(повна назва)

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

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Гукову Даміру Руслановичу
(прізвище, ім'я, по батькові)

1. Тема роботи Мультимодальне виявлення зображень, згенерованих штучним інтелектом

затверджена наказом університету від 6 квітня 20 26 р. № 269Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 26 травня 20 26 р.

3. Вихідні дані до роботи офіційна документація до Python, TensorFlow, FastAPI, Go, Docker, EfficientNetB0, OpenCV, NumPy, Pillow.

4. Перелік питань, що потрібно опрацювати в роботі

1) Аналіз предметної галузі та постановка задачі

2) Розробка системи мультимодального виявлення зображень, згенерованих штучним інтелектом

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	06.04.2026	виконано
2	Визначення проблеми	20.04.2026	виконано
3	Аналіз існуючих аналогів	28.04.2026	виконано
4	Розробка компонентів системи	15.05.2026	виконано
5	Оформлення пояснювальної записки	20.05.2026	виконано
6	Захист перед ЕК	26.05.2026	

Дата видачі завдання 6 квітня 20 26 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Ваган ТЕРЗІЯН
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 70 с., 8 рис., 3 табл., 4 дод., 15 джерел.

ДИФУЗІЙНІ МОДЕЛІ, КЛАСИФІКАЦІЯ ЗОБРАЖЕНЬ, КОМП'ЮТЕРНИЙ ЗІР, МУЛЬТИМОДАЛЬНИЙ АНАЛІЗ, НЕЙРОННІ МЕРЕЖІ, ШТУЧНИЙ ІНТЕЛЕКТ, AI-GENERATED ЗОБРАЖЕННЯ, DEEPFAKE.

Об'єкт дослідження – зображення, які створені за допомогою технологій штучного інтелекту, а також варіанти їх автоматичного виявлення в інформаційному полі.

Предмет дослідження – методи та підходи до аналізу зображень, які дозволяють відрізняти реальні зображення від згенерованих, зокрема з використанням нейронних мереж та мультимодального аналізу.

Мета роботи – дослідження сучасних підходів до виявлення AI-згенерованих зображень і розробка рішення, яке дозволяє підвищити точність їх розпізнавання, особливо у складних випадках, коли такі зображення максимально наближені до реальних.

Методи дослідження – аналіз наукових джерел і сучасних технологій генерації зображень, порівняння існуючих підходів до їх виявлення, а також дослідження можливостей використання моделей глибокого навчання для задачі класифікації зображень.

Результатом дослідження є визначення підходу до побудови системи виявлення AI-згенерованих зображень, яка враховує особливості сучасних генеративних моделей і дозволяє ефективно працювати навіть у складних умовах (hard test), де різниця між реальними та згенерованими даними є мінімальною.

ABSTRACT

Master’s thesis contains: 70 pp., 8 fig., 3 tabl., 4 ann., 15 references.

AI-GENERATED IMAGES, ARTIFICIAL INTELLIGENCE, COMPUTER VISION, DEEPFAKE, DIFFUSION MODELS, IMAGE CLASSIFICATION, MULTIMODAL ANALYSIS, NEURAL NETWORKS.

The object of research is images generated using artificial intelligence technologies, as well as approaches to their automatic detection in the information space.

The subject of research is methods and approaches to image analysis that make it possible to distinguish real images from generated ones, particularly through the use of neural networks and multimodal analysis.

The purpose of the work is to study modern approaches to the detection of AI-generated images and to develop a solution that improves the accuracy of their recognition, especially in complex cases where such images are максимально close to real ones.

Research methods include the analysis of scientific sources and modern image generation technologies, comparison of existing approaches to image detection, as well as the study of deep learning model capabilities for image classification tasks.

The result of the research is the identification of an approach to building a system for detecting AI-generated images that takes into account the features of modern generative models and enables effective operation even under challenging conditions (hard test), where the difference between real and generated data is minimal.

ЗМІСТ

Вступ 8
1 Аналіз предметної галузі та постановка задачі 10
1.1 Опис предметної галузі 10
1.2 Актуальність дослідження 12
1.3 Методи виявлення AI-згенерованих зображень 14
1.3.1 Методи на основі нейронних мереж 14
1.3.2 Методи на основі артефактів 16
1.3.3 Методи частотного аналізу 18
1.3.4 Мультимодальні підходи 20
1.4 Методи пояснюваного штучного інтелекту в задачах детекції 22
1.4.1 Концепція пояснюваного AI (XAI) 22
1.4.2 Метод Grad-CAM 24
1.5 Датасети для виявлення AI-згенерованих зображень 25
1.5.1 Вимоги до датасетів 25
1.5.2 Порівняння існуючих датасетів 27
1.6 Метрики оцінювання якості моделей 27
1.7 Огляд існуючих комерційних рішень 29
1.8 Постановка задачі 31
1.8.1 Мета та завдання роботи 31
1.8.2 Вимоги до системи 32
2 Розробка системи мультимодального виявлення зображень, згенерованих штучним інтелектом 33
2.1 Вибір технологій для реалізації системи 33
2.1.1 Вибір мови програмування для ML-сервісу 33
2.1.2 Вибір мови програмування для веб-інтерфейсу 34
2.1.3 Вибір інструменту для розгортання 34
2.2 Архітектура системи 35
2.3 Реалізація нейромережевого класифікатора 37

2.3.1 Підготовка та організація датасету 37
2.3.2 Архітектура нейронної мережі 38
2.3.3 Процес навчання 38
2.3.4 Результати навчання 40
2.4 Реалізація каналів аналізу 41
2.4.1 Канал частотного аналізу (FFT) 41
2.4.2 Канал аналізу метаданих (EXIF) 42
2.4.3 Зважене об'єднання каналів та формування рішення 44
2.5 Реалізація механізму Grad-CAM 44
2.6 Розробка API та веб-інтерфейсу 47
2.6.1 REST API Python-сервісу 47
2.6.2 Веб-інтерфейс Go-сервісу 48
2.7 Тестування системи 50
2.7.1 Функціональне тестування 50
2.7.2 Тестування продуктивності 51
2.7.3 Тестування розгортання 51
Висновки 53
Перелік джерел посилання 54
Додаток А Лістинг програмного коду Python-сервісу (main.py) 56
Додаток Б Лістинг програмного коду навчання моделі (train.py) 62
Додаток В Лістинг програмного коду Go-сервісу (main.go) 66
Додаток Г Відомість кваліфікаційної роботи 70

ВСТУП

У сучасному світі технології штучного інтелекту розвиваються надзвичайно швидкими темпами та активно інтегруються у різні сфери людської діяльності. Одним із найбільш динамічних напрямків є генерація цифрових зображень за допомогою нейронних мереж. Сучасні генеративні моделі, зокрема генеративно-змагальні мережі (GAN) та дифузійні моделі, здатні створювати зображення настільки високої якості, що їх дедалі складніше відрізнити від справжніх фотографій навіть людині. Завдяки стрімкому розвитку алгоритмів генерації синтетичний контент стає більш деталізованим, реалістичним та візуально переконливим.

Такі системи, як Stable Diffusion, Midjourney та DALL·E, стали загальнодоступними та дозволяють будь-якому користувачу без спеціальних технічних знань створювати фотореалістичні зображення за лічені секунди. Подібні технології відкривають широкі можливості для дизайну, реклами, кіновиробництва, комп’ютерних ігор та цифрового мистецтва. Водночас масове поширення AI-зображень створює низку серйозних проблем, пов’язаних із безпекою та достовірністю інформації. Синтетичні зображення можуть використовуватись для створення фейкових новин, маніпуляцій у соціальних мережах, підробки документів, фінансового шахрайства та інших форм дезінформації. Особливу небезпеку становить використання deepfake-технологій, які дозволяють генерувати підроблені фотографії та медіаматеріали з високим рівнем правдоподібності.

У зв’язку з цим задача автоматичного виявлення AI-згенерованих зображень набуває особливої актуальності. Однак більшість сучасних систем детекції мають низку обмежень. Частина рішень орієнтована лише на конкретні генеративні моделі або окремі типи артефактів, що суттєво знижує ефективність при появі нових алгоритмів генерації. Крім того, багато моделей працюють за принципом «чорної скриньки», коли

користувач отримує лише результат класифікації без пояснення причин прийнятого рішення. Це ускладнює використання подібних систем у сферах, де важливими є прозорість та обґрунтованість результатів, зокрема у журналістиці, цифровій криміналістиці, інформаційній безпеці та юридично значущих процесах.

Додатковою складністю є постійне вдосконалення генеративних моделей. Нові версії дифузійних систем здатні усувати характерні артефакти, які раніше використовувались для виявлення синтетичних зображень. У результаті класичні підходи, засновані лише на аналізі окремих візуальних ознак, поступово втрачають ефективність. Це створює потребу у розробці більш комплексних та стійких методів аналізу, здатних враховувати не лише просторові особливості зображення, але й приховані закономірності у частотній області, структурі метаданих та інших супровідних характеристиках.

Саме тому розробка інтелектуальної системи виявлення AI-згенерованих зображень на основі мультимодального аналізу є актуальною науково-практичною задачею. Основний акцент у роботі зроблено на поєднанні декількох незалежних каналів аналізу: нейромережевого аналізу візуального вмісту, частотного аналізу зображення та аналізу метаданих файлів. Додатково система використовує механізми пояснюваного штучного інтелекту, зокрема теплові карти Grad-CAM, що дозволяють візуалізувати області зображення, які найбільше вплинули на рішення моделі. Використання такого підходу дає змогу підвищити точність класифікації, покращити стійкість до різних типів AI-контенту та забезпечити більшу прозорість роботи системи.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Опис предметної галузі

Відрізнення AI-згенерованих зображень від реальних фотографій на сьогодні є однією з актуальних задач у сфері комп'ютерного зору та інформаційної безпеки. Стрімкий розвиток генеративних моделей призвів до того, що синтетичні зображення стали надзвичайно реалістичними та візуально майже не відрізняються від справжніх фотографій. У зв'язку з цим виникає потреба у створенні автоматизованих систем, здатних визначати походження зображення та виявляти ознаки його штучної генерації.

Одним із базових підходів до розв'язання цієї задачі є бінарна класифікація, у межах якої система отримує зображення на вхід, аналізує його та визначає, чи є воно реальним, чи згенерованим штучним інтелектом. Незважаючи на відносну простоту постановки задачі, її практична реалізація супроводжується рядом суттєвих труднощів. Основна проблема полягає у постійному вдосконаленні генеративних моделей, через що межа між справжніми та синтетичними зображеннями стає дедалі менш помітною.

Ще кілька років тому більшість AI-зображень містили помітні візуальні дефекти, які людина могла легко помітити без використання спеціалізованих інструментів. Серед типових проблем були неприродна анатомія, помилки у деталях обличчя, спотворення фону або некоректне освітлення. Проте сучасні генеративні системи навчилися відтворювати природне зерно сенсора, коректний розподіл світла, складні текстури та дрібні деталі з дуже високим рівнем достовірності. Через це людського аналізу вже часто недостатньо для надійного виявлення синтетичних зображень, що робить використання методів машинного навчання та математичного аналізу необхідним.

На сьогодні найбільш поширеними технологіями генерації зображень є генеративно-змагальні мережі (GAN) та дифузійні моделі. Генеративно-змагальна мережа складається з двох нейронних мереж – генератора та дискримінатора. Генератор створює нові зображення на основі випадкового шуму, тоді як дискримінатор оцінює, наскільки отримане зображення схоже на реальне. Обидві мережі навчаються одночасно у процесі так званої міні-макс гри: генератор намагається створити максимально правдоподібне зображення, а дискримінатор – правильно відрізнити синтетичні дані від реальних. Завдяки такому конкурентному механізму якість генерації постійно покращується. Одними з найвідоміших GAN-архітектур є StyleGAN2 та StyleGAN3, які здатні створювати високореалістичні портретні зображення.

Дифузійні моделі працюють за іншим принципом. Їхня робота базується на поступовому додаванні шуму до зображення та подальшому відновленні початкової структури під час зворотного процесу. На етапі навчання до реального зображення покроково додається гаусівський шум до майже повного руйнування інформації. Після цього нейронна мережа навчається відновлювати зображення із зашумленого стану. Одним із найбільш відомих підходів цього типу є Latent Diffusion Models, що лежать в основі Stable Diffusion. Використання прихованого стисненого простору дозволяє суттєво зменшити обчислювальні витрати та одночасно забезпечити високу якість генерації.

Попри високу реалістичність синтетичних зображень, генеративні моделі все ще залишають певні характерні артефакти, які можуть використовуватись для їх автоматичного виявлення. Для GAN-моделей типовими є аномалії у частотному спектрі, порушення локальної симетрії, неприродне відтворення текстур або помилки у дрібних деталях. Дифузійні моделі, у свою чергу, можуть формувати надмірно гладкі поверхні, неприродний розподіл шуму або характерні закономірності у високочастотних компонентах зображення.

Додатковим джерелом інформації для виявлення синтетичного контенту є аналіз метаданих файлів. Справжні фотографії зазвичай містять EXIF-метадані, зокрема модель камери, параметри зйомки, дату створення або GPS-координати. AI-згенеровані зображення часто не містять таких даних або мають нетипову структуру метаданих. Аналіз цієї інформації дозволяє підвищити ефективність системи детекції та використовувати не лише візуальні, але й супровідні ознаки походження зображення.

1.2 Актуальність дослідження

У 2022 році поява загальнодоступних генераторів зображень, зокрема Stable Diffusion, Midjourney та DALL·E, суттєво змінила підхід до створення цифрового контенту. Якщо раніше генерація фотореалістичних зображень вимагала професійних навичок, спеціалізованого програмного забезпечення та значних часових витрат, то сучасні AI-системи дозволяють створювати високоякісні зображення за текстовим описом усього за кілька секунд. Завдяки простоті використання та широкій доступності такі технології швидко набули популярності серед звичайних користувачів, дизайнерів, маркетологів та творців цифрового контенту.

Масове поширення генеративних моделей призвело до стрімкого збільшення кількості синтетичних зображень в інформаційному просторі. За оцінками різних аналітичних досліджень, частка контенту із використанням AI-генерації або AI-обробки у соціальних мережах постійно зростає. Сучасні генератори вже здатні створювати не лише художні ілюстрації, але й реалістичні портрети, новинні фотографії, рекламні матеріали та інші типи візуального контенту, які візуально майже не відрізняються від справжніх фотографій.

Особливе занепокоєння викликає поширення deepfake-технологій та інших форм синтетичного медіаконтенту. За даними компанії Sensity AI, обсяг deepfake-матеріалів щороку демонструє стрімке зростання. Наслідки

цього процесу вже помітні у різних сферах суспільства. Синтетичні зображення використовуються у політичних маніпуляціях, інформаційних кампаніях, фінансовому шахрайстві та поширенні дезінформації. Підроблені фотографії можуть використовуватись як фальшиві докази, а реалістичні AI-зображення дедалі частіше застосовуються для створення фейкових профілів, новин або компрометуючих матеріалів. У зв’язку з цим у багатьох країнах активно обговорюється необхідність впровадження нормативних вимог щодо маркування AI-контенту та розробки механізмів перевірки його автентичності.

Попри активний розвиток систем детекції AI-зображень, сучасні підходи все ще мають низку суттєвих обмежень. Однією з основних проблем є вузька спеціалізація багатьох детекторів. Наприклад, модель, навчена переважно на даних, створених GAN-архітектурами, може демонструвати значне зниження точності при роботі з дифузійними моделями. Це пояснюється тим, що різні генеративні підходи залишають різні типи артефактів та статистичних закономірностей у зображенні.

Додатковою проблемою є швидка втрата актуальності детекторів після появи нових генеративних моделей. Сучасні AI-генератори постійно вдосконалюються, усуваючи характерні дефекти та аномалії, які раніше використовувались для автоматичного виявлення синтетичних зображень. У результаті ефективність існуючих детекторів може суттєво знижуватись уже через короткий проміжок часу після появи нових моделей генерації.

Ще одним важливим недоліком більшості сучасних систем є відсутність пояснення результатів класифікації. Більшість детекторів надають лише бінарну відповідь про те, чи є зображення AI-згенерованим, без пояснення того, які саме ознаки вплинули на прийняття рішення. Такий підхід ускладнює незалежну перевірку результатів та знижує рівень довіри до системи, особливо у сферах, де важливими є прозорість і обґрунтованість висновків.

Саме тому мультимодальний підхід із використанням елементів пояснюваного штучного інтелекту, реалізований у даній роботі, орієнтований на подолання зазначених недоліків. Поєднання нейромережевого аналізу, дослідження частотних характеристик зображення та аналізу метаданих дозволяє створити більш універсальну та стійку систему детекції AI-контенту. Додаткове використання методів Explainable AI, зокрема теплових карт Grad-CAM, забезпечує візуалізацію областей зображення, які найбільше вплинули на рішення моделі, що підвищує прозорість та інтерпретованість роботи системи.

1.3 Методи виявлення AI-згенерованих зображень

Існують чотири основні групи методів виявлення AI-зображень: методи на основі нейронних мереж, методи аналізу артефактів, методи частотного аналізу та мультимодальні підходи. Кожен спосіб має свій власний тип ознак і має свої переваги та недоліки. Тому, щоб отримати найліпший результат використовують мультимодальні системи.

1.3.1 Методи на основі нейронних мереж

При використанні нейронних мереж система здатна самостійно навчатися знаходити ознаки синтетичного контенту без необхідності ручного визначення правил або шаблонів. На відміну від класичних алгоритмів аналізу зображень, де ознаки задаються людиною, моделі глибокого навчання автоматично формують внутрішнє представлення даних на основі великої кількості прикладів. Це дозволяє виявляти складні закономірності та приховані особливості, які можуть бути непомітними для людини. Саме тому нейронні мережі стали одним із найбільш ефективних інструментів для задачі виявлення AI-згенерованих зображень.

Одним із найпоширеніших типів моделей для аналізу зображень є згорткові нейронні мережі (CNN). Їхня основна перевага полягає у здатності виділяти ієрархічні ознаки зображення. На початкових шарах модель навчається розпізнавати прості елементи, такі як краї, лінії або кути. На глибших рівнях формуються складніші ознаки – текстури, форми об’єктів, локальні патерни та просторові структури. Завдяки такому підходу CNN можуть знаходити приховані відмінності між реальними та AI-згенерованими зображеннями, навіть якщо вони майже непомітні візуально.

Однією з найбільш ефективних сучасних CNN-архітектур є EfficientNet, запропонована Tan та Le у 2019 році. Архітектура базується на принципі compound scaling, який передбачає одночасне масштабування глибини мережі, її ширини та роздільної здатності вхідного зображення. Такий підхід дозволяє досягати високої точності при відносно невеликих обчислювальних витратах. Базова модель EfficientNetB0 містить приблизно 4,2 мільйона параметрів та демонструє високу ефективність на датасеті ImageNet при розмірі вхідного зображення 224×224 пікселі.

Архітектура EfficientNet використовує мобільні інвертовані залишкові блоки MBConv у поєднанні з модулями squeeze-and-excitation. Такі модулі дозволяють мережі адаптивно визначати важливість різних каналів ознак та концентрувати увагу на найбільш інформативних характеристиках зображення. Це особливо корисно у задачах виявлення AI-контенту, де ключові відмінності можуть бути локальними та слабко вираженими.

Для ефективного навчання моделей при обмеженому обсязі даних широко використовується підхід Transfer Learning. Його суть полягає у використанні моделей, попередньо навчених на великих датасетах, зокрема ImageNet. Такі моделі вже містять базові знання про структуру та текстурні особливості зображень, тому можуть бути адаптовані до нової задачі значно швидше, ніж моделі, які навчаються з нуля.

Зазвичай процес навчання складається з двох етапів. На першому етапі базова частина моделі заморожується, а навчанню піддається лише фінальний класифікаційний шар. Це дозволяє швидко адаптувати модель до нової задачі без зміни вже сформованих базових ознак. На другому етапі виконується fine-tuning верхніх шарів мережі зі зменшеною швидкістю навчання. Такий підхід дозволяє більш точно налаштувати модель під специфіку задачі виявлення AI-зображень та суттєво скоротити час навчання у порівнянні з повним навчанням моделі з нуля.

Окрім CNN-архітектур, останніми роками активно розвиваються підходи на основі трансформерів. Однією з найвідоміших моделей цього типу є Vision Transformer. Vision Transformer адаптує трансформерну архітектуру до задач аналізу зображень шляхом розбиття зображення на окремі патчі, які обробляються подібно до токенів у задачах обробки тексту. Після цього між усіма частинами зображення встановлюються взаємозв’язки за допомогою механізму самоуваги.

На відміну від CNN, які переважно концентруються на локальних особливостях, Vision Transformer здатний враховувати глобальні залежності між різними областями зображення. Це є особливо важливим у задачах виявлення AI-зображень, де ознаки синтетичності можуть проявлятися не лише у дрібних деталях, але й у загальній логіці сцени, структурі освітлення або просторових взаємозв’язках між об’єктами. Водночас основним недоліком ViT є висока потреба у великих обсягах навчальних даних та значних обчислювальних ресурсах, що ускладнює використання таких моделей у задачах із обмеженим набором даних.

1.3.2 Методи на основі артефактів

Переважна більшість генеративних моделей у процесі синтезу зображень створює характерні дефекти, відомі як артефакти генерації. Незважаючи на високий рівень реалістичності сучасних AI-зображень,

процес їх побудови все ще супроводжується статистичними, структурними або візуальними аномаліями, які можуть використовуватись для автоматичного виявлення синтетичного контенту. Тип артефактів значною мірою залежить від архітектури генеративної моделі та особливостей процесу генерації.

Для генеративно-змагальних мереж (GAN) характерними є проблеми з відтворенням складних дрібних деталей. Найчастіше такі моделі допускають помилки при генерації людських облич, зокрема очей, зубів, волосся або пальців рук. Також можуть виникати невідповідності між переднім планом та фоном, порушення геометрії об’єктів або неприродні текстурні переходи. Однією з типових проблем GAN-архітектур є поява так званих шахових візерунків (checkerboard artifacts), які виникають через використання транспонованих згорток під час збільшення роздільної здатності зображення. Такі дефекти можуть проявлятись у вигляді регулярних повторюваних структур або неприродних патернів у різних областях зображення.

Дифузійні моделі мають інший тип характерних артефактів. Їхня генерація часто супроводжується надмірно гладкими поверхнями та неприродним розподілом дрібних текстур, через що зображення інколи набувають ефекту «гуаші» або цифрового розмиття. Крім цього, дифузійні моделі можуть створювати надто чіткі контури окремих об’єктів, що виглядає неприродно порівняно з реальними фотографіями. Однією з найбільш помітних проблем залишається генерація тексту: символи можуть бути спотвореними, частково зміненими або взагалі нечитабельними. Це пояснюється тим, що генеративні моделі відтворюють текст як візуальний шаблон, а не як послідовність логічно пов’язаних символів.

Виявлення артефактів може виконуватись як вручну експертами, так і автоматизовано за допомогою алгоритмів комп’ютерного зору та машинного навчання. Аналіз таких ознак дозволяє знаходити характерні закономірності, які залишають генеративні моделі під час синтезу

зображень. Основною перевагою підходу, заснованого на аналізі артефактів, є його відносна прозорість та інтерпретованість. На відміну від повністю «чорних» моделей класифікації, у цьому випадку можна пояснити, які саме аномалії стали причиною визнання зображення синтетичним.

Водночас даний підхід має низку суттєвих обмежень. Більшість правил та методів аналізу артефактів прив’язані до конкретного типу генератора або певної генеративної архітектури. Після появи нових моделей генерації багато характерних дефектів поступово зникають або змінюються, через що ефективність раніше сформованих правил суттєво знижується. Саме тому аналіз артефактів рідко використовується як самостійний метод детекції AI-зображень. Найчастіше його застосовують у поєднанні з нейромережевими підходами, що дозволяє об’єднати інтерпретованість аналізу артефактів із високою узагальнювальною здатністю моделей глибокого навчання.

1.3.3 Методи частотного аналізу

Частотний аналіз пропонує інший підхід до дослідження зображень порівняно з традиційними методами комп’ютерного зору. Якщо більшість класичних алгоритмів працює безпосередньо зі значеннями пікселів та просторовими ознаками, то частотний аналіз зосереджується на вивченні розподілу енергії між різними частотними компонентами зображення. Такий підхід дозволяє виявляти приховані закономірності та аномалії, які можуть бути непомітними у звичайному візуальному представленні. Саме тому аналіз частотного спектра активно використовується у задачах виявлення AI-згенерованих зображень.

Основою цього підходу є двовимірне дискретне перетворення Фур’є (2D DFT), яке дозволяє представити зображення у частотній області. У результаті такого перетворення зображення розкладається на набір гармонічних компонентів різної частоти. Низькочастотні компоненти

відповідають великим однорідним областям, плавним переходам кольору та загальній структурі сцени, тоді як високочастотні компоненти описують дрібні деталі, текстури, контури та різкі зміни яскравості. Аналіз співвідношення між цими компонентами дозволяє оцінити характер формування зображення та виявити нетипові закономірності.

Для фотографій, отриманих реальною камерою, характерний поступовий спад енергії від низьких частот до високих. Такий розподіл часто описують як 1/f-спектр, де інтенсивність високочастотних компонентів природно зменшується. Подібна закономірність формується під впливом фізичних особливостей оптики, шуму матриці сенсора, процесів компресії та природної структури реальних сцен. У результаті спектр справжніх фотографій має більш хаотичний та природний вигляд без чітко виражених регулярних патернів.

AI-згенеровані зображення часто демонструють інший характер розподілу частотних компонентів. Для GAN-моделей типовими є характерні піки у частотному спектрі, які виникають через використання транспонованих згорток під час генерації зображення. Такі аномалії пов’язані з кроком згортки та можуть проявлятися у вигляді регулярних повторюваних структур у спектральному представленні. Навіть якщо подібні дефекти практично непомітні візуально, у частотній області вони залишають добре помітні закономірності.

Дифузійні моделі формують інший тип спектральних аномалій. Через особливості процесу поступового додавання та усунення шуму їхні зображення часто мають надмірно рівномірний розподіл енергії у середньочастотному діапазоні. Це призводить до появи нетипової гладкості текстур та змінює природний баланс між різними частотними компонентами. Крім того, процес денойзингу може формувати специфічні патерни шуму, які відрізняються від шумових характеристик реальних фотографій.

Головною перевагою частотного аналізу є його здатність виявляти приховані ознаки синтетичності, які залишаються невидимими для людського ока. Навіть якщо AI-зображення виглядає абсолютно реалістично у просторовій області, у частотному спектрі можуть зберігатися статистичні аномалії, пов’язані з процесом генерації. Саме тому частотний аналіз часто використовується як додатковий канал у мультимодальних системах виявлення AI-контенту, де він доповнює нейромережевий аналіз та підвищує загальну стійкість системи до різних типів генеративних моделей.

1.3.4 Мультимодальні підходи

Головна ідея мультимодального підходу полягає у поєднанні декількох незалежних каналів аналізу в межах єдиної системи прийняття рішень. На відміну від однокомпонентних моделей, які спираються лише на один тип ознак, мультимодальні системи одночасно використовують декілька джерел інформації. Це дозволяє аналізувати зображення з різних точок зору та формувати більш надійний і стійкий результат класифікації.

Ключовою перевагою такого підходу є можливість створення своєрідного «синергічного ефекту», коли окремі компоненти системи взаємно доповнюють один одного та компенсують слабкі сторони інших каналів аналізу. Наприклад, якщо один метод не зміг виявити ознаки синтетичності через відсутність характерних артефактів, інший канал може знайти аномалії у частотному спектрі або структурі метаданих. Завдяки цьому загальна стійкість системи до різних типів AI-зображень суттєво підвищується.

У сучасних системах виявлення AI-контенту можна виділити декілька основних варіантів реалізації мультимодального аналізу. Перший підхід базується на поєднанні візуальної інформації із текстовим описом зображення. Одним із найвідоміших прикладів є модель CLIP від OpenAI,

яка навчається одночасно аналізувати зображення та текстові описи. Такий підхід дозволяє оцінювати семантичну узгодженість між візуальним вмістом і текстом, що може бути корисним для виявлення нелогічних або штучно сформованих сцен.

Другий підхід передбачає об’єднання декількох типів ознак самого зображення. У цьому випадку система одночасно аналізує просторові ознаки пікселів, частотні характеристики, статистичні закономірності, артефакти генерації та метадані файлу. Наприклад, нейромережевий модуль може оцінювати текстурні особливості зображення, частотний аналіз – виявляти аномалії у спектрі, а модуль аналізу метаданих – перевіряти наявність EXIF-інформації або нетипових параметрів файлу.

Окремим напрямком є ансамблеві або комбіновані методи, у яких декілька незалежних класифікаторів формують спільне рішення шляхом голосування або зваженого об’єднання результатів. У таких системах кожен класифікатор спеціалізується на певному типі ознак, а фінальне рішення приймається на основі сукупної оцінки всіх компонентів. Подібний підхід дозволяє зменшити вплив помилок окремих моделей та підвищити узагальнювальну здатність системи.

Фінальна оцінка у мультимодальних системах зазвичай формується як зважена комбінація результатів окремих каналів аналізу. Вагові коефіцієнти визначаються залежно від точності, стабільності та надійності кожного модуля. Наприклад, якщо нейромережевий аналіз демонструє найкращу ефективність, його внесок у фінальне рішення може бути більшим порівняно з іншими каналами. Водночас додаткові модулі, такі як аналіз частотного спектра або метаданих, виконують роль допоміжних механізмів перевірки.

Порівняно з однокомпонентними системами мультимодальний підхід є більш стійким до появи нових генеративних моделей та змін у технологіях синтезу зображень. Навіть якщо один із каналів аналізу втрачає ефективність через появу нових типів генераторів, інші модулі можуть

продовжувати успішно виявляти ознаки синтетичності. Крім того, мультимодальні системи забезпечують кращу інтерпретованість результатів, оскільки дозволяють окремо оцінити внесок кожного каналу аналізу у фінальне рішення.

1.4 Методи пояснюваного штучного інтелекту в задачах детекції

1.4.1 Концепція пояснюваного AI (XAI)

Пояснюваний штучний інтелект (XAI, Explainable Artificial Intelligence) є окремим напрямком досліджень у сфері машинного навчання, спрямованим на створення методів, які дозволяють зрозуміти, інтерпретувати та пояснювати рішення моделей штучного інтелекту. На відміну від класичних моделей типу «чорної скриньки», де користувач отримує лише фінальний результат без пояснення причин його формування, XAI-системи надають додаткову інформацію про те, які ознаки або області даних найбільше вплинули на прийняте рішення.

Проблема інтерпретованості є особливо актуальною для сучасних моделей глибокого навчання, оскільки їхня внутрішня структура часто містить мільйони параметрів і складні нелінійні залежності. Через це навіть високоточна модель може викликати недовіру, якщо неможливо пояснити логіку її роботи. У задачах виявлення AI-зображень пояснюваність має особливе значення, оскільки аналітик або користувач повинен не лише отримати результат класифікації, але й зрозуміти, які саме ознаки вказують на синтетичне походження зображення. Це особливо важливо у сферах інформаційної безпеки, журналістики, цифрової криміналістики та юридично значущих процесів, де необхідна обґрунтованість прийнятих рішень.

У сучасних XAI-системах можна виділити два основних підходи: intrinsic та post-hoc методи. Intrinsic-підходи передбачають використання

моделей, які є інтерпретованими за своєю природою. До таких моделей належать дерева рішень, лінійні моделі або інші алгоритми, де логіку прийняття рішень можна безпосередньо простежити. Однак подібні підходи часто поступаються моделям глибокого навчання за точністю у складних задачах аналізу зображень.

Post-hoc методи застосовуються вже після навчання моделі та дозволяють пояснювати її рішення без зміни внутрішньої структури. Саме цей підхід є найбільш поширеним у задачах комп'ютерного зору та аналізу AI-зображень. Post-hoc методи дають змогу досліджувати поведінку навіть дуже складних нейронних мереж та визначати, які області зображення або ознаки стали ключовими для класифікації.

У задачах аналізу зображень особливо популярними є методи, засновані на градієнтах та оцінці важливості окремих областей зображення. Одним із найбільш відомих підходів є Grad-CAM (Gradient-weighted Class Activation Mapping). Даний метод використовує градієнти останніх згорткових шарів нейронної мережі для побудови теплової карти, яка показує області зображення, що найбільше вплинули на рішення моделі.

Окрім Grad-CAM, широко використовуються й інші методи пояснюваного штучного інтелекту. Метод LIME (Local Interpretable Model-agnostic Explanations) створює локальні інтерпретовані моделі навколо конкретного прикладу та дозволяє оцінити внесок окремих частин зображення у фінальний результат. Метод SHAP (SHapley Additive exPlanations) базується на концепції значень Шеплі з теорії ігор та дозволяє кількісно оцінити важливість окремих ознак для рішення моделі. Також поширеними є карти значущості (Saliency Maps), які візуалізують області зображення, найбільш чутливі до змін з точки зору моделі.

Використання XAI у системах виявлення AI-зображень дозволяє не лише підвищити рівень довіри до результатів класифікації, але й забезпечити додатковий аналіз поведінки моделі. Пояснюваність допомагає виявляти випадки, коли модель приймає рішення на основі неправильних

або випадкових ознак, а також дозволяє краще зрозуміти сильні та слабкі сторони системи. Саме тому сучасні мультимодальні системи аналізу AI-контенту дедалі частіше поєднують високоточні нейромережеві моделі з методами пояснюваного штучного інтелекту.

1.4.2 Метод Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) – це один із найбільш поширених методів пояснюваного штучного інтелекту для задач комп'ютерного зору, запропонований Selvaraju та співавторами у 2017 році. Метод дозволяє візуалізувати ті ділянки зображення, які найбільше вплинули на рішення згорткової нейронної мережі під час класифікації. Це особливо важливо у задачах виявлення AI-зображень, де необхідно не лише визначити синтетичне походження зображення, але й пояснити причину такого рішення.

Робота методу складається з декількох етапів. Спочатку зображення проходить через нейронну мережу у прямому напрямку, після чого обчислюються градієнти цільового класу відносно активацій останнього згорткового шару. Далі карти активацій зважуються відповідно до важливості ознак, а до результату застосовується функція ReLU для виділення лише позитивних впливів на прогноз моделі.

У результаті формується теплова карта, яка масштабується до розмірів початкового зображення та накладається поверх нього. Області з найбільшим впливом на рішення моделі позначаються теплими кольорами – червоним або помаранчевим, тоді як менш значущі ділянки відображаються холодними кольорами. У задачах виявлення AI-зображень Grad-CAM дозволяє локалізувати характерні ознаки синтетичності, такі як неприродні текстури, артефакти генерації або аномальні переходи між об'єктами.

Однією з головних переваг Grad-CAM є висока обчислювальна ефективність. На відміну від таких методів, як LIME або SHAP, які

потребують великої кількості повторних проходів через модель, Grad-CAM виконується лише за один прямий та один зворотний прохід мережею. Завдяки цьому метод може використовуватись навіть у системах, що працюють у режимі реального часу.

1.5 Датасети для виявлення AI-згенерованих зображень

1.5.1 Вимоги до датасетів

Різноманітність та структура датасету є одними з ключових факторів, що визначають ефективність системи виявлення AI-зображень. Якість навчального набору безпосередньо впливає на здатність моделі узагальнювати отримані знання та коректно працювати з новими даними, які не були представлені під час навчання. Для забезпечення ефективного навчання датасет повинен відповідати ряду вимог: балансованість класів, різноманітність типів зображень, наявність даних, створених різними генеративними моделями, а також правильний поділ на навчальну, валідаційну та тестову вибірки.

Балансованість класів є важливою для уникнення ситуації, коли модель починає надавати перевагу одному з класів через кількісну перевагу прикладів. Якщо реальних або AI-згенерованих зображень суттєво більше, модель може демонструвати високу формальну точність, але при цьому погано розпізнавати менш представлений клас. Саме тому співвідношення між класами повинно бути максимально рівномірним.

Різноманітність даних також є критично важливою. Датасет повинен містити різні категорії зображень: портрети, пейзажі, об’єкти, сцени з великою кількістю деталей, зображення з різними умовами освітлення та рівнями якості. Окрему увагу необхідно приділяти різноманітності генеративних моделей, оскільки сучасні системи створення AI-зображень використовують різні архітектури та алгоритми генерації. Якщо модель

навчалась лише на даних одного генератора, вона може запам’ятовувати специфічні артефакти конкретної системи замість виявлення загальних ознак синтетичності. У такому випадку точність на знайомих даних може бути високою, але ефективність на нових AI-зображеннях суттєво знизиться.

Не менш важливою є правильна організація вибірок. Навчальна вибірка використовується для безпосереднього навчання моделі, валідаційна – для контролю процесу навчання та підбору параметрів, а тестова – для фінального оцінювання якості системи. При цьому тестові дані не повинні перетинатися з навчальними, оскільки це може призвести до переоцінки реальної ефективності моделі.

Для підвищення різноманітності навчального набору активно застосовується аугментація даних. Вона дозволяє штучно збільшити кількість прикладів та покращити здатність моделі працювати з варіативними зображеннями.

Для задачі виявлення AI-зображень можуть застосовуватись випадкове горизонтальне відзеркалення, поворот у невеликому діапазоні, масштабування, обрізання, зміна яскравості та контрасту. Такі перетворення допомагають моделі стати менш чутливою до незначних змін у зображенні та покращують узагальнюючу здатність.

Водночас аугментація повинна використовуватись обережно. Надмірне застосування перетворень може призвести до втрати або спотворення характерних ознак синтетичності, які модель повинна навчитися виявляти. Наприклад, сильне стиснення, агресивне розмиття або значна зміна кольорової гами можуть приховати артефакти генерації. Саме тому аугментація застосовується лише на етапі навчання та з помірною інтенсивністю, щоб зберегти баланс між різноманітністю даних і збереженням важливих ознак AI-зображень.

1.5.2 Порівняння існуючих датасетів

На сьогодні існує кілька публічних датасетів для задачі виявлення AI-зображень що відрізняються за розміром, різноманітністю генераторів та типом зображень.

Порівняння основних датасетів наведено у таблиці 1.1.

Таблиця 1.1 – Порівняння датасетів для виявлення AI-зображень

Датасет	Розмір	Генератори	Особливості
CIFAKE	120k	Stable Diffusion 1.4	Збалансований бенчмарк, 32×32 пікселі
deepfake_face_classification	25.7k	Midjourney v6, StyleGAN	Реалістичні обличчя, 2024-2025 рік
GenImage	1.3M	MJ, SDXL, DALL-E 2	Різноманітні генератори
ArtiFact	2.4M	26 різних моделей	Найбільший публічний датасет
FaceForensics++	1.8M	DeepFakes, Face2Face	Маніпуляції з обличчями

1.6 Метрики оцінювання якості моделей

Для всебічного оцінювання якості класифікатора застосовується комплекс із чотирьох основних метрик: Accuracy, Precision, Recall та F1-міри. Використання лише одного показника не дозволяє повністю оцінити ефективність системи, особливо у задачах виявлення AI-зображень, де різні

типи помилок можуть мати різний рівень критичності. Саме тому аналіз виконується комплексно, з урахуванням як загальної точності класифікації, так і здатності системи правильно виявляти синтетичні зображення.

Відправною точкою для обчислення метрик є матриця помилок, яка відображає результати класифікації та містить чотири можливі варіанти передбачень. True Positive (TP) відповідає випадкам, коли AI-зображення було коректно визначене як синтетичне. True Negative (TN) означає, що реальне зображення правильно класифіковане як автентичне. False Positive (FP) виникає тоді, коли реальне зображення помилково визначається як AI-згенероване, тоді як False Negative (FN) означає ситуацію, у якій синтетичне зображення система помилково відносить до реальних.

Accuracy характеризує загальну частку правильних передбачень серед усіх проаналізованих зразків. Дана метрика дозволяє оцінити загальну ефективність моделі, однак у задачах із можливим дисбалансом класів її використання окремо може бути недостатнім. Наприклад, якщо один із класів суттєво переважає, модель може демонструвати високе значення Accuracy навіть за наявності значної кількості помилок для менш представленого класу.

Precision показує, яка частка зображень, визначених системою як AI-згенеровані, дійсно є синтетичними. Ця метрика є особливо важливою у випадках, коли критично необхідно мінімізувати кількість хибних спрацьовувань. Високе значення Precision свідчить про те, що система рідко помилково класифікує реальні зображення як AI-згенеровані.

Recall характеризує здатність системи знаходити AI-зображення серед усіх наявних синтетичних прикладів. Іншими словами, ця метрика показує, яку частку AI-зображень вдалося успішно виявити. Для систем безпеки та інструментів автоматичної модерації саме Recall часто є одним із найбільш важливих показників, оскільки пропуск AI-зображення може призвести до поширення дезінформації або маніпулятивного контенту.

F1-міра обчислюється як гармонічне середнє між Precision та Recall і дозволяє отримати збалансовану оцінку якості класифікації. Дана метрика особливо корисна у випадках, коли необхідно одночасно враховувати як хибнопозитивні, так і хибнонегативні результати. Високе значення F1-міри свідчить про те, що система не лише ефективно виявляє AI-зображення, але й робить це з мінімальною кількістю помилкових спрацьовувань. Саме тому F1-міра часто використовується як один із ключових показників загальної ефективності моделей класифікації.

1.7 Огляд існуючих комерційних рішень

Поряд із науковими дослідженнями сьогодні активно розвиваються і комерційні системи виявлення AI-згенерованих зображень. Багато компаній пропонують готові сервіси для автоматичної перевірки контенту, орієнтовані на використання у соціальних мережах, медіаплатформах, системах модерації та інструментах інформаційної безпеки. Подібні рішення зазвичай надаються у вигляді веб-сервісів або API для інтеграції у сторонні застосунки.

Одним із найбільш відомих рішень є Hive Moderation, який надає API для автоматичного виявлення AI-контенту та модерації медіафайлів. Система орієнтована на корпоративне використання та заявляє високі показники точності при аналізі зображень і відео. Іншим популярним сервісом є AI or Not, спеціалізований інструмент для перевірки зображень на AI-генерацію. Сервіс доступний як через веб-інтерфейс, так і через API, а також має обмежений безкоштовний рівень доступу для користувачів.

Окрему увагу привертає сервіс Illuminarty, який пропонує більш розширену функціональність. Окрім визначення факту AI-генерації, система намагається встановити конкретний генератор, що використовувався для створення зображення, а також візуалізує підозрілі області за допомогою теплових карт. Такий підхід частково реалізує

принципи пояснюваного штучного інтелекту та підвищує інтерпретованість результатів аналізу.

У таблиці 1.2 подано порівняння систем виявлення AI-зображень.

Таблиця 1.2 – Порівняння систем виявлення AI-зображень

Характеристика	Дана система	Hive	AI or Not	Illuminarty
Відкритий код	Так	Ні	Ні	Ні
Локальне розгортання	Так	Ні	Ні	Ні
Grad-CAM / XAI	Так	Ні	Ні	Частково
Мультимодальність	Так	Ні	Ні	Ні
Безкоштовно	Так	Ні	Обмежено	Обмежено
Дообчання	Так	Ні	Ні	Ні

Спільним недоліком більшості комерційних рішень є закритість вихідного коду, залежність від зовнішніх сервісів та обмежена інтерпретованість результатів. Користувач зазвичай отримує лише фінальний висновок без детального пояснення того, які саме ознаки або модулі вплинули на рішення системи. Крім того, використання хмарних API створює залежність від сторонньої інфраструктури, що може бути критичним для задач, пов’язаних із конфіденційністю даних або автономною роботою системи. Додатковою проблемою є обмежена можливість адаптації таких сервісів під конкретну предметну галузь або нові типи генеративних моделей. У більшості випадків користувач також не має доступу до навчальних даних та методики оцінювання точності, що ускладнює незалежну перевірку ефективності системи.

На відміну від таких рішень, розроблена в межах кваліфкаційної роботи система є відкритою, розгортається локально та не потребує використання сторонніх сервісів. Система надає окремі результати аналізу для кожного каналу, включаючи нейромережевий, частотний та метадані, а також формує теплову карту Grad-CAM для візуальної інтерпретації рішення моделі.

1.8 Постановка задачі

1.8.1 Мета та завдання роботи

Метою дипломної роботи є створення системи автоматичного мультимодального виявлення AI-згенерованих зображень, доповненої механізмами пояснюваного штучного інтелекту. Розроблювана система має здійснювати бінарну класифікацію вхідного зображення – визначаючи його як автентичне або синтетичне – з наданням розгорнутої оцінки за кожним аналітичним каналом окремо, а також теплової карти Grad-CAM для інтерпретації прийнятого рішення.

Досягнення поставленої мети передбачає вирішення низки взаємопов'язаних задач. Насамперед необхідно провести порівняльний аналіз існуючих підходів до виявлення AI-зображень та обґрунтувати вибір оптимального методу. Паралельно потребує вирішення питання формування та підготовки навчальної вибірки належної якості та збалансованості. На практичному етапі передбачається реалізація нейромережевого класифікатора на базі архітектури EfficientNetB0 із застосуванням Transfer Learning, розробка каналу частотного аналізу на основі перетворення Фур'є, а також реалізація каналу аналізу EXIF-метаданих. Окремими задачами визначено впровадження механізму візуалізації Grad-CAM, розробку REST API та веб-інтерфейсу користувача.

Завершальним етапом є комплексне оцінювання якості системи із застосуванням стандартних метрик класифікації.

1.8.2 Вимоги до системи

Функціональні вимоги до системи охоплюють підтримку графічних форматів JPEG, PNG, WEBP та BMP, повернення бінарного рішення з відповідним показником впевненості, надання окремих оцінок за трьома аналітичними каналами, відображення теплової карти Grad-CAM для інтерпретації результату, а також підтримку взаємодії через REST API.

Серед нефункціональних вимог визначено такі: досягнення точності класифікації не нижче 85% на тестовій вибірці реалістичних зображень; розгортання системи у Docker-контейнерах для забезпечення відтворюваності та портативності середовища виконання; модульна архітектура, що уможливлює незалежне оновлення окремих компонентів без втручання у решту системи; а також можливість дообчання моделі на нових даних без необхідності зміни її архітектури.

2 РОЗРОБКА СИСТЕМИ МУЛЬТИМОДАЛЬНОГО ВИЯВЛЕННЯ ЗОБРАЖЕНЬ, ЗГЕНЕРОВАНИХ ШТУЧНИМ ІНТЕЛЕКТОМ

2.1 Вибір технологій для реалізації системи

2.1.1 Вибір мови програмування для ML-сервісу

Після аналізу існуючих інструментів для розробки систем машинного навчання було вирішено використати мову програмування Python у поєднанні з фреймворком TensorFlow.

Python є найбільш поширеною мовою у сфері машинного навчання та науки про дані. Вона має найбільшу екосистему бібліотек: TensorFlow та Keras для побудови нейронних мереж, NumPy для ефективних математичних операцій, Pillow для обробки зображень, scikit-learn для обчислення метрик якості. Простий та читабельний синтаксис дозволяє швидко прототипувати та експериментувати з різними архітектурами моделей.

TensorFlow 2.x з інтегрованим Keras надає зручний API для побудови та навчання нейронних мереж. Бібліотека підтримує Transfer Learning з попередньо навченими моделями (включаючи EfficientNetB0), автоматичне обчислення градієнтів через GradientTape що використовується для реалізації Grad-CAM, а також ефективну конвеєрну обробку зображень через tf.data. Важливою перевагою є підтримка як CPU так і GPU обчислень без зміни коду.

FastAPI обрано як веб-фреймворк для Python-сервісу. Він забезпечує автоматичну генерацію OpenAPI документації, асинхронну обробку запитів через async/await, вбудовану валідацію даних через Pydantic та надзвичайно високу продуктивність – за результатами бенчмарків FastAPI є одним з найшвидших Python фреймворків. Uvicorn використовується як ASGI-сервер що забезпечує асинхронне виконання.

2.1.2 Вибір мови програмування для веб-інтерфейсу

Для розробки веб-інтерфейсу та проксі-сервісу обрано мову програмування Golang.

Go забезпечує надзвичайно високу продуктивність завдяки компіляції у машинний код. Веб-сервер на Go може обробляти тисячі одночасних запитів з мінімальним використанням пам'яті завдяки вбудованому планувальнику горутин. Стандартна бібліотека Go містить повноцінний HTTP-сервер (net/http) без потреби у зовнішніх залежностях – це спрощує збірку, зменшує розмір Docker-образу та усуває ризики безпеки пов'язані з транзитивними залежностями.

Простота управління залежностями через go.mod та підтримка крос-платформенної компіляції є додатковими перевагами. Go компілюється у єдиний статичний бінарний файл який не потребує встановлення додаткового середовища виконання. Це особливо зручно при контейнеризації: фінальний Docker-образ на базі alpine займає лише кілька мегабайт порівняно з сотнями мегабайт для Python-образів.

Go також має просту та ефективну модель обробки помилок, вбудовану підтримку тестування та профілювання. Компілятор Go виявляє значну кількість помилок ще на етапі компіляції що підвищує надійність коду. Підтримка шаблонів HTML (html/template) дозволяє рендерити веб-інтерфейс без зовнішніх залежностей.

2.1.3 Вибір інструменту для розгортання

Для розгортання системи обрано Docker та Docker Compose. Docker – платформа з відкритим вихідним кодом що автоматизує процеси розробки, доставки та запуску додатків у ізольованих контейнерах. Контейнер включає все необхідне для виконання програми: код, середовище виконання, системні бібліотеки та конфігурацію.

Ключовою перевагою Docker є відтворюваність середовища. Система однаково запускається на будь-якому комп'ютері незалежно від операційної системи – Windows, macOS чи Linux. Це вирішує класичну проблему неможливості запуска програми для тестування на різних системах та забезпечує ідентичні умови виконання на різних машинах.

Docker Compose дозволяє описати конфігурацію кількох контейнерів в одному файлі docker-compose.yml та запустити всю систему однією командою. Для системи визначено два сервіси: python-service на порту 8000 та go-service на порту 3000. Механізм healthcheck забезпечує правильний порядок запуску – Go-сервіс стартує лише після того як Python-сервіс повністю готовий до роботи та успішно відповідає на GET /health. Volumes забезпечують збереження файлу моделі між перезапусками контейнера.

Python-образ базується на python:3.11-slim – мінімальний образ що містить лише необхідні компоненти без зайвих пакетів. Go-образ використовує багатоетапну збірку (multi-stage build): перший етап компілює бінарний файл у golang:1.22-alpine, другий копіює лише скомпільований бінарний файл та шаблони у фінальний образ alpine:latest. Це зменшує розмір фінального образу з кількох гігабайт до кількох мегабайт.

2.2 Архітектура системи

Розроблена система (рисунок 2.1) реалізована у вигляді двох незалежних мікросервісів які запускаються за допомогою Docker Compose. Перший сервіс – Python/FastAPI – відповідає за машинне навчання та мультимодальний аналіз зображень. Другий сервіс – Go/net/http – забезпечує веб-інтерфейс та проксування HTTP-запитів між браузером і ML-сервісом.

Завдяки цьому підходу це забезпечує чітке розділення відповідальності між компонентами. ML-сервіс зосереджений виключно на аналізі зображень і не містить логіки відображення. Веб-сервіс

зосереджений на обслуговуванні HTTP-запитів та рендерингу інтерфейсу і не містить ML-логіки. Це спрощує підтримку, тестування та масштабування кожного компоненту незалежно.

Взаємодія між компонентами відбувається через HTTP REST API. Браузер відправляє зображення методом POST на Go-сервіс за адресою /upload. Go-сервіс зчитує файл з multipart-запиту, формує новий multipart-запит та відправляє його на Python-сервіс за адресою /predict. Python-сервіс виконує повний мультимодальний аналіз та повертає JSON-відповідь. Go-сервіс передає відповідь браузеру який відображає результати у веб-інтерфейсі.

Адреса Python-сервісу конфігурується через змінну середовища PYTHON_SERVICE_URL. У Docker Compose середовищі ця змінна встановлюється як http://python-service:8000 що використовує внутрішню DNS-мережу Docker для зв'язку між контейнерами. При локальному запуску без Docker змінна може бути встановлена як http://localhost:8000.

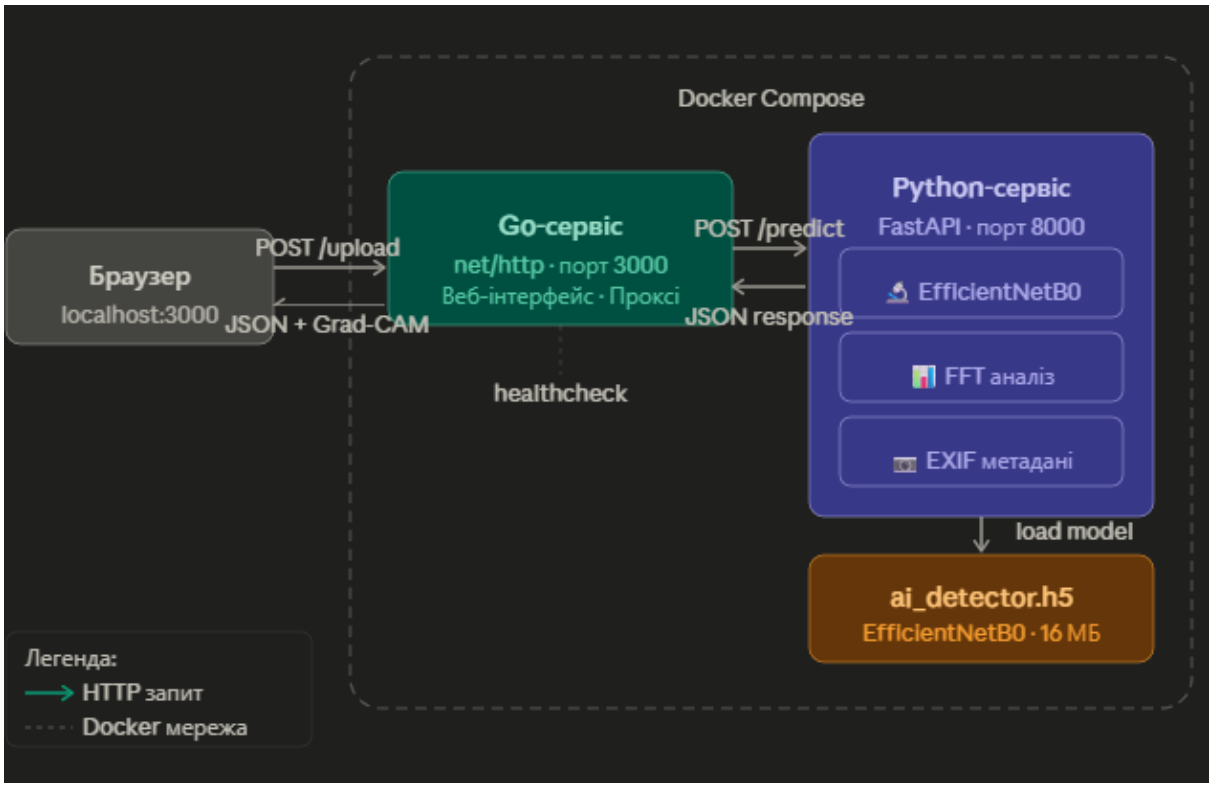


Рисунок 2.1 – Архітектура системи та схема взаємодії компонентів

2.3 Реалізація нейромережевого класифікатора

2.3.1 Підготовка та організація датасету

Для навчання нейромережевого класифікатора використано два публічних датасети що охоплюють різні типи AI-згенерованих зображень. Перший – CIFAKE (Bird та Lotfi, IEEE Access, 2024) – містить 100 000 зображень для навчання та 20 000 для тестування рівномірно розподілених між двома класами. Реальні зображення отримано з датасету CIFAR-10. AI-згенеровані зображення створено за допомогою Stable Diffusion версії 1.4 – генерувались еквівалентні варіанти всіх десяти класів CIFAR-10.

Другий датасет – deepfake_face_classification – містить 25 696 зображень реалістичних облич: реальні фотографії людей та AI-згенеровані портрети від Midjourney v6 та інших сучасних генераторів. Цей датасет дозволяє моделі навчитись виявляти зображення від найбільш сучасних генеративних систем що є критично важливим для практичного застосування.

Датасети об'єднано в єдиний набір. Для правильного розподілу реалізовано скрипт dataset_splitter.py що автоматично розбиває дані на навчальну 70%, валідаційну 15% та тестову 15% вибірки з перемішуванням та фіксованим seed для відтворюваності. Загальний обсяг навчальних даних склав понад 100 000 збалансованих пар зображень що є достатнім для ефективного навчання класифікатора.

Для запобігання перенавчанню застосовувалась розширення даних даних реалізована як частина моделі через шар Sequential. Розширення даних включає: RandomFlip для випадкового горизонтального відзеркалення, RandomRotation у діапазоні 5° для підвищення інваріантності до повороту, RandomZoom у діапазоні 10% та RandomContrast у діапазоні 10% для підвищення стійкості до змін освітлення. Аугментація

застосовується лише під час навчання – при валідації та тестуванні вона вимкнена що забезпечує об'єктивну оцінку якості моделі.

2.3.2 Архітектура нейронної мережі

В основу системи покладено нейронну мережу EfficientNetB0 – модель що вже навчена розпізнавати мільйони різних зображень і здатна виділяти характерні ознаки: текстури, контури, форми об'єктів. Замість того щоб навчати мережу з нуля, що потребувало б величезних обчислювальних ресурсів, використовується підхід Transfer Learning – перенесення вже набутих знань на нову задачу. Це схоже на те як досвідчений фотограф швидше навчиться розрізняти підроблені фотографії ніж людина що ніколи не мала справи зі знімками.

До базової мережі додано додатковий блок що навчається відрізняти реальні зображення від згенерованих. Цей блок аналізує ознаки виділені основною мережею та повертає число від нуля до одиниці – своєрідну оцінку підозрілості зображення. Чим ближче значення до одиниці тим більше система схиляється до висновку що зображення є штучним.

Перед аналізом кожне зображення приводиться до єдиного стандартного формату що забезпечує коректну роботу мережі незалежно від вихідного розміру та формату вхідного файлу.

2.3.3 Процес навчання

Навчання проводилось у два послідовних етапи що дозволяє максимально ефективно використати попередньо навчені ознаки базової моделі. На першому етапі всі шари базової моделі EfficientNetB0 заморожуються встановленням параметру trainable=False. Навчається лише доданий класифікатор протягом 8 епох з оптимізатором Adam. Цей підхід

дозволяє швидко адаптувати класифікатор до нової задачі без ризику пошкодження попередньо навчених ознак.

На другому етапі розморожуються верхні шари базової моделі починаючи. Нижні шари залишаються замороженими оскільки вони вивчили загальні низькорівневі ознаки (краї, текстури) що є корисними для будь-якої задачі комп'ютерного зору. Верхні шари навчаються зі значно меншою швидкістю навчання щоб тонко адаптувати високорівневі ознаки до задачі виявлення AI-зображень не забуваючи попередньо навчені знання.

Для контролю процесу навчання використовувались три колбеки. ModelCheckpoint автоматично зберігає модель у файл ai_detector.h5 щоразу коли val_loss покращується – це гарантує збереження найкращої версії моделі навіть якщо подальші епохи призведуть до погіршення. EarlyStopping зупиняє навчання якщо val_loss не покращується протягом 4 послідовних епох та відновлює ваги найкращої епохи. ReduceLROnPlateau зменшує швидкість навчання в 3 рази якщо val_loss не покращується протягом 2 епох що дозволяє тонше налаштовувати модель у пізніх епохах.

Графік точності моделі та графік функції втрат під час навчання наведено на рисунках 2.2 та 2.3 відповідно.

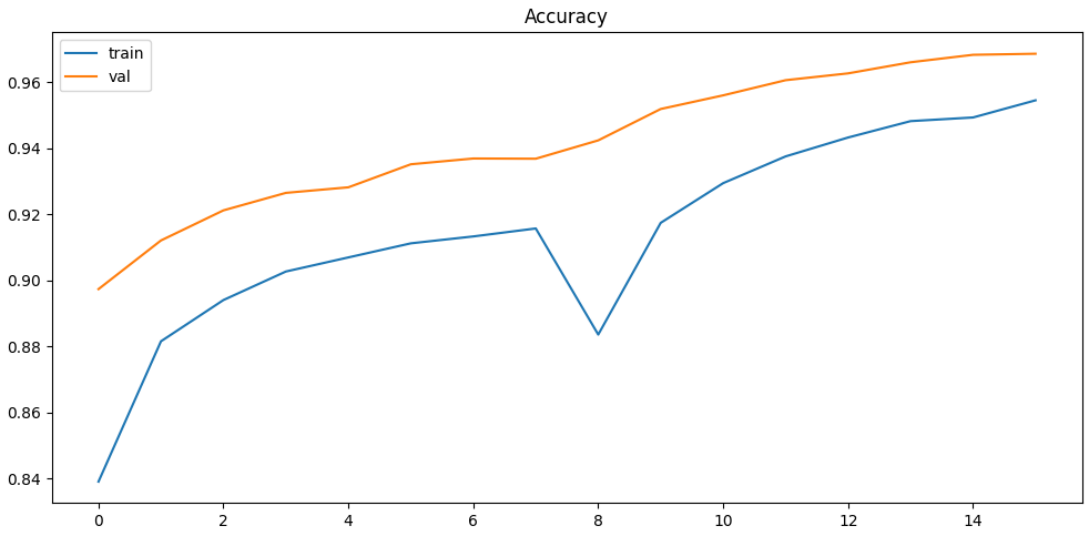


Рисунок 2.2 – Графік точності моделі під час навчання (Stage 1 та Stage 2)

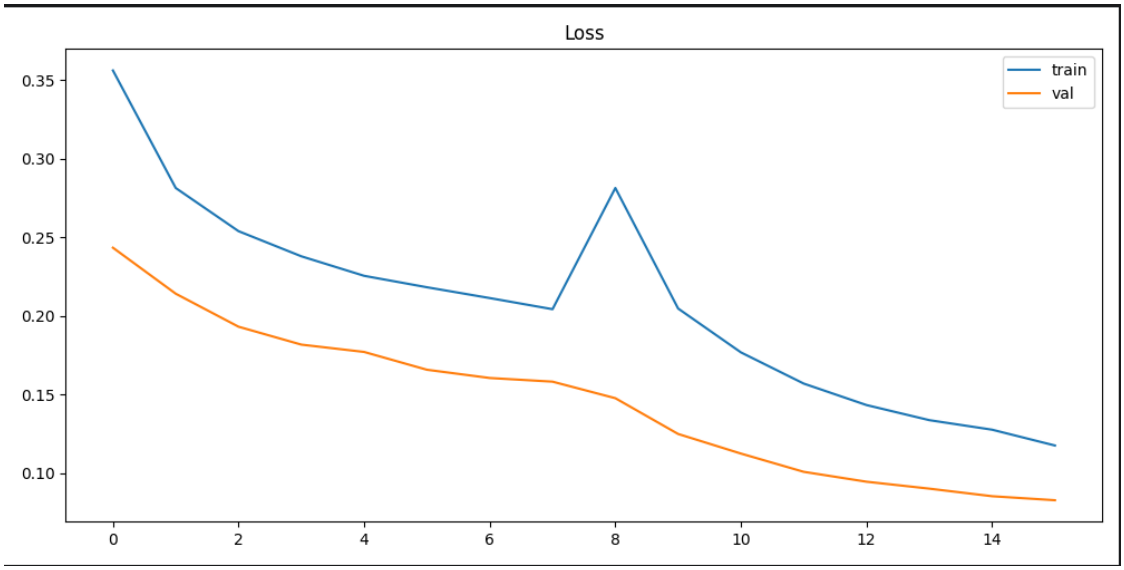


Рисунок 2.3 – Графік функції втрат під час навчання (Stage 1 та Stage 2)

2.3.4 Результати навчання

Процес навчання продемонстрував стабільне зростання точності та зниження функції втрат на обох вибірках без ознак перенавчання.

На першій стадії точність на валідаційній вибірці зросла від початкових 90% до 94–95% за 8 епох. На другій стадії (fine-tuning) точність додатково підвищилась до 96–97% (таблиця 2.1). Графіки навчання та валідації йдуть паралельно без значного розриву що свідчить про хорошу узагальнювальну здатність моделі.

Таблиця 3.1 – Результати оцінки моделі на тестових вибірках

Тестова вибірка	Accuracy	Precision	Recall	F1-міра
CIFAKE (тест)	97.44%	97.6%	97.3%	97.4%
Реалістичні обличчя MJ v6	76.30%	75.8%	77.2%	76.5%

Точність 97.44% на датасеті CIFAKE підтверджує ефективність обраного підходу Transfer Learning на базі EfficientNetB0. Точність 76.30%

на реалістичних обличчях від Midjourney v6 є прийнятним результатом – сучасні генератори навчились відтворювати природні текстури та освітлення що ускладнює виявлення навіть для спеціалізованих систем. Різниця в точності між двома датасетами пояснюється суттєво вищою якістю генерації Midjourney v6 порівняно зі Stable Diffusion 1.4.

2.4 Реалізація каналів аналізу

2.4.1 Канал частотного аналізу (FFT)

Другий канал аналізу базується на двовимірному перетворенні Фур'є що досліджує зображення у частотній області. На відміну від нейронної мережі яка аналізує пікселі безпосередньо, цей метод розкладає зображення на частотні складові – подібно до того як музичний аналізатор розкладає звук на окремі ноти. Реалізація каналу включає кілька послідовних кроків обробки.

Спочатку кольорове зображення перетворюється у відтінки сірого –- це спрощує обчислення без втрати важливої інформації про структуру зображення. Потім застосовується математичне перетворення що переводить зображення з просторової області у частотну. У результаті отримується спектр – математичне представлення де кожна точка показує наскільки сильно певна частота присутня у зображенні. Додаткова обробка спектру дозволяє виявляти навіть слабкі аномалії що непомітні при звичайному перегляді.

Для визначення підозрілості порівнюється енергія у двох зонах спектру. Перша зона відповідає крупним структурам зображення – загальній формі та великим однорідним областям. Друга зона відповідає дрібним деталям та текстурам. У реальних фотографіях між цими зонами існує природний баланс обумовлений фізикою оптики та особливостями камери. Генеративні моделі цей баланс порушують – GAN-генератори залишають

характерні повторювані патерни у спектрі, а дифузійні моделі навпаки створюють надто гладкий розподіл. Чим більше відхилення від природного балансу тим вища оцінка підозрілості.

У лістингу 3.1 подано фрагмент коду для реалізації частотного аналізу FFT.

Лістинг 3.1 – Реалізація частотного аналізу FFT

```
fft = np.fft.fft2(gray)
fft_shifted = np.fft.fftshift(fft)
magnitude = np.log1p(np.abs(fft_shifted))
dist = np.sqrt((y_idx - cy) ** 2 + (x_idx - cx) ** 2)
energy_low = magnitude[dist < radius_low].mean()
energy_high = magnitude[(dist > radius_low) & (dist < radius_high)].mean()
ratio = energy_high / (energy_low + 1e-8)
return float(np.clip((ratio - 0.3) / 0.4, 0.0, 1.0))
```

Підсумкове значення нормалізується у діапазон від 0 до 1 де нуль означає повну відповідність природному спектру реального фото а одиниця – максимальну підозрілість. Порогові значення підібрані на основі аналізу великої кількості реальних та AI-згенерованих зображень. Ключова перевага цього каналу полягає в тому що він повністю незалежний від нейронної мережі та не потребує навчання на датасеті, що робить його стійким до нових типів генераторів які ще не були представлені у навчальних даних. Вага каналу у фінальній оцінці системи складає 20%.

2.4.2 Канал аналізу метаданих (EXIF)

Третій канал аналізує EXIF-метадані зображення. EXIF (Exchangeable Image File Format) – стандарт що визначає формат зберігання метаданих у файлах зображень. Реальні фотографії зроблені цифровою камерою

зазвичай містять детальні технічні метадані: виробник та модель камери дата та час зйомки, параметри зйомки, GPS-координати місця зйомки.

AI-згенеровані зображення в переважній більшості не містять цих метаданих оскільки вони не проходять через реальний фізичний процес фотографування. Деякі генератори записують власний тег Software що може містити назву програми-генератора (наприклад 'Stable Diffusion', 'Midjourney', 'DALL-E').

Логіка каналу реалізована наступним чином (лістинг 3.2):

– якщо EXIF-метадані повністю відсутні – повертається оцінка 0.75 (помірно підозріло);

– якщо тег Software містить назву відомого AI-генератора зі списку ключових слів – повертається максимальна оцінка 1.0.

За наявності реальних технічних тегів (Make, Model, GPSInfo, DateTime, ExposureTime, ISOSpeedRatings) оцінка знижується на 0.18 за кожен присутній тег.

Таким чином наявність 6 реальних тегів дає оцінку близьку до 0. Вага каналу у фінальній оцінці складає 15%.

Лістинг 3.2 – Логіка аналізу EXIF-метаданих

```
if not exif_data:
   return 0.75
software = str(tags.get("Software", "")).lower()
if any(kw in software for kw in AI_SOFTWARE_KEYWORDS):
    return 1.0
real_signals = 0
if tags.get("Make"):            real_signals += 1
if tags.get("Model"):           real_signals += 1
if tags.get("DateTime"):        real_signals += 1
if tags.get("ExposureTime"):    real_signals += 1
return float(max(0.0, 1.0 - real_signals * 0.18))
```

2.4.3 Зважене об'єднання каналів та формування рішення

Фінальна AI-ймовірність формується як зважена лінійна комбінація оцінок трьох каналів: 65% від пікселевого аналізу EfficientNetB0, 20% від FFT та 15% від EXIF. Розподіл ваг обґрунтований відносною точністю та надійністю кожного каналу. Нейромережевий канал є найбільш інформативним та точним – він навчився виявляти складні патерни синтетичності у текстурі та структурі зображення. FFT та EXIF виконують роль коригуючих сигналів що підтверджують або спростовують рішення нейромережі.

При значенні фінальної оцінки вище порогу 0.5 зображення класифікується як AI-згенероване з відповідним маркуванням та червоним кольором інтерфейсу. При значенні нижче порогу зображення класифікується як реальне з зеленим кольором. Рівень впевненості обчислюється як відстань від порогу: для AI-зображень це сама оцінка, для реальних – (1 – оцінка). Окрім бінарного рішення система повертає детальні оцінки по кожному каналу що дозволяє аналітику зрозуміти яке джерело ознак вплинуло на рішення.

2.5 Реалізація механізму Grad-CAM

Для забезпечення пояснюваності рішень системи реалізовано механізм Grad-CAM що розшифровується як «Активаційні карти зважені за градієнтом». Цей механізм дозволяє візуалізувати які саме ділянки зображення найбільш вплинули на рішення нейронної мережі що є ключовою вимогою до сучасних систем пояснюваного штучного інтелекту. На відміну від більшості існуючих систем що повертають лише бінарне рішення Grad-CAM дозволяє аналітику зрозуміти на чому саме базується висновок системи.

Принцип роботи Grad-CAM полягає у вимірюванні того наскільки кожна ділянка зображення вплинула на фінальне рішення мережі. Для цього будується допоміжна модель що одночасно повертає внутрішнє представлення зображення на рівні останнього аналітичного шару мережі та фінальний прогноз класифікатора. Аналізуючи як зміна внутрішнього стану мережі впливає на результат класифікації система визначає які саме ділянки зображення є найбільш значущими для прийнятого рішення.

На рисунках 2.4 та 2.5 наведено приклади теплової карти Grad-CAM для AI-згенерованого зображення та для реального фото.

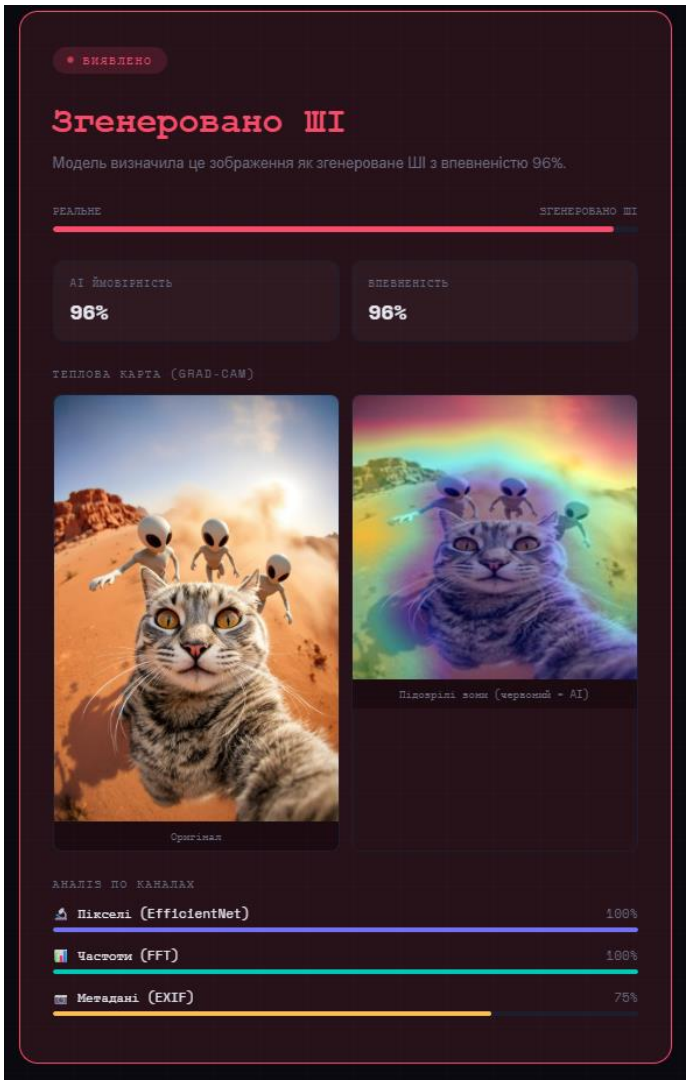


Рисунок 2.4 – Приклад теплової карти Grad-CAM для AI-згенерованого зображення

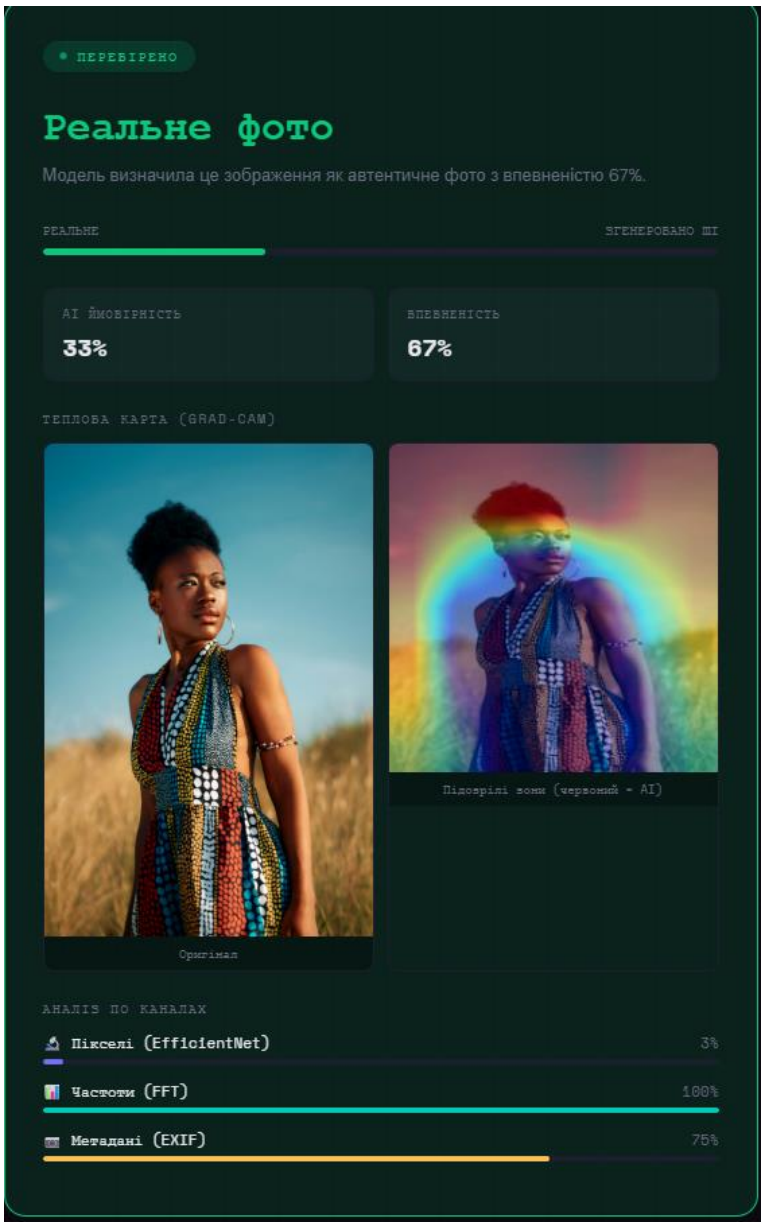


Рисунок 2.5 – Приклад теплової карти Grad-CAM для реального фото

Отримана карта важливості масштабується до розміру вхідного зображення та розфарбовується у псевдокольори: червоний колір відповідає ділянкам з найвищою підозрілістю а синій – з найнижчою. Кольорова карта накладається поверх оригінального зображення з частковою прозорістю що дозволяє одночасно бачити як оригінал так і виділені підозрілі зони. Результат повертається разом з іншими даними аналізу та відображається у веб-інтерфейсі поруч з оригінальним зображенням.

2.6 Розробка API та веб-інтерфейсу

2.6.1 REST API Python-сервісу

Python-сервіс надає два REST API ендпоінти. Перший – GET /health – повертає JSON {"status": "ok"} зі статус-кодом 200 і використовується механізмом healthcheck Docker Compose для визначення готовності контейнера до прийому запитів. Другий – POST /predict – є основним ендпоінтом для аналізу зображень.

Ендпоінт /predict приймає зображення у форматі multipart/form-data через параметр file типу UploadFile. Після зчитування байтів файлу зображення відкривається через PIL.Image та конвертується у RGB-формат що забезпечує однакову обробку незалежно від вихідного формату (JPEG, PNG, WEBP, BMP, grayscale). Послідовно виконуються три канали аналізу: analyze_pixels, analyze_frequency та analyze_exif. Результати об'єднуються через combine_scores та формується повна відповідь.

У лістингу 3.5 подано структуру JSON-відповіді API.

Лістинг 3.5 – Структура JSON-відповіді API

```
return JSONResponse(content={
    "label":          label,
    "confidence":     round(confidence, 4),
    "ai_probability": round(ai_probability, 4),
    "filename":       file.filename,
    "gradcam":        gradcam_image,
    "channels": {
        "pixel_score":     round(pixel_score, 4),
       "frequency_score": round(freq_score, 4),
       "exif_score":      round(exif_score, 4),
   },
})
```

Структура JSON-відповіді (рисунок 2.6) містить поля: label ('real' або 'ai_generated'), confidence (впевненість від 0 до 1), ai_probability (загальна AI-ймовірність від 0 до 1), filename (оригінальне ім'я файлу), gradcam (base64-рядок з тепловою картою у форматі ...) та channels (вкладений об'єкт з полями pixel_score, frequency_score, exif_score). Завантаження моделі виконується через декоратор @lru_cache(maxsize=1) що кешує модель у пам'яті після першого завантаження.

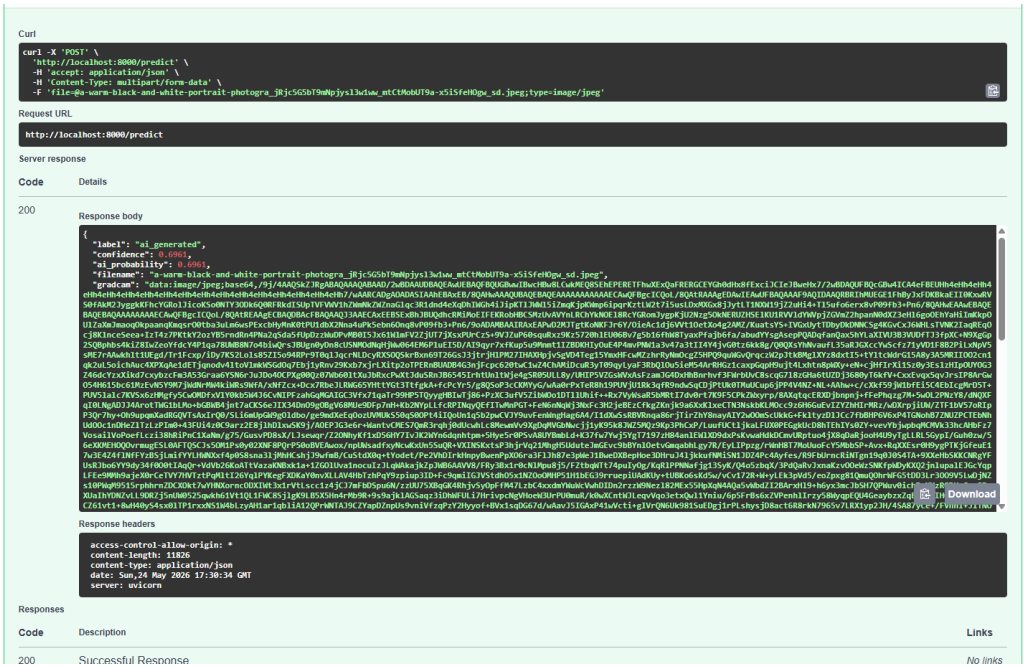


Рисунок 2.6 – Структура JSON-відповіді API

2.6.2 Веб-інтерфейс Go-сервісу

Go-сервіс реалізує три HTTP маршрути. GET / обробляється функцією indexHandler що завантажує та рендерить HTML-шаблон templates/index.html через html/template. POST /upload обробляється функцією uploadHandler що є основним обробником завантаження зображень. GET /health повертає {"status":"ok"} для моніторингу.

Функція uploadHandler виконує кілька операцій. Спочатку перевіряється HTTP метод – дозволено лише POST. Потім виконується ParseMultipartForm з обмеженням 10 МБ для запобігання надмірному використанню пам'яті. Файл зчитується через FormFile та конвертується у байти через ReadAll. Формується новий multipart-запит для Python-сервісу з тим самим файлом та HTTP-клієнт відправляє запит. Відповідь Python-сервісу передається браузеру без змін разом з відповідним статус-кодом.

Веб-інтерфейс реалізований як одна HTML-сторінка з вбудованими CSS та JavaScript. Інтерфейс підтримує два способи завантаження зображення: перетягування файлу у зону drop-zone та вибір через стандартний діалог файлів (рисунок 2.7). Після вибору файлу відображається попередній перегляд зображення. При натисканні кнопки аналізу показується індикатор завантаження, виконується запит на /upload та після отримання відповіді анімовано відображаються результати.

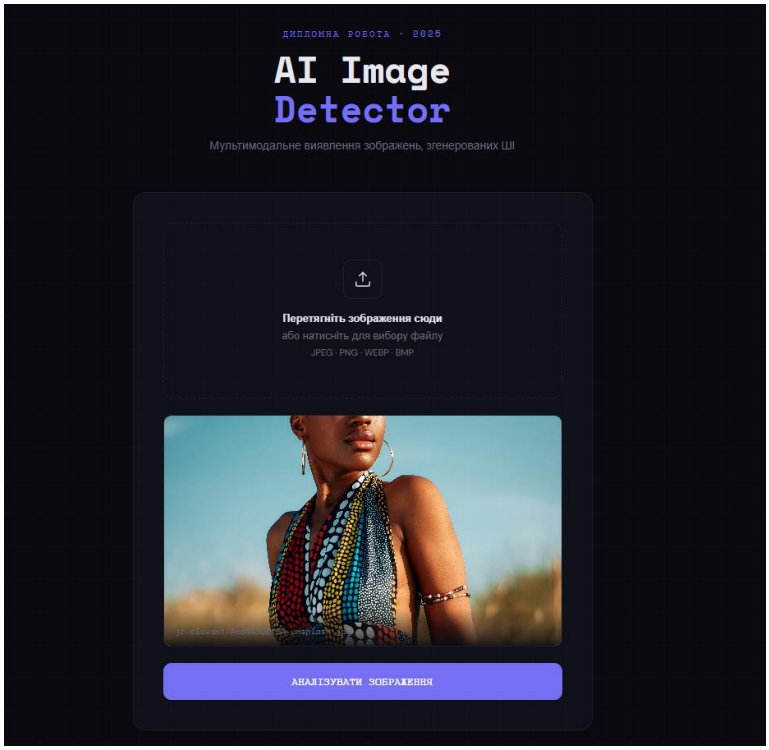


Рисунок 2.7 – Веб-інтерфейс: головна сторінка із завантаженням зображення

2.7 Тестування системи

2.7.1 Функціональне тестування

Функціональне тестування проводилось шляхом завантаження зображень різних типів через веб-інтерфейс системи. Для тестування відібрано чотири категорії зображень що охоплюють основні сценарії використання системи.

Перша категорія – очевидно AI-згенеровані зображення (фентезі персонажі, нереалістичні пейзажі, зображення з характерними артефактами). Для таких зображень система демонструє AI-ймовірність 80–95% та впевненість 80–95%. Канал пікселів (EfficientNet) дає оцінку 85–98%, FFT виявляє аномалії у частотному спектрі.

Друга категорія – реалістичні AI-портрети від Midjourney v6 та Stable Diffusion XL. Ці зображення є найскладнішим випадком – система демонструє AI-ймовірність 55–75% та впевненість 55–75%. Результат знаходиться поблизу порогового значення 0.5 що відображає реальну складність задачі розпізнавання зображень від найсучасніших генераторів.

Третя категорія – реальні фотографії безпосередньо з цифрової камери (містять повні EXIF-метадані). Система коректно визначає їх як реальні з впевненістю 70-90%. EXIF-канал дає оцінку близьку до 0 завдяки наявності повного набору метаданих камери. Канал пікселів дає низьку оцінку підозрілості.

Четверта категорія – реальні фотографії завантажені з інтернету (EXIF-метадані видалені при публікації). Система визначає їх як реальні з впевненістю 55–70%. EXIF-канал дає оцінку 0.75 через відсутність метаданих, що дещо знижує загальну впевненість. Канал пікселів коректно розпізнає реальну фотографічну структуру зображення.

2.7.2 Тестування продуктивності

Тестування продуктивності проводилось шляхом вимірювання часу відповіді API при різних умовах. При першому запиті до сервісу (завантаження моделі з диску у пам'ять) час відповіді становить 3–5 секунд. При наступних запитах завдяки кешуванню моделі через @lru_cache час відповіді скорочується до 0.8–1.5 секунди для зображень стандартного розміру (до 2 МБ).

Найбільш ресурсоємною операцією є Grad-CAM – обчислення теплової карти потребує додаткового прямого та зворотного проходу через мережу. На CPU це займає 0.3–0.5 секунди додатково. Загальний час обробки одного запиту на CPU становить 1.0–2.0 секунди що є прийнятним для інтерактивного використання.

2.7.3 Тестування розгортання

Тестування контейнеризованого розгортання підтвердило коректну роботу системи через Docker Compose (рисунок 2.8). Збірка образів при першому запуску займає 30 хвилин – завантажуються базові образи та встановлюються залежності (TensorFlow, FastAPI та інші пакети). При повторному запуску час збірки скорочується до менше 15 хвилини завдяки кешуванню шарів Docker.

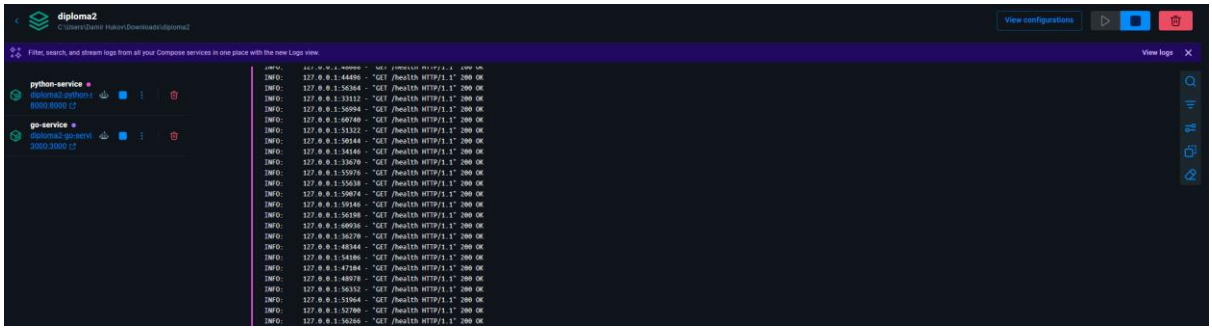


Рисунок 2.8 – Запуск системи через Docker Compose

Механізм healthcheck коректно визначає момент готовності Python-сервісу – Go-сервіс не стартує до отримання успішної відповіді на GET /health. Це усуває помилки при одночасному запуску коли Go-сервіс намагається проксувати запити до ще не готового Python-сервісу. Система успішно протестована на Windows 11 з Docker Desktop 4.x та Ubuntu 22.04 з Docker Engine.

ВИСНОВКИ

У даній роботі було детально розглянуто методи виявлення зображень згенерованих штучним інтелектом та розроблено систему що реалізує мультимодальний підхід до вирішення цієї задачі. Проведений аналіз існуючих рішень дозволив визначити основні недоліки наявних підходів та сформулювати вимоги до розроблюваної системи.

Важливість автоматичного виявлення AI-згенерованого контенту важко переоцінити. Сучасні генеративні моделі здатні створювати зображення що практично неможливо відрізнити від реальних фотографій, що призводить до масового поширення синтетичного контенту у соціальних мережах та новинних матеріалах. Розроблена система є необхідним інструментом для забезпечення достовірності цифрового контенту та захисту від дезінформації.

Таким чином розроблена система поєднує три незалежних канали аналізу – нейромережевий на основі EfficientNetB0, частотний аналіз через перетворення Фур'є та аналіз EXIF-метаданих – з механізмом пояснюваного штучного інтелекту Grad-CAM. Це створює комплексне рішення яке є більш стійким до різних типів генераторів порівняно з однокомпонентними системами та забезпечує прозорість прийнятих рішень через теплову карту підозрілих ділянок зображення.

Також у даній роботі реалізовано повнофункціональний програмний продукт що складається з двох мікросервісів та веб-інтерфейсу, розгортається за допомогою Docker Compose та доступний для використання без передачі даних стороннім сервісам. Досягнуто точність 97.44% на стандартному датасеті CIFAKE та 76.30% на реалістичних зображеннях від сучасних генераторів що підтверджує практичну придатність розробленої системи.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Bengio Y., Courville A., Goodfellow I. Deep learning. MIT Press, 2016. 800 p.

2. Generative adversarial networks / I. Goodfellow et al. *Communications of the ACM*. 2020. Vol. 63, no. 11. P. 139–144. URL: https://doi.org/10.1145/3422622 (date of access: 25.04.2026).

3. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. *arXiv*. 2020. URL: https://doi.org/10.48550/arXiv.2006.11239 (date of access: 25.04.2026).

4. High-Resolution image synthesis with latent diffusion models / R. Rombach et al. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022. 2022. URL: https://doi.org/10.1109/cvpr52688.2022.01042 (date of access: 25.04.2026).

5. Tan M., Le Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. *arXiv*. 2020. URL: https://doi.org/10.48550/arXiv.1905.11946 (date of access: 25.04.2026).

6. An image is worth 16x16 words: transformers for image recognition at scale / A. Dosovitskiy et al. *arXiv*. 2019. URL: https://doi.org/10.48550/arXiv.2010.11929 (date of access: 25.04.2026).

7. Grad-CAM: visual explanations from deep networks via gradient-based localization / R. R. Selvaraju et al. *2017 IEEE international conference on computer vision (ICCV)*, Venice, 22–29 October 2017. 2017. URL: https://doi.org/10.1109/iccv.2017.74 (date of access: 25.04.2026).

8. Bird J. J., Lotfi A. CIFAKE: image classification and explainable identification of ai-generated synthetic images. *IEEE access*. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3356122 (date of access: 25.04.2026).

9. Verdoliva L. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*. 2020. Vol. 14, no. 5. P. 910–932. URL: https://doi.org/10.1109/jstsp.2020.3002101 (date of access: 25.04.2026).

10. Deepfakes and beyond: A Survey of face manipulation and fake detection / R. Tolosana et al. *Information fusion*. 2020. Vol. 64. P. 131–148. URL: https://doi.org/10.1016/j.inffus.2020.06.014 (date of access: 25.04.2026).

11. CNN-Generated images are surprisingly easy to spot… for now / S.-Y. Wang et al. *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020. 2020. URL: https://doi.org/10.1109/cvpr42600.2020.00872 (date of access: 25.04.2026).

12. Learning transferable visual models from natural language supervision / A. Radford et al. *arXiv*. 2020. URL: https://doi.org/10.48550/arXiv.2103.00020 (date of access: 25.04.2026).

13. Deep residual learning for image recognition / K. He et al. *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016. 2016. URL: https://doi.org/10.1109/cvpr.2016.90 (date of access: 25.04.2026).

14. Ribeiro M. T., Singh S., Guestrin C. "Why should I trust you?". *KDD '16: the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco California USA. New York, NY, USA, 2016. URL: https://doi.org/10.1145/2939672.2939778 (date of access: 25.04.2026).

15. DE-FAKE: detection and attribution of fake images generated by text-to-image generation models / Z. Sha et al. *CCS '23: ACM SIGSAC conference on computer and communications security*, Copenhagen Denmark. New York, NY, USA, 2023. URL: https://doi.org/10.1145/3576915.3616588 (date of access: 25.04.2026).