МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ «КИЇВСЬКИЙ АВІАЦІЙНИЙ ІНСТИТУТ»

ДОПУСТИТИ ДО ЗАХИСТУ
Завідувачка кафедри

_____Олена НЕЧИПОРУК
“_____”_____2025 р.

КВАЛІФІКАЦІЙНА РОБОТА

(ПОЯСНЮВАЛЬНА ЗАПИСКА)

ЗДОБУВАЧА ОСВІТНЬОГО СТУПЕНЯ «МАГІСТР»

Тема: Інтелектуальна система виявлення *deepfakes* у відео з використанням глибокого навчання

Виконавець: Катерина СТАРЕНЬКА

Керівник: Ольга СУПРУН

Нормоконтролер: Євгеній ТУПОТА

Київ 2025

ДЕРЖАВНИЙ УНІВЕРСИТЕТ «КИЇВСЬКИЙ АВІАЦІЙНИЙ ІНСТИТУТ»

Факультет комп'ютерних наук та технологій

Кафедра інтелектуальних кібернетичних систем

Спеціальність 126 "Інформаційні системи та технології"

(шифр, найменування)

Освітньо професійна програма «Інтелектуальні системи та технології»

Форма навчання денна

ЗАТВЕРДЖУЮ
Завідувачка кафедри
_____Олена НЕЧИПОРУК
"_____"_____2025 р.

ЗАВДАННЯ
на виконання кваліфікаційної роботи
Старенької Катерини Олександрівни
(прізвище, ім'я, по батькові випускника в родовому відмінку)

1. **Тема кваліфікаційної роботи:** «Інтелектуальна система виявлення *deepfakes* у відео з використанням глибокого навчання» затверджена наказом ректора від «29» вересня 2025р. №1575/ст.
2. **Термін виконання роботи**: з 29.09.2025 по 31.12.2025
3. **Вихідні дані до роботи (проєкту)**: датасет «*Deepfake and Real Images*», тестові відеофайли, попередньо навчена модель *EfficientNet-B0*, програмні бібліотеки комп'ютерного зору та глибокого навчання.
4. **Зміст пояснювальної записки**: аналіз існуючих методів та систем виявлення *deepfake*, проєктування архітектури інтелектуальної системи та розробка методу *Multi-Frame Statistical Analysis*, реалізація моделі з використанням трансферного навчання, розробка вебінтерфейсу системи.
5. **Перелік обов'язкового графічного (ілюстративного) матеріалу:**
   1. Діаграма прецедентів використання системи;
   2. Архітектура інтелектуальної системи;
   3. Схема алгоритму *flowchart* процесу аналізу відео;
   4. Інтерфейс розробленого вебзастосунку з результатами аналізу.

6. Календарний план–графік

№ пор.	Завдання	Термін виконання	Відмітка про виконання
1	Дослідження феномену *deepfake*, його класифікації, соціальних та етичних наслідків	29.09.2025-05.10.2025	
2	Огляд методів створення та поширення *deepfake*, існуючих підходів їх виявлення. Написання Вступу та Розділу 1	06.10.2025-12.10.2025	
3	Аналіз аналогів, формування загального опису системи	13.10.2025-17.10.2025	
4	Формулювання вимог, їх пріоритезація, проєктування сценаріїв використання системи та написання Розділу 2	18.10.2025-22.10.2025	
5	Вибір підходу навчання, розроблення архітектури інтелектуальної системи	23.10.2025-29.10.2025	
6	Структурування даних для навчання, реалізація модулів попередньої обробки, детектора обличчя (*MTCNN*) та просунутої аугментації	30.10.2025-05.11.2025	
7	Налаштування та впровадження алгоритму *Fine-tuning*. Початок тренувального процесу	06.11.2025-09.11.2025	
8	Проведення навчання моделі, аналіз метрик ефективності	10.11.2025-20.11.2025	
9	Реалізація статистичного аналізу та алгоритму визначення фінального рішення	21.11.2025-30.11.2025	
10	Розроблення вебзастосунку, інтеграція всіх модулів системи та опис обраних інструментів. Написання Розділу 3.	01.12.2025-07.12.2025	
11	Огляд та опис прототипу, визначення подальших удосконалень. Написання Розділу 4 та Висновку.	08.12.2025-10.12.2025	
12	Фінальне форматування пояснювальної записки, підготовка презентації та доповіді.	11.12.2025-25.12.2025	
13	Захист кваліфікаційної роботи	26.12.2025	

7. Дата видачі завдання: «29» вересня 2025р

Керівник кваліфікаційної роботи _____ Ольга СУПРУН
(підпис керівника) (П.І.Б.)

Завдання прийняв до виконання _____ Катерина СТАРЕНЬКА
(підпис випускника) (П.І.Б.)

РЕФЕРАТ

Кваліфікаційна робота «Інтелектуальна система виявлення *deepfakes* у відео з використанням глибокого навчання» містить 90 сторінок, 23 рисунків, 6 таблицю, 27 використаних джерел.

Об'єкт дослідження: процеси автоматизованого виявлення синтетичних змін у відеопотоці, створених за допомогою генеративних технологій штучного інтелекту.

Предмет дослідження: інтелектуальна система виявлення *deepfake*-відео на основі трансферного навчання з використанням архітектури *EfficientNet-B0* та багатокадрового статистичного аналізу темпоральної послідовності.

Мета кваліфікаційної роботи: дослідити, розробити та реалізувати інтелектуальну систему автоматизованого виявлення *deepfake*-відео з використанням методів глибокого навчання та статистичного аналізу для забезпечення інформаційної безпеки та протидії поширенню дезінформації.

Метод дослідження: аналіз існуючих підходів до виявлення цифрових підробок, застосування методу трансферного навчання та тонкого налаштування нейронної мережі, попередня обробка даних за допомогою алгоритмів комп’ютерного зору, статистичний аналіз часових рядів та експериментальне тестування на збалансованому датасеті.

Результат проєкту: інтелектуальна програмна система з вебінтерфейсом для автоматизованої класифікації відеоконтенту на предмет наявності *deepfake*-маніпуляцій, що поєднує покадрову оцінку з аналізом темпоральної стабільності, забезпечує точність діагностики та надає детальну візуалізацію метрик достовірності.

DEEPFAKE, ГЛИБОКЕ НАВЧАННЯ, *EFFICIENTNET*, КОМП'ЮТЕРНИЙ ЗІР, НЕЙРОННІ МЕРЕЖІ, ТРАНСФЕРНЕ НАВЧАННЯ, СТАТИСТИЧНИЙ АНАЛІЗ, ШТУЧНИЙ ІНТЕЛЕКТ.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ, ТЕРМІНІВ

GAN – Генеративно-змагальна мережа.

CN – Згорткова нейронна мережа.

RNN – Рекурентна нейронна мережа.

LSTM – Довга коротка пам'ять.

VAE – Варіаційний автокодувальник.

DSA – Аналіз доменних ознак.

GDPR – Загальний регламент про захист даних.

DFDC – Датасет *DeepFake Detection Challenge*

AUC – Площа під *ROC*-кривою

ШІ/AI – Штучний інтелект

PPG – Фотоплетизмографія

MP4 – Формат мультимедійних файлів *MPEG*-4

AVI – Переплетення аудіо-відео

MOV – Формат файлів *QuickTime*

WebM – Контейнер *WebM*

API – Інтерфейс прикладного програмування

PDF – Портативний формат документів

JSON – Нотація об'єктів *JavaScript*

GPU – Графічний процесор

CPU – Центральний процесор

MTCNN – Багатозадачна каскадна згорткова мережа

AMP – Автоматична змішана точність

ВСТУП

Розвиток сучасних інформаційних технологій докорінно змінив спосіб, у який створюється та споживається цифровий контент. Штучний інтелект став невід'ємною частиною повсякденного життя, дозволяючи автоматизувати складні завдання: від покращення якості фотографій до створення цілих віртуальних світів у кіноіндустрії. Однак поруч із корисними інноваціями виникли й серйозні загрози, пов'язані з можливістю легкої маніпуляції візуальними даними. Сьогодні технології глибокого навчання дозволяють генерувати настільки реалістичні зображення та відео, що відрізнити їх від справжніх стає дедалі важче навіть для досвідчених фахівців.

Поява діпфейків стала викликом не лише для окремих осіб, чию репутацію можна зіпсувати за лічені хвилини, а й для стабільності суспільства загалом. Можливість створення фальшивих виступів відомих людей чи підробки відеодоказів створює атмосферу недовіри до будь-якої інформації, отриманої з інтернету. Оскільки обсяги такого контенту зростають щодня, ручна перевірка кожного файлу стає фізично неможливою. Саме тому вкрай актуальним є створення розумних автоматизованих систем, які здатні самостійно аналізувати відео, помічати найменші технічні помилки штучного інтелекту та захищати користувачів від дезінформації.

Об'єктом дослідження є процеси автоматизованого виявлення штучних маніпуляцій у відеоматеріалах. У межах роботи розглядається те, як саме змінюється структура цифрового зображення під впливом генеративних технологій і які специфічні сліди залишаються у відеопотоці після втручання штучного інтелекту.

Предметом дослідження виступає інтелектуальна система детекції, що базується на поєднанні аналізу окремих кадрів та перевірки того, наскільки плавно змінюється зображення протягом усього ролика. Основна увага приділена виявленню аномалій та технічних помилок, які виникають у процесі створення

підробок і проявляються у вигляді мікроскопічних дефектів або порушення стабільності картинки.

Метою роботи є створення та практична реалізація надійного програмного засобу для автоматичної перевірки автентичності відео. Розроблена система має забезпечувати високу точність розпізнавання маніпуляцій та допомагати користувачам швидко відрізняти оригінальний контент від професійно створених фальшивок.

Методи дослідження включають комплексний підхід, що поєднує аналіз сучасних технологій машинного зору та глибокого навчання. Для розпізнавання ознак підробки використано метод тонкого налаштування нейронної мережі на великих наборах даних. Окрім візуального аналізу зображень, застосовано статистичні методи для перевірки часової стабільності відео, що дозволяє відстежувати різкі стрибки та відхилення у послідовності кадрів. Надійність системи підтверджено шляхом експериментального тестування на спеціалізованих базах даних.

Результатом проєкту є функціональний вебзастосунок із доступним інтерфейсом. Програма дозволяє автоматично обробляти завантажені відеофайли, знаходити на них обличчя та проводити їх поглиблений аналіз. Після завершення перевірки система надає користувачеві не лише остаточний висновок, а й детальний звіт із графіками та поясненнями, які обґрунтовують прийняте рішення на основі виявлених підозрілих ознак.

Наукова новизна результатів полягає в удосконаленні підходу до виявлення підробок завдяки впровадженню аналізу темпоральної (часової) стабільності відео. На відміну від більшості підходів, які оцінюють кожен кадр як окреме зображення, запропоноване рішення використовує спеціальний комплексний показник стабільності. Це дозволяє враховувати динаміку змін у часі та ідентифікувати приховане «мерехтіння» або хаотичність підробки, що значно підвищує стійкість системи до високоякісних маніпуляцій, які важко розпізнати на рівні поодиноких кадрів.

РОЗДІЛ 1
АНАЛІЗ ПРОБЛЕМАТИКИ ВИЯВЛЕННЯ *DEEPFAKE*-ВІДЕО

1.1. Проблеми достовірності інформації у цифровому середовищі

У сучасному світі інформація поширюється з надзвичайною швидкістю і у великих об’ємах, що спричиняє серйозні виклики для її достовірності. Зростання ролі соціальних мереж, месенджерів, платформ зі згенерованим контентом, а також автоматизованих систем поширення формує умови, за яких фейкова, спотворена чи однобока інформація здатна швидко охоплювати широку аудиторію без належної перевірки. Довіра до медіа та інформаційних джерел поступово знижується, оскільки користувачі стикаються не тільки з відверто неправдивими матеріалами, а й з таким контентом, що формально походить від надійних джерел, проте містить викривлення або поданий у хибному контексті.

Однією з основних проблем є те, що в публічному дискурсі та медійному середовищі часто плутають поняття дезінформації та неправдивої інформації. Уряд Великої Британії визначає дезінформацію як «навмисне створення та поширення неправдивої та/або маніпульованої інформації, спрямованої на обман та введення в оману людей з метою заподіяння шкоди або для політичної, особистої чи фінансової вигоди». Це відрізняється від неправдивої інформації, яка визначається, як ненавмисне поширення неправдивої інформації» [3].

Діаграма «Поширення дезінформації» (рис.1.1) показує, як неправдиві відомості поширюються у сучасному інформаційному середовищі. Основним джерелом є спеціальні сайти, що імітують справжні медіа та створюють фейковий контент. Далі ця інформація розходиться через різні канали: традиційні ЗМІ, відомих осіб, соціальні мережі, зашифровані месенджери та навіть особисте спілкування офлайн. Важливо, що поширення відбувається не в одному напрямку,

а у вигляді кола, коли повідомлення багаторазово повторюється та підсилюється. Така мережа робить дезінформацію стійкою та важкою для виявлення, тому протидія їй вимагає системного підходу та постійного контролю різних каналів комунікації.

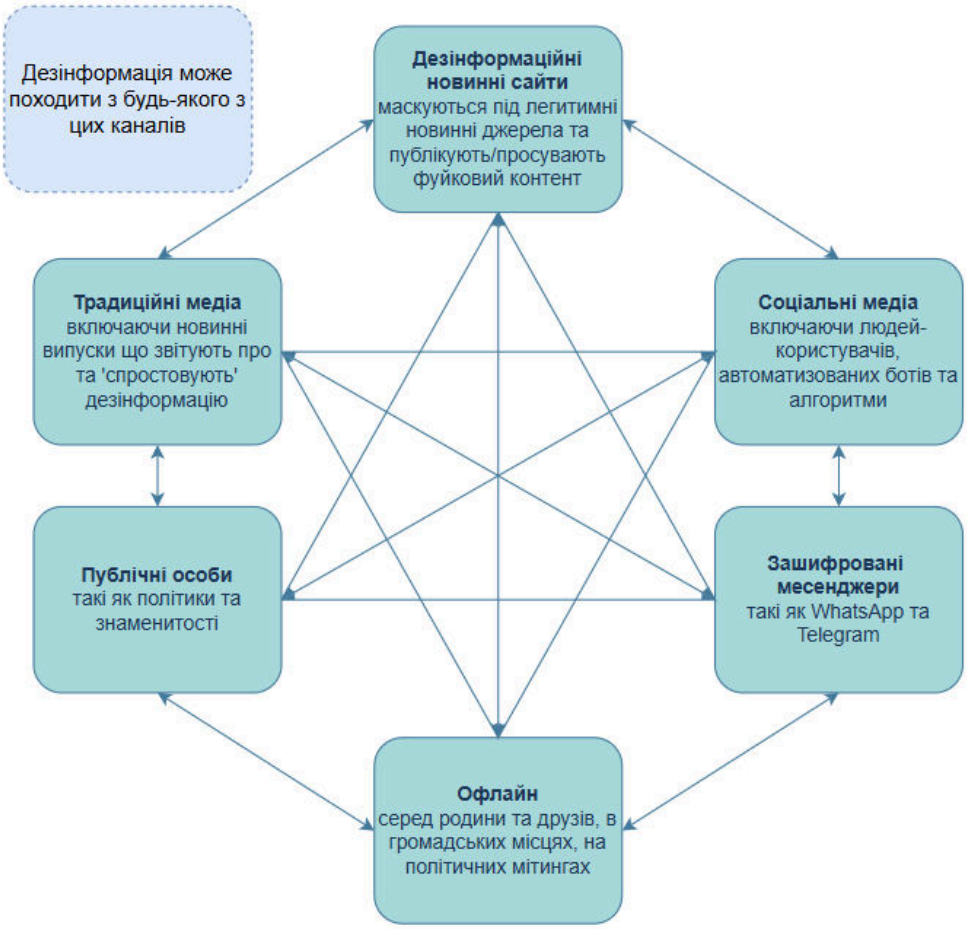


Рис. 1.1. Діаграма поширення дезінформації

Крім того, психологічні та когнітивні фактори відіграють важливу роль у тому, якій інформації довіряють люди. Дослідження показують, що емоційна зарядженість, схильність вірити інформації, яка відповідає вже сформованим переконанням, сприяють тому, що неправдиві твердження легко поширюються та сприймаються як істинні навіть за наявності спростувань. Віра в дезінформацію не лише може призвести до негативних суджень та прийняття рішень, але й має тривалий вплив на міркування людей навіть після того, як її було виправлено – ефект, відомий як ефект постійного впливу [4].

Ще одним елементом кризи довіри є те, що технологічні зміни, особливо пов'язані зі штучним інтелектом та генеративними системами, значно полегшує і прискорює створення контенту з ознаками маніпуляції. *Deepfakes* – фотографії, відео та аудіокліпи, які здаються реальними, але створені інструментами штучного інтелекту, це знижує рівень довіри користувачів до контенту, який вони бачать в Інтернеті. Оскільки контент, створений штучним інтелектом, зростає в обсязі та складності, він може використовуватися зловмисниками для поширення дезінформації та вчинення шахрайства. Соціальні мережі були переповнені таким контентом, що призвело до широкого скептицизму та занепокоєння. У дослідженні *Connected Consumer Study* компанії *Deloitte* за 2024 рік половина респондентів заявили, що вони більш скептично ставляться до точності та надійності онлайн-інформації, ніж рік тому. Серед респондентів, знайомих з генеративним штучним інтелектом або тих, хто його використовує, 68% висловили занепокоєння тим, що синтетичний контент може бути використаний для обману або шахрайства, а 59% повідомили, що їм важко відрізнити медіа, створене людьми, від медіа, створеного штучним інтелектом. Вісімдесят чотири відсотки респондентів, знайомих з поколінням штучного інтелекту, погодилися з тим, що такий контент завжди повинен бути чітко позначений [5]. Такі дані свідчать про зростання невизначеності: навіть коли інформація подається авторитетно, існує підозра, що вона може бути підроблена або змінена.

Узагальнюючи, можна стверджувати, що проблеми достовірності інформації у цифровому середовищі – це не лише питання "фейків" в класичному розумінні, а складний комплекс соціальних, психологічних, технологічних та етичних чинників. І саме у цьому контексті виникає технологія *deepfake*, як один із найбільш радикальних прикладів того, як маніпуляція може переходити з меж "помітного монтажу" до змін, які ледь можна розрізнити з оригіналом. *Deepfake*-технології стають кульмінацією тенденцій, що вже існують: потреба візуальної достовірності, емоційного впливу, швидкого поширення та недовіри до перевірених джерел. Вони створюють новий рівень загрози достовірності, адже кількість і якість маніпуляцій,

здатних обдурити не тільки публіку, а й автоматичні системи перевірки, значно зростає.

1.2. Поняття та класифікація *deepfake*-технологій

У сучасному цифровому середовищі з розвитком штучного інтелекту з'явився феномен *deepfake* – синтетичних мультимедійних матеріалів, створених за допомогою методів глибокого навчання.

Термін «*deepfake*» походить від комбінації слів «*deep*» (що стосується глибокого навчання) та «*fake*» (підробка). Зазвичай він використовується для позначення маніпуляцій з існуючими медіа (зображення, відео та/або аудіо) або створення нових (синтетичних) медіа за допомогою підходів на основі глибокого навчання. Найбільш поширеними прикладами *deepfake*-даних є підроблені зображення обличь, фальсифіковані аудіозаписи мовлення та відео, що поєднують як підроблені зображення, так і фальсифіковану мову. Хоча наявність слова «*fake*» вказує на оброблені або синтезовані медіа, існує безліч корисних застосувань *deepfake*-технології, наприклад, для розваг та творчого мистецтва. У зв'язку з цим було запропоновано інший термін «глибокий синтез» як більш нейтральну альтернативу. Однак цей новий термін не отримав широкого поширення [6].

З наукової точки зору *deepfake* можна визначити, як результат роботи генеративних моделей, що відтворюють реалістичні властивості даних (обличчя, голос, рухи), імітуючи автентичність оригінального контенту. Базовими технологічними підходами до створення *deepfake* є генеративно-змагальні мережі (*GANs*), варіаційні автоенкодери (*VAE*) та їхні модифікації, які забезпечують можливість відтворення високоякісних і правдоподібних мультимедійних матеріалів.

У сучасних дослідженнях виділяють кілька основних підходів до класифікації *deepfake*-технологій: за типом маніпуляції, за модальністю контенту та за рівнем складності.

Таблиця 1.1

Класифікація *deepfake*-технологій

Критерій класифікації	Підкатегорія	Характеристика
За типом маніпуляції	Заміна ідентичності (*Face Swap, Identity Swap*)	Підміна обличчя однієї людини обличчям іншої у відео чи на зображенні
	Маніпуляція атрибутами (*Attribute Manipulation*)	Зміна віку, статі, виразу обличчя, кольору шкіри чи інших характеристик
	Генерація нових обличь (*Face Synthesis*)	Створення абсолютно нових, неіснуючих у реальності облич
	Маніпуляція рухом і мімікою (*Motion Transfer, Expression Swap*)	Перенесення жестів та виразів обличчя з однієї особи на іншу
	Аудіо-*deepfake*	Синтез або трансформація голосу, імітація мовлення конкретної людини
За модальністю контенту	Одномодальні *deepfake*	Використовують лише один тип даних (відео або аудіо)
	Мультимодальні *deepfake*	Поєднують різні модальності (відео + аудіо + текст), підвищуючи реалістичність та складність детекції [*arXiv*]
За рівнем складності	*Shallowfake (cheapfake)*	Прості маніпуляції без глибинних моделей (монтаж, зміна швидкості відтворення)
	Deepfake	Складні підробки на основі глибинних нейронних мереж із високою реалістичністю

Таким чином, поняття *deepfake* охоплює широкий спектр технологій, які варіюються від відносно простих маніпуляцій до високоякісних синтетичних матеріалів, що практично не відрізняються від оригіналу. Класифікація цих технологій є важливим елементом дослідження, адже різні типи підробок потребують застосування різних алгоритмічних і методологічних підходів для їх виявлення.

1.3. Соціальні, етичні та правові наслідки використання *deepfake*

Технології *deepfake* поступово перетворюються на одну з ключових загроз сучасного інформаційного простору. Вони порушують сталі уявлення про достовірність цифрових медіа та підривають довіру до комунікаційних каналів, які ще донедавна вважалися відносно надійними. Соціальний вимір проблеми полягає насамперед у тому, що *deepfake* сприяють поширенню дезінформації та дестабілізації суспільних процесів. Коли аудиторія стикається з контентом, який важко відрізнити від справжнього, формується так звана “криза довіри” – людина починає сумніватися навіть у правдивості перевірених джерел.

Особливо небезпечним є застосування *deepfake* у політичній сфері. Політики є головними мішенями для цієї технології. Діпфейкове відео кандидата, який говорить щось обурливе або пропагує шкідливі ідеї, може швидко поширитися та вплинути на виборців, часто без можливості вчасно це спростувати. Може знадобитися лише один згенерований кліп, щоб зруйнувати кампанію, заплямувати репутацію або змінити хід виборів. Емоційний вплив на кандидатів та їхніх прихильників може бути руйнівним. У міжнародній політиці ставки ще вищі. *Deepfake* може бути використаний для створення дипломатичної кризи або навіть розпалювання конфлікту між країнами. Це більше, ніж просто незручність. Це питання глобальної стабільності. Якщо діпфейки продовжуватимуть розвиватися, вони можуть стати інструментом цифрової війни, підриваючи довіру не лише між окремими особами, а й між цілими країнами [7].

Етична проблема використання технології *deepfake* у сфері заміни акторів полягає насамперед у питанні згоди, авторського права та впливу на професійну діяльність. Коли обличчя або голос актора замінюється штучно згенерованим контентом без його дозволу, це порушує особисті права людини на власне зображення та ідентичність. Такий підхід може використовуватися для створення сцен, у яких актор ніколи не знімався, що вводить глядача в оману й підриває довіру до кіноіндустрії.

Крім того, заміна акторів *deepfake*-технологіями ставить під загрозу професійні можливості артистів. Продюсери можуть віддавати перевагу цифровим копіям, що зменшує потребу в акторській праці та призводить до економічних і соціальних наслідків для творчих професій. Водночас виникає небезпека зловживань – наприклад, створення матеріалів, які дискредитують актора або використовують його образ у неприйнятному контексті.

Таким чином, використання *deepfake* для заміни акторів вимагає чіткого етичного регулювання, прозорості у виробничих процесах та законодавчого захисту особистих і професійних прав митців.

Правові наслідки поширення *deepfake* не менш складні. Чинні законодавчі системи здебільшого не встигають адаптуватися до темпів технологічного розвитку. У багатьох юрисдикціях відсутні чіткі механізми, які б дозволяли ефективно захищати громадян від шкоди, завданої використанням *deepfake*. Існуючі правові бази, хоча й враховують деякі аспекти синтетичних медіа, залишаються недостатніми для регулювання створення та розповсюдження діпфейків. Різні країни застосували різні підходи до регулювання цього питання, але залишається потреба в міжнародній співпраці для встановлення універсальних правових стандартів [8].

Важливою є й проблема відповідальності: хто має відповідати за шкоду, спричинену поширенням *deepfake* – автор алгоритму, користувач, який створив відео, чи платформа, що надала йому доступ до аудиторії? Оскільки поширення таких матеріалів здебільшого відбувається у глобальному цифровому середовищі, питання юрисдикції ускладнює притягнення винних до відповідальності. У низці країн, зокрема у США, Китаї уже запроваджено спеціальні нормативні акти, спрямовані на обмеження шкідливого використання *deepfake*, однак міжнародної єдиної правової бази досі не існує.

У Сполучених Штатах як федеральний так і уряди штатів запровадили заходи для вирішення проблем, пов'язаних з діпфейками. На федеральному рівні одним із важливих кроків є запровадження «Закону про відповідальність за діпфейки», який має на меті вимагати використання цифрових водяних знаків на синтетичних

носіях, щоб вказати, що вони створені штучно, та пропонує правові наслідки за зловмисне використання діпфейків. Європейський Союз вирішує проблему діпфейків у рамках своєї ширшої нормативної бази. Одним з ключових нормативних актів є Загальний регламент про захист даних («*GDPR*»), який може застосовуватися до діпфейків. Згідно з *GDPR*, несанкціоноване створення діпфейків може порушувати закони про конфіденційність та захист даних шляхом обробки персональних даних без їхньої згоди. Ще однією важливою ініціативою є запропонований Закон про цифрові послуги («*DSA*»), метою якого є модернізація правової бази для цифрових послуг. *DSA* прагне притягнути онлайн-платформи до відповідальності за розміщення незаконного контенту, включаючи діпфейки, та вимагає прозорості в практиці модерації контенту. Китай активно намагався регулювати технологію діпфейків, запровадивши комплексні заходи, включаючи створення Адміністрації кіберпростору Китаю («ККП»), яка забезпечує дотримання правил щодо онлайн-аудіовізуальних інформаційних послуг, вимагаючи, щоб синтетичні медіа, включаючи діпфейки, були марковані таким чином, щоб інформувати глядачів про те, що контент є штучним. Ці правила розроблені для запобігання зловживанню діпфейками способами, які можуть завдати шкоди національній безпеці або порушити суспільний порядок. Окрім встановлення цих правил, ККП активно контролює онлайн-платформи для забезпечення їх дотримання [9].

Таким чином, соціальні, етичні та правові наслідки використання *deepfake* мають комплексний характер і взаємопов’язані між собою. Вони одночасно зачіпають довіру в суспільстві, підривають основи міжособистісної та публічної комунікації, порушують базові права людини та ставлять під сумнів ефективність чинних правових механізмів. У цьому контексті особливо важливим є розроблення інтегрованих підходів, що поєднують технологічні інструменти детекції, правове регулювання та етичні стандарти цифрової взаємодії. Лише така багаторівнева стратегія дозволить мінімізувати шкідливі наслідки *deepfake* і водночас зберегти потенціал цієї технології для легітимних та корисних застосувань.

1.4. Сучасні методи створення та поширення *deepfake*-відео

Сучасні підходи до генерації підроблених відеоматеріалів ґрунтуються на швидкому розвитку генеративних моделей штучного інтелекту, котрі дозволяють отримувати фотореалістичні зображення, узгоджені послідовності кадрів та правдоподібну синхронізацію звуку й руху губ. З часом розвиток технологій пройшов шлях від методів, що ґрунтувалися на аналізі та зміні ключових параметрів обличчя, до сучасних рішень на основі глибинних нейронних мереж – зокрема автоенкодерів, *GAN*, а останнім часом і дифузійних моделей, а також їх поєднання з методами нейронного рендерингу. Генерацію діпфейків можна загалом розділити на чотири основні напрямки досліджень:

1) Обмін обличчями – обмін ідентифікаторами між двома зображеннями осіб;

2) Реконструкція обличчя – передачf вихідних рухів та поз;

3) Генерація розмовного обличчя – досягнення природного узгодження рухів роту з текстовим контентом під час генерації персонажів;

4) Редагування атрибутів обличчя – зміна певних атрибутів обличчя цільового зображення [10].

У контексті розвитку технологій важливим є те, що перші інструменти створення *deepfake* ґрунтувалися на автоенкодерах, які навчалися відтворювати риси обличчя однієї людини та переносити їх на іншу. Згодом впровадження генеративно-змагальних мереж забезпечило значно вищий рівень реалістичності, дозволяючи моделі синтезувати узгоджені текстури, освітлення та деталі обличчя. Схема створення *deepfake* з використанням *GAN* зображена на рис.2.2.

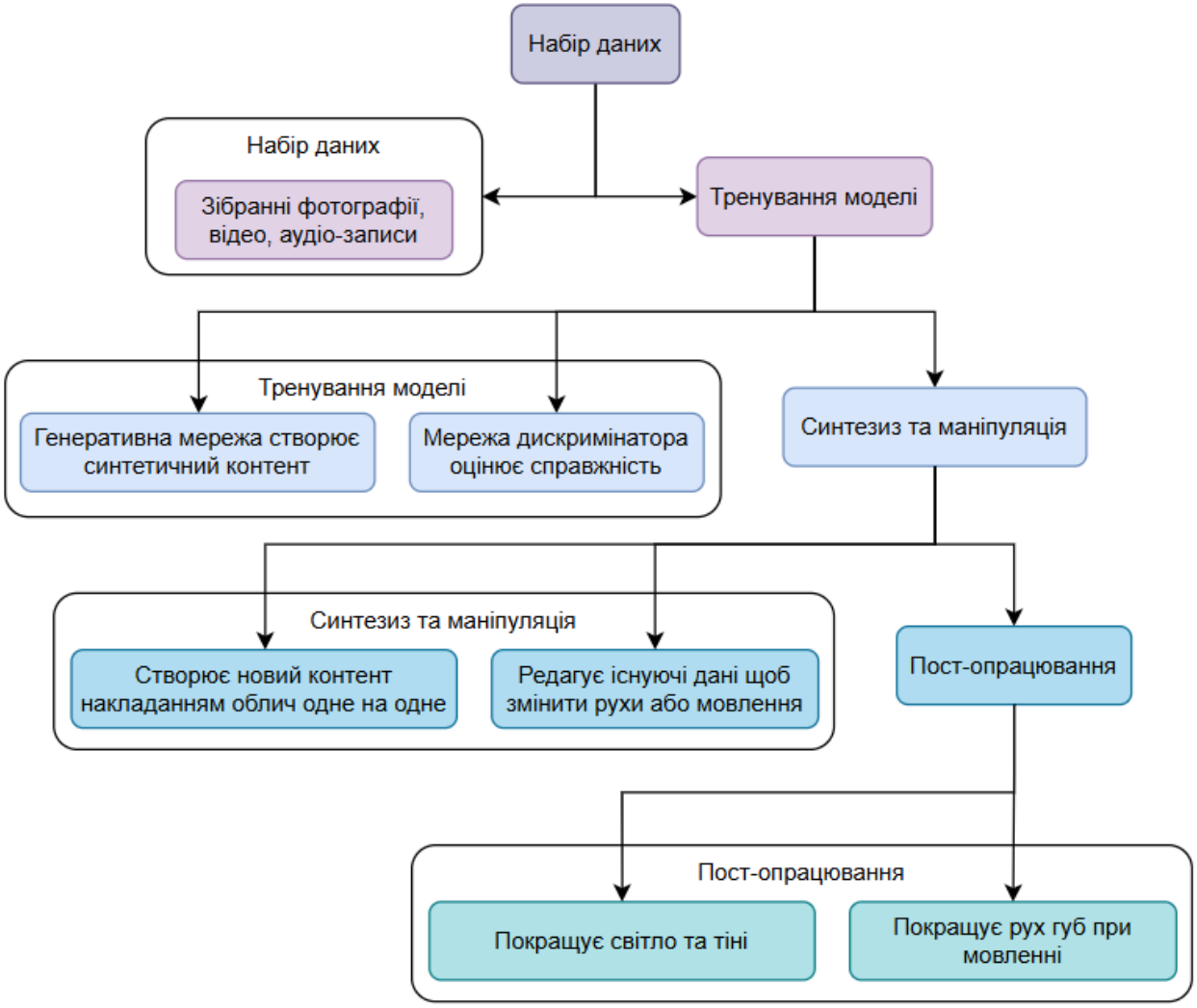


Рис. 1.2. Етапи створення *deepfake*-контенту за допомогою генеративно-змагальних мереж

Подальшим кроком стали дифузійні моделі та архітектури на основі трансформерів, здатні відтворювати не лише статичні обличчя, а й динамічні сцени, що охоплюють як невербальні жести, так і природні мікрорухи. Дифузійні моделі з'явилися як потужне нове сімейство глибоких генеративних моделей з рекордною продуктивністю в багатьох застосуваннях, включаючи синтез зображень, генерацію відео та дизайн молекул [11].

Водночас слід наголосити, що стрімкий прогрес у галузі генеративних моделей поєднується з дедалі більшою доступністю інструментів для кінцевого користувача, що робить процес створення дипфейків не лише сферою професійних досліджень, але й масовим явищем цифрової культури. Програмні засоби на зразок

DeepFaceLab, FaceSwap, а також сервіси на базі *Stable Diffusion* та подібних моделей забезпечують можливість генерування високоякісних відеопідробок навіть користувачами без спеціалізованої технічної підготовки.

DeepFaceLab є одним із найпоширеніших інструментів для створення відео-діпфейків, що спеціалізується на заміні обличь у відео. Програмне забезпечення дозволяє досягати високого рівня реалістичності завдяки широким можливостям налаштування нейронних мереж та підтримці різних алгоритмів тренування. Використовується переважно ентузіастами та дослідниками, оскільки вимагає значних обчислювальних ресурсів і певної технічної підготовки.

FaceSwap – це безкоштовне програмне забезпечення з відкритим кодом, розроблене для виконання базових завдань підміни облич у фото та відео. Завдяки відкритій архітектурі воно активно підтримується спільнотою та сумісне з різними платформами. Попри зручність для новачків, результативність *FaceSwap* поступається більш складним інструментам, особливо в аспекті якості та швидкості обробки.

Stable Diffusion-based сервіси орієнтовані на генерацію графічних зображень та створення простих діпфейків на основі текстових запитів користувачів. Вони доступні у вигляді вебсервісів, що робить їх зручними навіть для людей без технічної підготовки. Хоча ці системи ефективно справляються із завданнями синтезу зображень і можуть використовуватись для візуальних маніпуляцій, можливості для створення відео-діпфейків у них поки що обмежені.

Zao – це мобільний додаток китайського походження, який здобув популярність завдяки надзвичайно швидкому та простому процесу створення відео з підміною облич. Достатньо завантажити власне фото, щоб за кілька секунд отримати відеокліп із реалістично інтегрованим обличчям. Його ключова перевага полягає у зручності використання для широкої аудиторії, проте він обмежений у гнучкості налаштувань та викликає занепокоєння щодо конфіденційності даних.

Таблиця 1.2

Переваги та недоліки програмних засобів для створення *deepfake*

Назва	Переваги	Недоліки
DeepFaceLab	Висока якість підробок; розширені можливості налаштування; велика база навчальних матеріалів	Складність освоєння; потреба у потужному обладнанні
FaceSwap	Безкоштовність і відкритий код; кросплатформеність; зручність для новачків	Нижча якість і швидкість; обмежені можливості порівняно з професійними
Stable Diffusion-сервіси	Простота використання; доступність через онлайн-платформи; швидке генерування зображень	Обмеженість у створенні відео; ризики зловживання та етичні питання
Zao	Миттєвий результат; орієнтованість на масового користувача; мінімальні вимоги до навичок	Низька гнучкість; ризики для приватності; залежність від серверної обробки

Поширення *deepfake*-відео відбувається переважно через соціальні мережі, стрімінгові сервіси та месенджери, що характеризуються високою швидкістю циркуляції інформації, вірусністю контенту та слабкими механізмами превентивного контролю. Це призводить до ситуації, коли навіть локально створені підробки здатні упродовж лічених годин набути глобального резонансу та стати частиною інформаційних маніпуляцій. У такий спосіб сучасні технології генерації дипфейків необхідно розглядати не лише в контексті технічних інновацій у сфері штучного інтелекту, але й як явище, що змінює характер функціонування інформаційного простору, ускладнює забезпечення кібербезпеки та створює нові виклики для правового регулювання й етичних стандартів.

1.5. Труднощі автоматичного виявлення маніпуляцій у відеоконтенті

Виявлення *deepfake*-контенту почали досліджувати ще з моменту його першої появи. Класичні підходи до виявлення підробок головним чином зосереджуються на внутрішній статистиці та спеціально визначених ознаках, таких як моргання

очей, положення голови та візуальні артефакти, щоб проаналізувати просторові шаблони маніпуляції зображенням [12]. Еволюція генеративних моделей призвела до того, що синтетичні зображення й відео стали більш реалістичними: зникли характерні візуальні артефакти, а статистичні властивості кадрів наблизилися до природних. Унаслідок цього класичні методи виявлення, які покладаються на пошук однакових ознак чи помітних слідів рендерингу, поступово втрачають ефективність. Автоматичне розпізнавання підробок у відео стикається з багаторівневими технічними й методологічними викликами, що значно ускладнює застосування таких систем у практиці:

1) Важливим фактором є специфіка поширення відео в інтернеті: стиснення, перекодування, фільтрація платформ і зниження роздільної здатності спотворюють або затирають найдрібніші сліди маніпуляцій, на які зазвичай орієнтуються алгоритми. Моделі, навчені на якісних даних, демонструють суттєве падіння точності при роботі з відео із соціальних мереж. Таким чином, виявлення маніпуляцій після обробки платформами потребує окремих методик і ретельної підготовки даних.

2) Обмеження, пов’язані з датасетами та методологією оцінювання. Хоча з’явилися великі набори даних для тренування (наприклад, *FaceForensics++* та *DFDC*), вони охоплюють лише частину можливих сценаріїв – зокрема певні методи генерації, типові ракурси, освітлення та демографічні характеристики; це породжує ризик упередження моделей та їхньої непридатності до нових умов. Більш того, стандартизовані метрики (наприклад, *AUC* чи *accuracy*) не завжди корелюють із реальними потребами – наприклад, чутливість до визначення діпфейку у журналістському чи судовому контексті може мати зовсім інші вимоги, ніж верифікація контенту у соціальних мережах. У зв’язку з цим наукова спільнота підкреслює необхідність розробки більш репрезентативних бенчмарків, складніших метрик та тестових сценаріїв, здатних оцінювати стійкість моделей до компресії, постобробки та нових генеративних методів.

3) Переконливі підробки часто комбінують відео й аудіо, а також можуть містити супутні метадані й контекстні підказки (наприклад, підписи, текст у кадрі),

тому ефективна система виявлення має оперувати не лише просторовими ознаками окремого кадру, але й тимчасовими патернами (послідовність мікро-рухів, синхронність артикуляції та спектральні особливості голосу) та міжмодальними невідповідностями. Багатомодальні детектори обіцяють підвищену стабільність, проте вони складніші в реалізації, вимагають великих узгоджених датасетів і є більш чутливими до технічних похибок, що створює додаткові інженерні й дослідницькі бар'єри.

Крім того, високий рівень хибнопозитивних спрацьовувань може шкодити як користувачам, так і платформам, тоді як занадто «обережні» алгоритми здатні пропускати небезпечний контент. Відсутність прозорості у поясненні рішень також знижує довіру фахівців і ускладнює використання таких результатів у правовому полі.

У підсумку можна стверджувати, що проблеми детекції мають системний характер і виходять за межі суто технічних питань. Вони охоплюють динаміку розвитку контенту, алгоритмічні й економічні механізми поширення, а також етичні та юридичні аспекти, що вимагає комплексного підходу до створення надійних і масштабованих систем виявлення маніпуляцій у відео.

1.6. Існуючі підходи та алгоритми виявлення *deepfake*

Виявлення підробленого візуального контенту у вигляді *deepfake* стало одним із ключових напрямів сучасних досліджень у сфері інформаційної безпеки. Методи, що застосовуються для цієї мети, умовно можна поділити на дві групи: традиційні підходи, які орієнтуються на аналіз явних артефактів та фізіологічних особливостей, і сучасні методи, що ґрунтуються на використанні моделей глибинного навчання та мультимодальних алгоритмів.

Традиційні методи виявлення *deepfake* спираються на пошук низькорівневих ознак і невідповідностей у візуальному чи аудіоряді. Серед них виділяють аналіз частоти кліпання очей, мікрорухів м'язів обличчя, асиметрії в освітленні та тінях, а також статистичних особливостей окремих пікселів чи спектрів. Такі підходи

мають відносну простоту реалізації та потребують менше обчислювальних ресурсів. Водночас вони мають низьку стійкість до вдосконалених генеративних моделей: із розвитком технологій штучного інтелекту більшість візуальних артефактів стали малопомітними, що значно знижує ефективність класичних методів у реальних умовах, особливо при роботі зі стисненим відео або після обробки платформами.

Методи на основі штучного інтелекту базуються переважно на використанні глибинних нейронних мереж, зокрема згорткових (*CNN*), рекурентних (*RNN*), генеративно-змагальних мереж (*GAN*) та дифузійних моделей. Такі алгоритми здатні автоматично виявляти складні патерни у зображеннях і відео, які є недоступними для традиційного аналізу. Вони враховують просторово-часові особливості, наприклад синхронність артикуляції та голосу, плавність рухів обличчя, стабільність текстур шкіри. В рамках роботи *Intel* над відповідальним штучним інтелектом компанія випустила технологію *FakeCatcher*, яка здатна виявляти підроблені відео з точністю до 96%. Коли наше серце перекачує кров, то капіляри змінюють колір. Ці сигнали кровотоку збираються з усього обличчя, і алгоритми переводять ці сигнали в просторово-часові карти [13]. Потім, використовуючи глибоке навчання *FakrCatcher* миттєво розуміє чи це діпфейк.

Окремо виділяють мультимодальні підходи, які поєднують візуальні й аудіо сигнали, що дозволяє знижувати рівень помилкових результатів. Основними перевагами таких методів є висока точність і здатність адаптуватися до нових типів маніпуляцій. Проте їхнє впровадження у практику ускладнюють потреба у великих і різноманітних наборах даних для тренування, значні обчислювальні витрати та ризик перенавчання на вузьких вибірках.

Для систематизації основних характеристик традиційних та ШІ-орієнтованих методів виявлення *deepfake* доцільно представити їх порівняння у табличній формі (табл. 1.3).

Таблиця 1.3

Порівняння підходів виявлення технології *deepfake*

Підхід	Ознаки, на які спирається	Сильні сторони	Слабкі сторони
Традиційні методи	Артефакти рендерингу, мікрорухи, освітлення, статистика пікселів	Простота реалізації, невеликі обчислювальні витрати, інтерпретованість результатів	Низька стійкість до сучасних генеративних моделей, залежність від якості відео
Методи на основі ШІ	Просторово-часові патерни, текстури, синхронність голосу та рухів, мультимодальні ознаки	Висока точність, здатність до узагальнення, ефективність для нових сценаріїв	Потреба у великих датасетах, значні обчислювальні ресурси, складність інтерпретації результатів

У зв’язку з цим дослідники наголошують на необхідності комплексного підходу до побудови систем виявлення. Це передбачає інтеграцію різних типів ознак (піксельних, фізіологічних, часових), використання адаптивного навчання, впровадження стратегій захисту від адверсаріальних атак, а також розробку стандартизованих протоколів оцінювання, які враховують реальні умови поширення контенту.

1.7. Обмеження та виклики сучасних систем детекції

Незважаючи на інтенсивний розвиток технологій автоматичного виявлення маніпуляцій у відеоконтенті, сучасні системи детекції *deepfake* стикаються з низкою принципових обмежень. Однією з ключових проблем є узагальнюваність результатів: моделі, що досягають високих показників точності на контрольованих наборах даних, таких як *FaceForensics++* або *DFDC*, часто демонструють значне падіння ефективності при зіткненні з новими прикладами, створеними іншими методами чи за відмінних умов освітлення та стиснення. Це свідчить про залежність алгоритмів від специфіки навчальних датасетів і слабку здатність до переносу знань на «дикі» сценарії. До того ж, значну загрозу становлять

адверсаріальні атаки, коли штучно внесені малопомітні зміни дозволяють обійти навіть найсучасніші моделі, що підкреслює вразливість існуючих підходів.

Ще одним серйозним викликом є обмеженість наявних датасетів, які використовуються для тренування та тестування систем. Більшість з них містять відносно прості приклади підробок, створені подібними генеративними моделями, що не відображає реальної різноманітності медіа-середовища. У результаті показники точності, отримані в лабораторних умовах, є завищеними та не відповідають реальній ефективності у відкритих сценаріях використання . Крім того, у випадках, коли відео поширюється через соціальні мережі та піддається багатократному стисненню, артефакти стають ще менш помітними, що додатково ускладнює роботу алгоритмів. Таким чином, постає необхідність створення більш різноманітних і репрезентативних наборів даних, які б відображали повний спектр можливих підробок.

Не менш важливою проблемою є висока ресурсомісткість моделей і низький рівень інтерпретованості їхніх рішень. Сучасні алгоритми виявлення базуються на глибоких нейронних мережах із мільйонами параметрів, що потребує значних обчислювальних потужностей і робить їх непридатними для використання на пристроях із обмеженими ресурсами чи у режимі реального часу. Водночас у практичних сферах, таких як судочинство чи журналістика, важливим є не лише саме визначення «оригінал/підробка», а й пояснення, на основі яких характеристик алгоритм зробив свій висновок. Нестача зрозумілості знижує довіру до результатів і обмежує можливість офіційного застосування систем у критично важливих контекстах . Подолання цих обмежень вимагає не лише технічних інновацій у сфері алгоритмів, але й розробки комплексних стратегій, які включатимуть нові методи тренування, підвищення стійкості до атак та інтеграцію принципів.

1.8. Висновки до розділу

Проведений аналіз показав, що проблема достовірності цифрового контенту є однією з ключових викликів сучасного інформаційного суспільства. Масштабне

поширення соціальних мереж і платформ для обміну медіафайлами створює умови, за яких недостовірні дані можуть швидко набувати глобального характеру, впливаючи на суспільну думку та навіть політичні процеси. У цьому контексті феномен *deepfake* виступає яскравим прикладом нової хвилі цифрових загроз, що поєднують у собі високий рівень технологічності та потенційну небезпеку для суспільної безпеки й особистих прав громадян.

Детальний розгляд поняття та класифікації *deepfake*-технологій дозволив визначити основні напрями їхнього розвитку та застосування. Незважаючи на наявність потенційно корисних сфер використання, таких як освіта, мистецтво чи розважальна індустрія, найбільший суспільний резонанс спричиняють саме випадки зловживання цими технологіями. Соціальні, етичні та правові наслідки включають порушення права на приватність, поширення дезінформації, шантаж та маніпуляцію громадською свідомістю. Це робить проблему детекції підробленого контенту не лише технічною, а й соціальною та правовою, що вимагає комплексного підходу.

Разом із тим, аналіз сучасних методів створення *deepfake*-відео дозволив виявити, що технології генерації стають дедалі більш досконалими: використання генеративно-змагальних мереж, дифузійних моделей та нейронного рендерингу забезпечує надзвичайний рівень реалістичності підробок. Хоча сучасні алгоритми детекції пропонують різні, кожен із них має власні обмеження та вразливості. Коло проблем, що потребують подальших наукових досліджень окреслюють виклики, пов’язані з узагальнюваністю моделей, їхньою стійкістю до атак, потребою в масштабних обчислювальних ресурсах і низьким рівнем інтерпретованості.

Таким чином, у розділі обґрунтовано, що питання виявлення *deepfake* є багатогранним і вимагає не лише вдосконалення алгоритмічних рішень, а й створення нормативно-правових механізмів, міждисциплінарної взаємодії та формування суспільної обізнаності.

РОЗДІЛ 2
ПРОЄКТУВАННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ВИЯВЛЕННЯ *DEEPFAKES* У ВІДЕО

2.1. Аналіз існуючих систем

У сучасному світі активно розробляються та впроваджуються різноманітні системи для автоматичного виявлення *deepfake*-контенту. Ці рішення відрізняються за архітектурою, методологією, сферами застосування та рівнем доступності. Деякі з них орієнтовані на академічні дослідження і забезпечують високу точність в контрольованих умовах, інші призначені для комерційного використання та інтегруються у платформи соціальних мереж чи медіа-компанії. Розглянемо три ключові приклади існуючих систем детекції та їхні характерні особливості, зосередившись на принципах їхньої роботи з відеоматеріалами.

Microsoft Video Authenticator – це комерційне рішення, розроблене корпорацією *Microsoft*. Може аналізувати фотографію або відео, щоб визначити відсоток ймовірності або показник достовірності того, що медіафайл штучно маніпульований [14]. Система призначена для використання медіа-організаціями, політичними кампаніями та іншими установами, що потребують швидкої верифікації контенту.

Принцип роботи системи базується на покадровому аналізі відеоматеріалу з використанням глибоких нейронних мереж, навчених на великих наборах як автентичних, так і маніпульованих зображень. При обробці відео система послідовно аналізує кожен кадр, виявляючи ледь помітні відхилення у градієнтах яскравості, межах об'єктів, текстурах шкіри та інших візуальних характеристиках, які можуть свідчити про цифрові зміни. Алгоритм звертає особливу увагу на області навколо обличчя, де найчастіше виникають артефакти через складність

відтворення природних рухів м'язів, мікроекспресій та взаємодії зі світлом. Система також аналізує тимчасову узгодженість між послідовними кадрами, оскільки *deepfake*-алгоритми часто створюють невеликі нестиковки у русі або освітленні, які стають помітними при детальному покадровому порівнянні. Для фотографій застосовується схожий підхід, але з акцентом на статичні артефакти та аномалії у частотних характеристиках зображення.

Intel FakeCatcher – це перший у світі детектор діпфейків у реальному часі, який повертає результати за мілісекунди, що дозволяє визначити підроблені відео з точністю до 96%. Технологія використовує апаратне та програмне забезпечення Intel, працює на сервері та взаємодіє через вебплатформу [15]. На відміну від більшості існуючих рішень, що фокусуються на пошуку візуальних артефактів, *FakeCatcher* використовує принципово інший підхід, заснований на аналізі біологічних сигналів людського організму. Може бути придатною для інтеграції у платформи прямих трансляцій та системи відеоконференцій.

Принцип роботи *Intel FakeCatcher* базується на аналізі фотоплетизмографічних сигналів (*PPG*) – змін кольору шкіри обличчя, спричинених пульсацією крові у капілярах під час серцевих скорочень. Коли серце перекачує кров через судини, гемоглобін поглинає світло, що призводить до мікроскопічних змін відтінку шкіри, непомітних неозброєним оком, але вловлюваних сучасними сенсорами та алгоритмами обробки зображень. Система використовує просторово-часовий аналіз відео для побудови карт розподілу *PPG*-сигналів по всій поверхні обличчя. У справжньому відео ці сигнали мають характерну ритмічну структуру, синхронізовану з частотою серцевих скорочень, та специфічний просторовий розподіл, що відповідає анатомії судинної системи обличчя. У *deepfake*-відео, незалежно від якості візуальної підробки, відтворення цих складних біологічних патернів є надзвичайно складним завданням, оскільки генеративні моделі зазвичай не моделюють такі тонкі фізіологічні процеси. Алгоритм аналізує як амплітуду, так і фазу *PPG*-сигналів, їхню просторову узгодженість та відповідність типовим характеристикам живої тканини.

Sensity AI – це передова платформа для виявлення дипфейків, яка ідентифікує медіа, маніпульовані штучним інтелектом, для боротьби з діпфейками та синтетичним шахрайством, захищаючи від таких загроз, як шахрайство з особистими даними, видавання себе за знаменитостей та обхід перевірки особи[16]. Система орієнтована на великі організації, урядові установи та медіа-компанії, що потребують постійного контролю за поширенням потенційно шкідливого контенту у цифровому просторі. На відміну від точкових рішень для перевірки окремих файлів, *Sensity AI* пропонує екосистемний підхід до проблеми детекції *deepfake*, поєднуючи автоматизований моніторинг, аналітику та інструменти реагування.

Принцип роботи *Sensity AI* базується на багатомодальному аналізі, що поєднує обробку візуальної інформації, аудіодоріжок та супутніх метаданих для досягнення максимальної точності виявлення. При аналізі відео система одночасно застосовує декілька спеціалізованих нейронних мереж: перша аналізує просторові характеристики кожного кадру, шукаючи візуальні артефакти, характерні для різних методів генерації *deepfake* (*GAN*-артефакти, проблеми з краями об'єктів, аномалії у текстурах); друга мережа досліджує тимчасові патерни, виявляючи нестиковки у послідовності рухів, мікровібрації та іншу динаміку, що може свідчити про маніпуляції; третя компонента аналізує аудіодоріжку, перевіряючи синхронізацію руху губ з мовленням, виявляючи артефакти синтезованого голосу та аномалії у спектральних характеристиках звуку. Окремий модуль обробляє метадані файлу, історію його поширення у соціальних мережах, параметри кодування та інші технічні характеристики, які можуть надати додаткові підказки про автентичність. Система використовує ансамблевий підхід, де рішення приймається на основі зваженого голосування всіх компонент, що значно знижує ймовірність помилкових спрацьовувань. Для фотографій застосовується адаптована версія алгоритму з акцентом на просторовий аналіз та метадані.

Таблиця 2.1

Порівняльна характеристика систем виявлення *deepfake*

Система	Основні переваги	Основні недоліки	Область застосування
Microsoft Video Authenticator	Висока точність на відомих типах *deepfake*; інтеграція з екосистемою *Microsoft*; зручний інтерфейс; покадровий аналіз відео	Закрите рішення без можливості модифікації; потребує хмарної інфраструктури; зниження ефективності на стисненому контенті; обмежена гнучкість налаштувань	Корпоративне середовище; медіа-організації; політичні кампанії; урядові установи
Intel FakeCatcher	Унікальний підхід на основі *PPG*-сигналів; висока точність (96%); робота в реальному часі; стійкість до стиснення; складність обходу системи	Потребує високоякісного відео з чітким обличчям; обмежена ефективність при поганому освітленні; не підходить для відео низької якості або з великої відстані	Верифікація відео в реальному часі; платформи прямих трансляцій; відеоконференції; медіа-платформи; системи безпеки
Sensity AI	Багатомодальний аналіз (відео, аудіо, метадані); моніторинг у режимі реального часу; висока масштабованість; детальна аналітика та звітність; ансамблеві методи	Висока вартість підписки; складність початкового налаштування та інтеграції; залежність від зовнішніх *API*; можливі затримки у обробці	Моніторинг соціальних мереж; корпоративна безпека; великі організації; боротьба з дезінформацією; урядові агенції

Аналіз трьох провідних систем виявлення *deepfake* демонструє різноманітність технологічних підходів та їхню адаптацію до специфічних потреб користувачів. *Microsoft Video Authenticator* представляє традиційний підхід, що базується на виявленні візуальних артефактів через покадровий аналіз нейронними мережами, забезпечуючи надійне рішення для корпоративного середовища з акцентом на інтеграцію з існуючою екосистемою продуктів. *Intel FakeCatcher* демонструє інноваційний напрямок розвитку технологій детекції, використовуючи

фундаментальні біологічні характеристики, які складно імітувати штучно, що робить цю систему особливо перспективною у контексті постійної еволюції генеративних моделей. *Sensity AI*, у свою чергу, пропонує найбільш комплексний підхід, поєднуючи множинні модальності аналізу та орієнтуючись на широкомасштабний моніторинг цифрового простору.

Важливо зазначити, що жодна з систем не забезпечує абсолютної гарантії виявлення всіх типів *deepfake*, особливо тих, що створені з використанням найновіших генеративних моделей або спеціально адаптовані для обходу конкретних алгоритмів детекції. Це підкреслює необхідність комплексного підходу до верифікації контенту, що поєднує автоматизовані системи з експертною оцінкою та критичним мисленням.

2.2. Загальний опис системи

На основі проведеного аналізу існуючих систем виявлення *deepfake* та виявлених обмежень було сформовано концепцію нової інтелектуальної системи. Процес формування архітектури представлено у вигляді інтелект-карти (рис. 2.1).

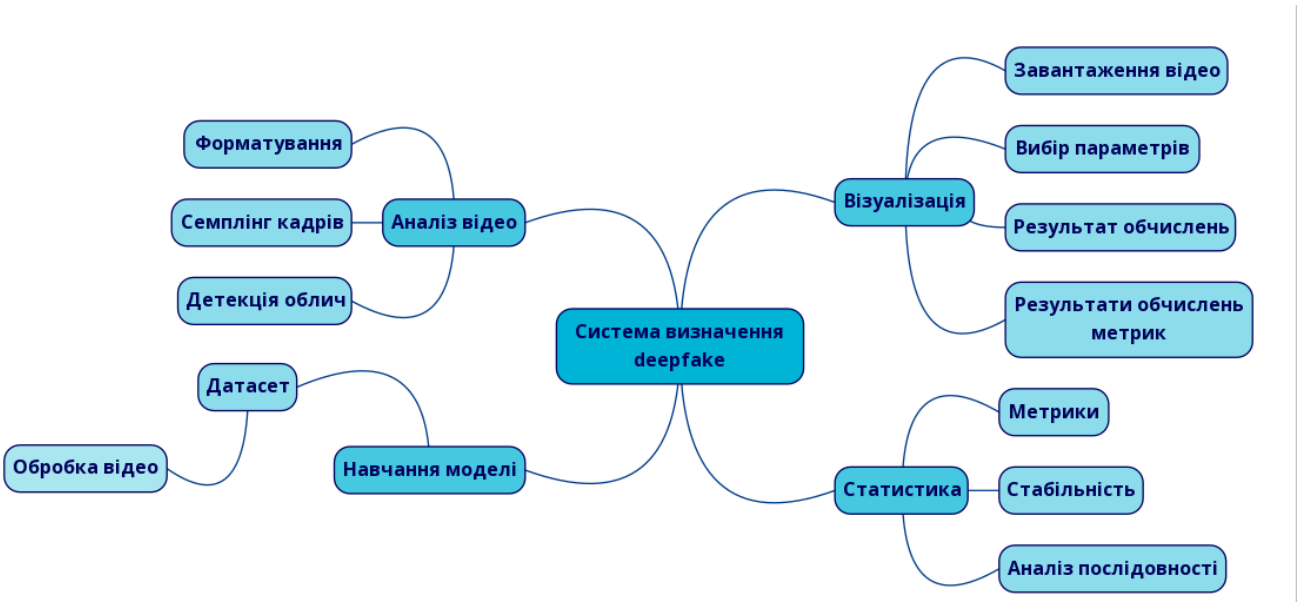


Рис. 2.1. *Mind-map* проєктування системи виявлення *deepfake*

Розроблена система призначена для автоматизованого виявлення синтетичних відеоматеріалів, створених за допомогою технологій глибинного навчання. Система здійснює комплексний аналіз відеопослідовностей та надає обґрунтований висновок щодо автентичності контенту з детальною візуалізацією результатів.

Робота системи починається з завантаження відеофайлу, який може бути у різних форматах, таких як *MP4, AVI, MOV* та інших поширених відеоформатах. Після завантаження система автоматично витягує з відео певну кількість кадрів для подальшого аналізу, рівномірно розподіляючи їх по всій тривалості відеоматеріалу. Це дозволяє отримати репрезентативну вибірку, яка охоплює весь відеоконтент без необхідності обробки кожного кадру, що значно прискорює процес аналізу.

На кожному витягнутому кадрі система виконує пошук та виділення зон, де присутні обличчя людей. Для цього використовуються спеціалізовані алгоритми детекції, які можуть розпізнавати обличчя навіть при складних умовах освітлення, різних кутах зйомки та частковому перекритті об'єктами. Після виявлення обличчя система вирізає його з кадру та підготовляє для подальшого аналізу, при цьому враховуючи певний відступ навколо обличчя для збереження контексту.

Кожне виділене зображення обличчя піддається аналізу за допомогою навченої нейронної мережі, яка визначає ймовірність того, що дане обличчя є синтетичним або маніпульованим. Нейронна мережа базується на сучасних архітектурах глибинного навчання та була навчена на великому датасеті реальних та підроблених зображень облич. Для кожного кадру система видає числову оцінку у діапазоні від нуля до одиниці, де значення ближче до нуля вказує на високу ймовірність автентичності, а значення ближче до одиниці свідчить про можливу підробку.

Ключовою особливістю розробленої системи є додатковий етап аналізу, який оцінює не лише окремі кадри, але й зміну результатів класифікації в часі. Цей підхід базується на спостереженні, що підроблені відео часто демонструють нестабільність у результатах аналізу послідовних кадрів, тоді як справжні відео характеризуються більш консистентними показниками. Система обчислює

різноманітні статистичні характеристики отриманої послідовності оцінок, включаючи міру їхньої варіативності, частоту різких змін між сусідніми кадрами, рівень передбачуваності послідовності та загальну плавність змін показників.

На основі покадрового аналізу та статистичних метрик система формує фінальне рішення щодо автентичності відео. При цьому враховується не лише середнє значення оцінок по всіх кадрах, але й показники стабільності цих оцінок у часі. Якщо система виявляє високу варіативність результатів між кадрами або різкі стрибки в оцінках, це додатково підвищує ймовірність того, що відео є підробленим, навіть якщо окремі кадри мають невисокі показники підозрілості.

Результати аналізу представляються користувачу через зручний вебінтерфейс, який не потребує спеціальних технічних знань для використання. Інтерфейс відображає обрані ключові кадри з відео, на яких виділені виявлені обличчя та вказані оцінки для кожного кадру. Система також генерує графіки, які показують зміну оцінок протягом всього відео, що дозволяє візуально оцінити стабільність результатів. Фінальне рішення представлене у вигляді зрозумілого висновку з категоріями "справжнє", "підроблене" або "невизначене", якщо система не може впевнено класифікувати відео.

Система розроблена з використанням сучасних технологій машинного навчання та комп'ютерного зору, що забезпечує високу точність детекції при прийнятній швидкості обробки. Модульна структура програмного забезпечення дозволяє легко налаштовувати параметри аналізу та адаптувати систему під різні вимоги та умови використання. Всі налаштування централізовані та можуть бути змінені без необхідності втручання в основний програмний код, що спрощує підтримку та розвиток системи.

2.3. Аналіз функціональних та нефункціональних вимог

Для забезпечення ефективної роботи системи виявлення *deepfake* необхідно чітко визначити вимоги, яким вона має відповідати. Вимоги поділяються на дві основні категорії: функціональні, які описують що саме система повинна робити,

та нефункціональні, які визначають як система має це робити. Розглянемо детально кожну з цих категорій.

Функціональні вимоги визначають конкретні функції та можливості, які повинна виконувати система для забезпечення виявлення *deepfake* відеоматеріалів.

Система повинна забезпечувати завантаження відеофайлів різних форматів, включаючи *MP4, AVI, MOV, MKV* та *WebM*. При цьому система має коректно обробляти відео з різними кодеками та роздільними здатностями. Максимальна тривалість відео для аналізу визначена як 60 секунд для забезпечення прийнятної швидкості обробки.

Система повинна автоматично виявляти обличчя на кадрах відео з мінімальним розміром 50 пікселів. Детекція має працювати при різних умовах освітлення та кутах повороту обличчя. При виявленні кількох обличь система обирає найбільше для подальшого аналізу.

Кожне виявлене обличчя повинно класифікуватися як реальне або синтетичне з наданням числової оцінки ймовірності підробки від 0 до 1. Класифікація враховує характерні артефакти *deepfake*-генерації, такі як нереалістичне освітлення, нечіткість меж та візуальні аномалії.

Система повинна здійснювати статистичний аналіз послідовності оцінок для виявлення темпоральних аномалій. Аналіз включає обчислення дисперсії, детекцію різких стрибків між кадрами, визначення ентропії та оцінку автокореляції. На основі цих метрик формується інтегральний показник стабільності.

Фінальне рішення формується на основі покадрових оцінок та показників темпоральної стабільності. Система класифікує відео як справжнє, підроблене або невизначене, надаючи числові показники впевненості.

Система повинна надавати візуалізацію результатів через вебінтерфейс, включаючи ключові кадри з виділеними обличчями, часовий графік оцінок, гістограму розподілу та панель статистичних метрик.

Система має підтримувати можливість налаштування порогових значень та параметрів аналізу через конфігураційний файл без зміни програмного коду.

Система повинна забезпечувати аутентифікацію користувачів з різними рівнями доступу для захисту від несанкціонованого використання та розмежування прав адміністраторів і звичайних користувачів.

Система має підтримувати *API* для інтеграції з іншими програмними продуктами та сервісами, дозволяючи автоматизувати процес верифікації відеоконтенту в сторонніх системах.

Система повинна генерувати детальні звіти з результатами аналізу у форматах *PDF* та *JSON* з можливістю експорту всіх метрик, візуалізацій та метаданих відео для архівування та документування.

Система має підтримувати пакетну обробку кількох відеофайлів з формуванням зведеного звіту та можливістю встановлення пріоритетів обробки для різних файлів.

Система повинна надавати можливість порівняльного аналізу декількох відео одночасно з візуалізацією спільних та відмінних характеристик для виявлення серій пов'язаних *deepfakes*.

Нефункціональні вимоги визначають якісні характеристики роботи системи, які не стосуються безпосередньо її функціональності, але є важливими для практичного використання.

Продуктивність системи має забезпечувати обробку відео тривалістю до 30 секунд не більше ніж за 2 хвилини на апаратному забезпеченні з *GPU*. Без *GPU* час обробки не повинен перевищувати 5 хвилин. Система має ефективно використовувати обчислювальні ресурси та пам'ять.

Точність детекції повинна становити не менше 90 відсотків на тестовому датасеті, з *precision* не нижче 88 відсотків та *recall* не нижче 92 відсотків.

Надійність системи передбачає стабільну роботу протягом тривалого часу без збоїв. Система повинна коректно обробляти помилкові ситуації, такі як пошкоджені відеофайли або відсутність облич на кадрах, з видачею зрозумілих повідомлень про помилки.

Масштабованість системи повинна дозволяти розгортання як на локальному комп'ютері, так і на серверній інфраструктурі. Архітектура має підтримувати

горизонтальне масштабування для обробки великої кількості запитів одночасно при розгортанні в хмарному середовищі.

Зручність використання передбачає інтуїтивно зрозумілий інтерфейс, який не потребує спеціальної технічної підготовки. Час навчання роботі з системою для нового користувача не повинен перевищувати 15 хвилин.

Сумісність системи має забезпечувати роботу на операційних системах *Windows, Linux та macOS*. Система повинна підтримувати сучасні веббраузери, включаючи *Chrome, Firefox, Safari* та *Edge* останніх версій.

Безпека системи передбачає захист завантажених відеофайлів та результатів аналізу від несанкціонованого доступу. Всі дані користувачів мають зберігатися з шифруванням, а передача даних здійснюватися через захищені протоколи.

Підтримуваність коду забезпечується модульною архітектурою, детальною документацією та дотриманням стандартів кодування *Python*. Код має бути структурованим з чіткими коментарями для спрощення подальшого розвитку та підтримки системи.

Для ефективної реалізації функціональних та нефункціональних вимог до системи виявлення *deepfake* доцільно використати метод *MoSCoW*. Даний підхід дозволяє чітко визначити пріоритетність впровадження різних вимог проєкту, враховуючи їх важливість для досягнення основної мети системи.

Таблиця 2.2

Пріоритетність вимог

Статус	Вимоги
Must have (Обов'язкові)	Завантаження відеофайлів різних форматів Автоматична детекція облич на кадрах Класифікація облич на *real/fake* Статистичний аналіз темпоральної послідовності Формування фінального рішення з категоріями
Should have (Важливі)	Налаштування параметрів через конфігурацію Візуалізація результатів через вебінтерфейс Модульна архітектура та документація коду Підтримка сучасних веббраузерів Інтуїтивно зрозумілий інтерфейс Висока точність класифікації

Закінчення табл. 2.2

Could have (Бажані)	Пакетна обробка кількох відео Сумісність з різними операційними системами Коректна обробка помилкових ситуацій Час обробки до 2 хвилин для 30-секундного відео
Won't have (Не плануються на даному етапі)	Генерація детальних звітів у *PDF/JSON* Аутентифікація користувачів API для інтеграції з іншими системами Порівняльний аналіз декількох відео Шифрування даних користувачів Горизонтальне масштабування для хмари

Такий метод дозволяє чітко визначити, які вимоги є критично важливими для першої версії системи, а які можуть бути реалізовані на пізніших етапах розробки. Це дає змогу зосередитись на ключових функціях, необхідних для запуску мінімально життєздатного продукту, що забезпечує високу точність виявлення *deepfake* та можливість інтерпретації результатів. Поступове розширення функціональності згідно з пріоритетами дозволить адаптувати систему до конкретних потреб користувачів та впроваджувати нові можливості на основі зворотного зв'язку від експертів у галузі медіа-безпеки.

2.4. Проєктування сценаріїв використання системи

Для наочного представлення взаємодії користувача з системою виявлення *deepfake* відео була розроблена діаграма прецедентів (рис. 2.2), яка ілюструє основні функціональні можливості та зв'язки між компонентами системи.

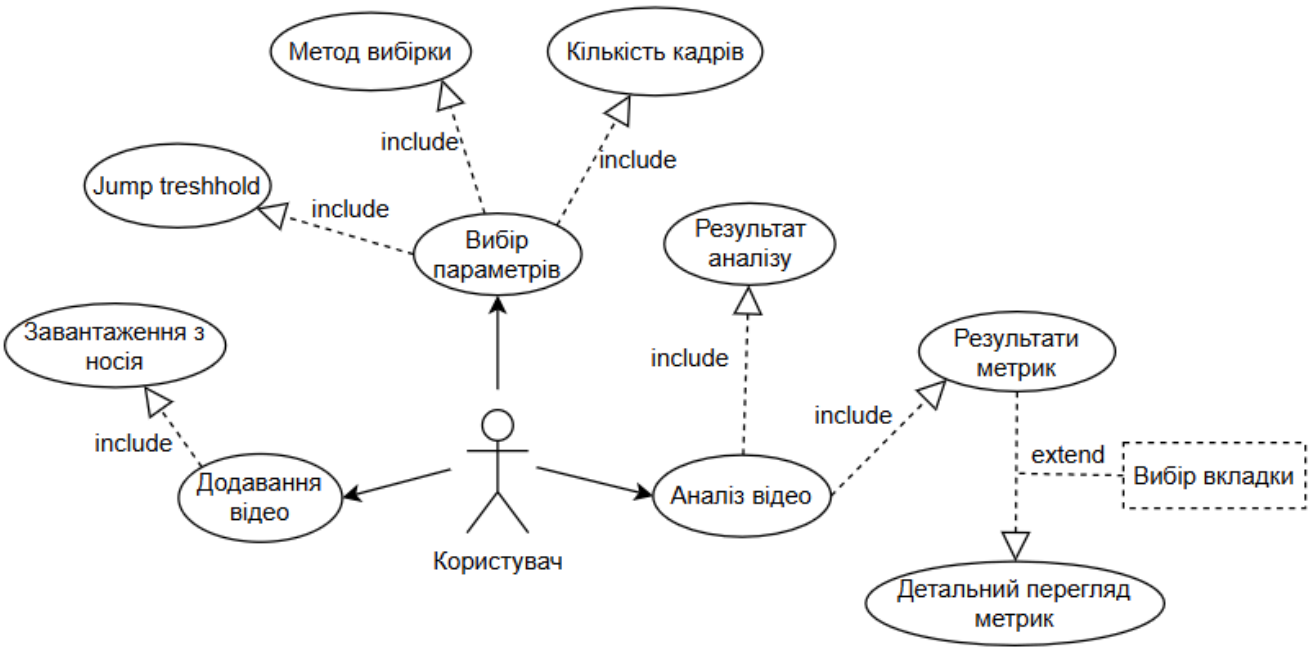


Рис. 2.2. Діаграма прецедентів системи виявлення *deepfake*

На діаграмі представлено основного актора – Користувача, який ініціює процес аналізу відео. Центральним прецедентом є "Аналіз відео", який включає в себе низку допоміжних прецедентів та залежностей. Прецедент "Додавання відео" передує основному аналізу та включає функцію "Завантаження з носія", що дозволяє користувачу вибрати відеофайл з локального сховища або іншого джерела.

Сценарій 1: Базовий аналіз відео

Мета: Перевірити автентичність відеоматеріалу за допомогою системи виявлення *deepfake*.

Передумови: Користувач має відеофайл, який потребує перевірки на наявність *deepfake* маніпуляцій.

Основний потік подій:

– Користувач відкриває вебінтерфейс системи через браузер.

– На головній сторінці користувач натискає кнопку "Завантажити відео".

– Система відображає ім'я завантаженого файлу та його базові характеристики (тривалість, розмір).

– Користувач натискає кнопку "Аналізувати" для запуску процесу обробки.

– Система відображає індикатор прогресу обробки.

– Після завершення аналізу система представляє результати у вигляді кольорового індикатора (зелений для *Real*, червоний для *Fake*, помаранчевий для *Uncertain*).

– Користувач бачить фінальну оцінку впевненості та показник стабільності (*stability score*).

– Система відображає ключові кадри з відео з виділеними обличчями та індивідуальними оцінками для кожного кадру.

Альтернативні потоки:

– Якщо система не виявила жодного обличчя на кадрах, відображається повідомлення про неможливість проведення аналізу.

– Якщо формат відеофайлу не підтримується, система повідомляє про помилку та пропонує список підтримуваних форматів.

Результат: Користувач отримує рішення щодо автентичності відео з обґрунтуванням у вигляді метрик та візуалізацій.

Сценарій 2: Налаштування параметрів аналізу

Мета: Адаптувати процес аналізу під конкретні вимоги користувача через зміну параметрів обробки.

Передумови: Користувач розуміє базові принципи роботи системи та хоче налаштувати параметри для більш точного або швидкого аналізу.

Основний потік подій:

– Користувач завантажує відеофайл до системи.

– У розділі "Додаткові параметри" користувач обирає метод вибірки кадрів, встановлює кількість кадрів для аналізу за допомогою повзунка та налаштовує параметр *Jump threshold* – чутливість виявлення різких змін між кадрами.

– Після встановлення всіх параметрів користувач натискає "Аналізувати".

– Система виконує обробку з урахуванням обраних налаштувань.

Результат: Користувач отримує результати аналізу, оптимізовані під його конкретні потреби та вимоги до швидкості або точності.

Сценарій 3: Детальний перегляд метрик та статистики

Мета: Надати користувачу можливість поглибленого аналізу результатів через детальний перегляд статистичних метрик та візуалізацій.

Передумови: Користувач завершив базовий аналіз відео та хоче детальніше вивчити результати для прийняття більш обґрунтованого рішення.

Основний потік подій:

– Після завершення аналізу відео користувач бачить стислий результат на головному екрані.

– Користувач обирає вкладку з детальними метриками для перегляду поглибленої інформації.

Результат: Користувач отримує всебічне розуміння результатів аналізу з можливістю самостійно оцінити надійність висновків системи на основі детальних статистичних показників та візуалізацій.

Представлені сценарії охоплюють основні варіанти використання системи від простого базового аналізу до поглибленого дослідження результатів з налаштуванням параметрів. Кожен сценарій розроблений з урахуванням різних рівнів технічної підготовки користувачів та різних цілей використання системи.

2.5. Висновки до розділу

У другому розділі кваліфікаційної роботи виконано комплексний аналіз проблематики виявлення *deepfake* відеоматеріалів та розроблено концептуальні основи системи детекції синтетичних відео. Дослідження показало, що сучасні методи генерації *deepfakes* базуються на технологіях глибинного навчання, зокрема генеративних мережах та автокодувальниках, які досягли високого рівня реалістичності синтезованого контенту.

Проведений огляд існуючих підходів виявив, що більшість систем зосереджується на покадровому аналізі візуальних артефактів. Однак така стратегія має обмеження, оскільки генеративні моделі постійно вдосконалюються і їхня ефективність знижується при застосуванні до нових методів генерації.

Розроблена концепція системи пропонує інноваційний підхід, що поєднує традиційну покадрову класифікацію на основі глибинного навчання з методом статистичного аналізу темпоральної послідовності *predictions*. Ключова ідея полягає в тому, що навіть якщо окремі кадри *deepfake* відео мають високу візуальну якість, генеративні моделі часто не здатні підтримувати стабільність характеристик протягом всієї відеопослідовності. Запропонований метод *Multi-Frame Statistical Analysis* аналізує варіативність *predictions*, частоту різких стрибків, ентропію розподілу та автокореляцію, що дозволяє виявляти темпоральні артефакти, характерні для синтетичних відеоматеріалів.

Визначені функціональні вимоги охоплюють весь цикл обробки відео від завантаження до представлення результатів з візуалізацією метрик. Нефункціональні вимоги встановлюють якісні характеристики системи: обробка тридцятисекундного відео за дві хвилини, точність класифікації не менше 90 відсотків та зручність використання через вебінтерфейс. Пріоритезація вимог дозволила визначити мінімально життєздатний продукт, який було повністю реалізовано, та сформувати дорожню карту подальшого розвитку системи.

Розроблена діаграма прецедентів та детальні сценарії використання описують основні варіанти взаємодії користувача з системою від базового аналізу до поглибленого дослідження статистичних метрик. Таким чином, у другому розділі було сформовано повну концептуальну основу для розробки системи виявлення *deepfakes*, що включає аналіз проблемної області, опис оригінального підходу до детекції, визначення вимог та сценарії використання. Запропонований підхід має потенціал підвищити надійність виявлення *deepfakes* порівняно з традиційними методами покадрової класифікації.

РОЗДІЛ 3
РОЗРОБЛЯННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ВИЯВЛЕННЯ *DEEPFAKES* У ВІДЕО

3.1. Вибір підходу навчання моделі

Для задачі виявлення *deepfake* відеоматеріалів існує декілька основних підходів до навчання моделі класифікації, кожен з яких має свої переваги та обмеження.

Навчання з нуля (*Training from scratch*) передбачає створення нейронної мережі з випадковою ініціалізацією ваг та її повне навчання на цільовому датасеті *deepfake* зображень. Цей підхід дозволяє моделі повністю адаптуватися під специфіку задачі без будь-яких попередніх припущень. На конкурентних ринках володіння власною моделлю штучного інтелекту може бути значним бізнес-активом. На відміну від ліцензування сторонньої моделі (до якої також мають доступ конкуренти), модель, навчена з нуля, стає частиною ексклюзивної інтелектуальної власності. Це не лише відрізняє її, але й дозволяє постійно вдосконалювати та оптимізувати модель без залежності від зовнішніх постачальників, надаючи конкурентну перевагу [17].

Однак навчання з нуля потребує дуже великих обсягів даних, зазвичай мільйонів зображень, та значних обчислювальних ресурсів. Для досягнення хорошої точності може знадобитися кілька тижнів або навіть місяців навчання на потужному обладнанні. Крім того, модель, навчена з нуля на обмеженому датасеті, схильна до перенавчання та погано генералізується на нові типи *deepfakes*, які не були представлені в навчальній вибірці.

Transfer Learning (трансферне навчання) – це метод машинного навчання, в якому знання, отримані в результаті виконання одного завдання або набору даних,

використовуються для покращення продуктивності моделі в іншому пов'язаному завданні або на іншому наборі даних. Іншими словами, трансферне навчання використовує те, що було вивчено в одному середовищі, для покращення узагальнення в іншому середовищі [18].

Transfer Learning використовує модель, яка вже була навчена на великому загальному датасеті зображень, таких як *ImageNet* з мільйонами фотографій різних об'єктів. Ідея полягає в тому, що нижні шари нейронної мережі вчаться розпізнавати універсальні візуальні паттерни, такі як краї, текстури, форми та кольори, які є корисними для будь-яких задач комп'ютерного зору. При трансферному навчанні ці попередньо навчені ваги використовуються як початкова точка, а потім модель дотреновується на специфічному датасеті *deepfakes*. Це значно скорочує час навчання та зменшує потребу в великій кількості даних, оскільки

Fine-tuning (тонке налаштування) в машинному навчанні – це процес адаптації попередньо навченої моделі для конкретних завдань або випадків використання. Це стало фундаментальною технікою глибокого навчання, особливо в процесі навчання базових моделей, що використовуються для генеративного штучного інтелекту [19]. *Fine-tuning* є різновидом трансферного навчання, де не вся модель заморожується, а лише частина шарів. Зазвичай заморожуються нижні шари, які відповідають за виявлення базових візуальних ознак, а верхні шари, які відповідають за більш специфічні характеристики, продовжують навчатися на цільовому датасеті. Це дозволяє моделі зберегти загальні знання про зображення, але адаптуватися до специфічних особливостей *deepfake* детекції.

Few-shot Learning (навчання на малій кількості прикладів) розроблений для ситуацій, коли доступна дуже обмежена кількість навчальних даних. Це метод машинного навчання, який дозволяє попередньо навченій моделі узагальнювати дані для нових категорій даних (з якими попередньо навчена модель не стикалася під час навчання), використовуючи лише кілька позначених зразків на клас. Це підпадає під парадигму метанавчання (метанавчання означає навчання навчанню)

[20]. Однак для *deepfake* детекції цей підхід менш популярний, оскільки існують достатньо великі публічні датасети з реальними та синтетичними обличчями.

Self-supervised Learning (самоконтрольоване навчання) передбачає навчання моделі на немаркованих даних шляхом створення псевдозадач, таких як передбачення повернутого зображення або відновлення замаскованих частин. Після такого попереднього навчання модель дотренується на маркованих даних *deepfakes*. Самостійне навчання особливо корисне в таких галузях, як комп'ютерний зір та обробка природної мови, які потребують великих обсягів маркованих даних для навчання сучасних моделей штучного інтелекту. Оскільки ці марковані набори даних вимагають трудомісткого анотування експертами-людьми, збір достатньої кількості даних може бути надзвичайно складним. Підходи із самостійним керуванням можуть бути більш ефективними з точки зору часу та витрат, оскільки вони замінюють частину або всю необхідність ручного маркування навчальних даних [21].

Для розробленої системи виявлення *deepfakes* було обрано підхід *Transfer Learning* з *Fine-tuning* на основі попередньо навченої архітектури *EfficientNet-B0*.

EfficientNet-B0 – це згорткова нейронна мережа, яка навчається на більш ніж мільйоні зображень з бази даних *ImageNet*. Мережа може класифікувати зображення за 1000 категоріями об'єктів, такими як клавіатура, миша, олівець та багато іншого. В результаті мережа вивчила багаті представлення ознак для широкого діапазону зображень [22].

Цей вибір обумовлений декількома важливими факторами, які забезпечують оптимальний баланс між точністю, швидкістю навчання та ефективністю використання ресурсів:

1) Використання попередньо навченої моделі на датасеті *ImageNet* надає значну перевагу в якості початкових представлень. *EfficientNet-B0* вже навчилася розпізнавати базові візуальні паттерни, текстури та форми на мільйонах зображень різних категорій. Ці знання є безпосередньо застосовними до задачі аналізу облич, оскільки модель вже розуміє структуру зображень, розподіл кольорів, характеристики освітлення та геометричні форми. Замість того, щоб витрачати

тижні на навчання цих базових представлень з нуля, система може одразу зосередитися на вивченні специфічних артефактів *deepfake* маніпуляцій.

2) Цей підхід значно скорочує вимоги до обсягу навчальних даних. Навчання моделі з нуля для задачі класифікації облич потребувало б мільйонів зразків для досягнення прийнятної точності. Натомість, використовуючи трансферне навчання, система може досягти високої точності на датасетах середнього розміру, таких як *FaceForensics++* з десятками тисяч зображень. Це критично важливо, оскільки створення великомасштабних датасетів *deepfakes* є трудомістким процесом, що потребує значних ресурсів для генерації та маркування синтетичних відео.

3) Час навчання при використанні трансферного навчання скорочується в десятки разів порівняно з навчанням з нуля. Типовий цикл навчання моделі з нуля для задачі класифікації зображень може тривати від кількох тижнів до місяця на потужному *GPU*. При використанні попередньо навченої моделі час навчання скорочується до кількох днів або навіть годин, залежно від розміру датасету та доступних обчислювальних ресурсів. Це дозволяє швидше експериментувати з різними конфігураціями, підбирати оптимальні гіперпараметри та ітеративно покращувати модель.

4) Трансферне навчання забезпечує кращу генералізацію моделі на невідомі типи *deepfakes*. Модель, навчена з нуля на обмеженому датасеті, може запам'ятати специфічні артефакти конкретних методів генерації, представлених у навчальних даних, але не здатна розпізнати нові типи маніпуляцій. Попередньо навчена модель має більш універсальні візуальні представлення, що дозволяє їй краще адаптуватися до нових варіантів *deepfakes*, які з'являються з розвитком генеративних технологій.

5) Архітектура *EfficientNet-B0* спеціально розроблена для забезпечення оптимального балансу між точністю та обчислювальною ефективністю. Порівняно з іншими популярними архітектурами, такими як *ResNet* або *VGG*, *EfficientNet* досягає порівнянної або вищої точності при значно меншій кількості параметрів та операцій. Це робить модель придатною для розгортання не лише на серверах, але

й на пристроях з обмеженими ресурсами, що розширює можливості практичного застосування системи.

Стратегія *fine-tuning*, застосована в системі, передбачає часткове заморожування нижніх шарів мережі, які відповідають за виявлення базових візуальних ознак, та активне навчання верхніх шарів разом з новим *classification head*, спеціально адаптованим для бінарної класифікації *real/fake*. Це дозволяє зберегти загальні знання про структуру зображень, накопичені при навчанні на *ImageNet*, і водночас адаптувати модель до специфічних характеристик *deepfake* детекції, таких як аномалії в текстурі шкіри, нереалістичне освітлення контурів обличчя та артефакти, що виникають при блендингу синтетичного обличчя з оригінальним фоном.

Додатково, в системі застосовується розширена аугментація даних, включаючи імітацію *JPEG* компресії, розмиття, зміну кольорів та випадкове видалення областей зображення. Ці техніки працюють синергетично з трансферним навчанням, допомагаючи моделі фокусуватися на справжніх характеристиках *deepfakes*, а не на артефактах конкретного датасету або умов зйомки. Використання *advanced augmentation techniques*, таких як *Mixup* та *CutMix*, де зображення з різних класів змішуються між собою, додатково покращує здатність моделі до генералізації та робить її більш стійкою до різних типів маніпуляцій.

Таким чином, обраний підхід *Transfer Learning з Fine-tuning* на основі *EfficientNet-B0* забезпечує оптимальне рішення для задачі виявлення *deepfakes*, поєднуючи високу точність класифікації, швидкість навчання, ефективне використання обмежених навчальних даних та хорошу генералізацію на нові типи синтетичних відеоматеріалів при прийнятних вимогах до обчислювальних ресурсів.

3.2. Розроблення архітектури інтелектуальної системи

Представлена система виявлення *deepfakes* реалізована за принципами модульної багатошарової архітектури з чітким розділенням відповідальності (*Modular Layered Architecture*). Це архітектурний підхід, який організовує код у логічно згруповані модулі та шари, де кожен компонент має чітко визначену відповідальність і мінімальну залежність від інших компонентів. Така архітектура забезпечує високу підтримуваність, масштабованість, тестованість та можливість паралельної розробки різних частин системи.

Основою архітектури є принцип поділу системи на вертикальні функціональні модулі та горизонтальні шари абстракції. Вертикальні модулі представлені директоріями верхнього рівня, кожна з яких інкапсулює певну функціональну область системи. Горизонтальні шари визначають рівень абстракції та близькість до користувача або до низькорівневих операцій. Такий підхід дозволяє розробникам легко орієнтуватися в кодовій базі, швидко локалізувати функціональність і вносити зміни без ризику порушити інші частини системи.

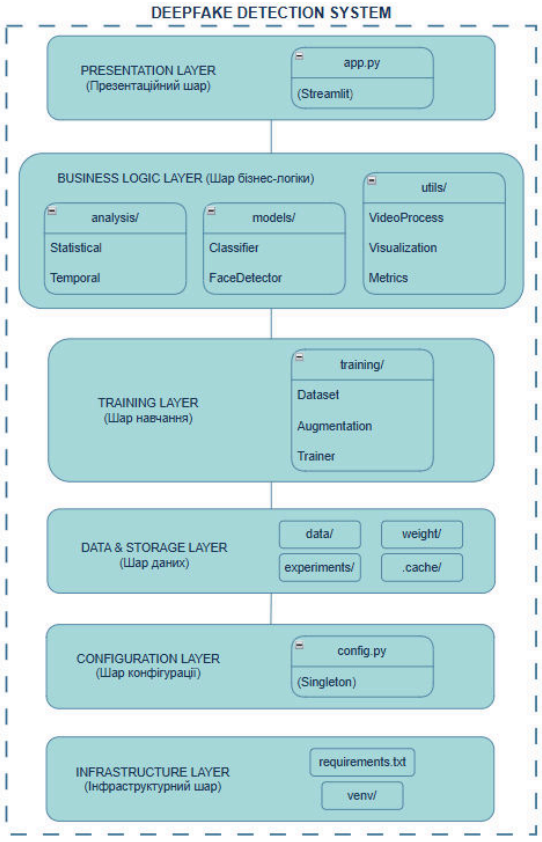


Рис. 3.1. Архітектура інтелектуальної системи

Структура проєкту організована навколо шести основних функціональних модулів:

1. Модуль *analysis* містить компоненти для аналізу даних, зокрема *statistical_analyzer.py* для статистичного аналізу послідовностей *predictions* та *temporal_analyzer.py* для темпорального аналізу часових патернів. Цей модуль представляє аналітичний шар системи і є незалежним від способу отримання даних, що дозволяє використовувати його алгоритми для аналізу будь-яких числових послідовностей.

2. Модуль *models* інкапсулює всі компоненти машинного навчання, включаючи classifier.py з реалізацією нейронної мережі для класифікації облич та *face_detector.py* з логікою виявлення облич на зображеннях. Цей модуль представляє шар моделей і може функціонувати незалежно від решти системи, що дозволяє тренувати, тестувати та розгортати моделі окремо.

3. Модуль *training* відповідає за весь процес навчання моделей і містить dataset.py для завантаження та організації тренувальних даних, augmentation.py для аугментації зображень з метою покращення генералізації моделі, та *train.py* з повним *pipeline* навчання. Цей модуль представляє шар підготовки даних та навчання, він активно використовує модуль *models* але залишається відокремленим від продакшн-коду, що дозволяє проводити експерименти без впливу на робочу систему.

4. Модуль *utils* містить допоміжні утиліти загального призначення, які можуть використовуватися будь-якими іншими модулями системи. Тут знаходяться *metrics.py* для обчислення метрик якості моделі, *video_processor.py* для роботи з відеофайлами (декодування, екстракція кадрів, метадані), та *visualization.py* для створення графіків та візуалізацій результатів.

Директорії *data* та *experiments* представляють організаційні шари для зберігання даних та результатів експериментів відповідно, хоча самі не містять програмного коду. Директорія *weights* призначена для зберігання натренованих ваг моделей, що дозволяє версіонувати та швидко перемикатися між різними версіями моделей. Технічні директорії *.cache* та *pycache* (присутні в кожному модулі) містять

кешовані дані та скомпільовані *Python* файли відповідно, що прискорює роботу системи.

На самому верхньому рівні розташовані ключові файли координації системи. Файл *app.py* є точкою входу для вебдодатку і представляє презентаційний шар архітектури. Він імпортує та координує роботу всіх нижчих модулів, надаючи користувацький інтерфейс через *Streamlit framework*. Файл *config.py* централізовано управляє всією конфігурацією системи через набір *dataclass* структур для різних аспектів роботи (*ModelConfig, StatisticalAnalysisConfig, VideoProcessingConfig, ThresholdConfig, TrainingConfig*). Цей файл реалізує патерн *Singleton* для глобального доступу до конфігурації з будь-якої частини системи. Файли *requirements.txt*, *deepfake_detector.pyproj* та *deepfake_detector.sln* представляють інфраструктурний шар для управління залежностями та конфігурацією проєкту.

Кожен модуль містить файл *init.py*, що робить директорію *Python* пакетом і визначає публічний *API* модуля через механізм *all*. Це реалізує принцип інкапсуляції, коли модуль експортує тільки ті компоненти, які призначені для використання ззовні, приховуючи внутрішні деталі реалізації. Така організація створює чіткі контракти між модулями і знижує зв'язаність системи.

Архітектура реалізує кілька важливих архітектурних патернів. *Layered Architecture* проявляється в чіткому поділі на презентаційний шар (*app.py*), бізнес-логіки (*analysis, utils*), моделей (*models*), даних (*training, data*) та конфігурації (*config.py*). Кожен шар залежить тільки від нижчих шарів, що запобігає циклічним залежностям. *Module Pattern* реалізований через організацію функціональності в окремі директорії-пакети з чітко визначеними публічними інтерфейсами. *Separation of Concerns* забезпечується тим, що кожен модуль відповідає за один аспект системи: *models* – за машинне навчання, *analysis* – за аналітику, *training* – за навчання, *utils* – за допоміжні функції. *Dependency Injection* реалізується через *config.py*, який надає всі налаштування як зовнішні залежності, а не хардкодить їх всередині модулів.

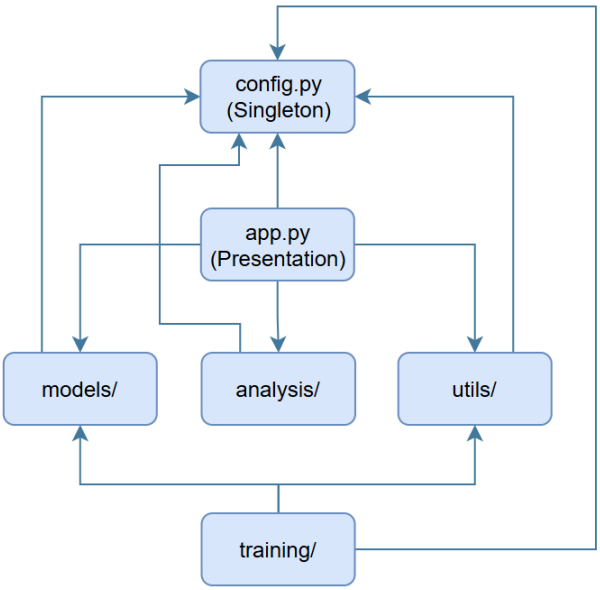


Рис. 3.2. Залежність компонентів інтелектуальної системи

Потоки залежностей в архітектурі організовані зверху вниз. Файл *app.py* на вершині імпортує з усіх модулів (*models, analysis, utils, config*), але сам не імпортується нікуди, що робить його кінцевою точкою залежностей. Модулі *models* та *analysis* є незалежними один від одного і не імпортують один одного, що дозволяє використовувати їх окремо. Модуль *training* залежить від *models* (використовує моделі для навчання) та *utils* (використовує утиліти для роботи з даними), але *models* не залежить від *training*, що дозволяє використовувати натреновані моделі без коду навчання. Модуль *utils* є незалежним і містить функції загального призначення, які можуть використовуватися в будь-яких інших модулях. Всі модулі залежать від *config.py* для отримання налаштувань, але *config.py* не залежить ні від кого, що робить його фундаментом системи.

Така архітектура надає суттєві переваги:

1) Модульність і можливість заміни компонентів – можна замінити алгоритм детекції облич в *models/face_detector.py* без впливу на інші частини, змінити методи статистичного аналізу в *analysis/statistical_analyzer.py* незалежно від моделей, або додати нові методи візуалізації в *utils/visualization.py* без модифікації основного коду.

2) Тестованість – кожен модуль можна тестувати ізольовано, що спрощує написання unit тестів, а наявність init.py з чітким *API* дозволяє легко створювати *mock* об'єкти для тестування.

3) Масштабованість – нову функціональність можна додавати як нові модулі без модифікації існуючих, структура підтримує як вертикальне масштабування (додавання нових можливостей в існуючі модулі), так і горизонтальне (додавання нових модулів).

4) Підтримуваність – чітка структура дозволяє швидко знайти потрібний код, зміни локалізовані в конкретних модулях і не розповсюджуються по всій системі, а нові розробники можуть швидко зрозуміти організацію проєкту.

Середовище розробки підтримується через *venv* директорію з ізольованим *Python* оточенням, *requirements.txt* для керування залежностями проєкту, та *.pyproj* і *.sln* файли для інтеграції з *Visual Studio*. Така організація забезпечує відтворюваність середовища розробки та спрощує онбординг нових членів команди.

Важливою характеристикою є слабка зв'язаність модулів. Модулі взаємодіють через чітко визначені інтерфейси (публічні класи та функції), не залежать від внутрішньої реалізації один одного, використовують спільний *config.py* як єдине джерело конфігурації замість прямих залежностей, та можуть функціонувати незалежно для тестування або використання в інших проєктах. Водночас висока зв'язаність всередині модулів забезпечує, що вся функціональність, пов'язана з однією задачею, знаходиться в одному місці, не розпорошена по різних частинах системи, легко знаходиться розробниками, та може бути модифікована без пошуку пов'язаного коду в інших місцях.

3.3. Структура даних, використаних для навчання моделі

Для успішного навчання моделі виявлення *deepfake* необхідно мати якісний та правильно структурований набір даних. У даному проєкті використовується датасет *Deepfake and Real Images*, який є загальнодоступним набором даних на

платформі *Kaggle*. Набір даних містить маніпульовані зображення та реальні зображення. Маніпульовані зображення – це обличчя, створені різними способами. Цей набір даних був оброблений з метою отримання максимального результату від цих зображень. Кожне зображення – це зображення людського обличчя у форматі *JPG* розміром 256 × 256, реальне або штучне [23]. Цей датасет було обрано завдяки його збалансованості, достатньому обсягу даних та зручній структурі, що дозволяє швидко розпочати процес навчання без складної попередньої обробки.

Датасет містить приблизно 190 тисяч зображень облич. Усі зображення рівномірно розподілені між двома класами: справжні обличчя реальних людей та синтетичні обличчя, згенеровані за допомогою генеративних змагальних мереж. Така збалансованість є критично важливою характеристикою для навчання нейронних мереж, оскільки незбалансовані дані можуть призвести до того, що модель буде віддавати перевагу домінуючому класу та погано розпізнавати менш представлений клас. У даному випадку кількість справжніх та підроблених зображень є приблизно однаковою, що забезпечує об'єктивне навчання класифікатора.

Структура датасету організована за принципом розділення на три незалежні підмножини: навчальну вибірку, валідаційну вибірку та тестову вибірку. Навчальна вибірка містить найбільшу кількість зображень і використовується безпосередньо для оновлення ваг нейронної мережі під час процесу навчання. Валідаційна вибірка застосовується для оцінки якості моделі після кожної епохи навчання та дозволяє контролювати процес навчання, виявляти перенавчання та обирати найкращу версію моделі. Тестова вибірка використовується лише один раз після завершення навчання для отримання фінальної оцінки якості моделі на даних, які вона ніколи не бачила під час навчання.

Кожна з трьох вибірок має ідентичну внутрішню структуру та складається з двох папок: *Real* для справжніх зображень та *Fake* для підроблених. Така організація дозволяє автоматично визначати мітки класів на основі назви папки, в якій знаходиться зображення. Це є стандартним підходом у глибокому навчанні,

який називається *ImageFolder* структурою та підтримується більшістю бібліотек машинного навчання, включаючи *PyTorch*.

Усі зображення у датасеті вже попередньо оброблені та представляють собою вирізані ділянки облич. Це означає, що етап детекції обличчя на зображенні вже виконано, і кожен файл містить лише область обличчя без зайвого фону. Такий підхід значно спрощує процес підготовки даних та дозволяє моделі зосередитися виключно на аналізі характеристик обличчя. Зображення збережені у форматі з достатньою роздільною здатністю для виявлення характерних ознак як справжніх, так і згенерованих облич.

Справжні зображення у датасеті походять з різноманітних джерел та містять обличчя реальних людей різного віку, статі, етнічної приналежності та з різними умовами освітлення. Така різноманітність є важливою для узагальнюючої здатності моделі, оскільки дозволяє їй навчитися розпізнавати справжні обличчя незалежно від індивідуальних характеристик людини чи умов зйомки. Підроблені зображення були згенеровані за допомогою технології *StyleGAN*, яка є однією з найпотужніших генеративних моделей для створення фотореалістичних зображень неіснуючих людей. Ці синтетичні обличчя виглядають надзвичайно переконливо для людського ока, однак містять характерні артефакти та несумісності на рівні пікселів, які можуть бути виявлені алгоритмами машинного навчання.

Загалом структура даних для навчання забезпечує ефективний та надійний процес підготовки моделі виявлення *deepfake*. Збалансованість класів гарантує об'єктивне навчання, розділення на три вибірки дозволяє контролювати якість та уникати перенавчання. Така організація даних є стандартною практикою у галузі комп'ютерного зору та дозволяє досягти високої точності моделі при відносно невеликому обсязі початкових даних.

3.4. Процес навчання моделі

Процес розроблення та налаштування інтелектуальної системи виявлення *deepfakes* є багатоетапною процедурою, яка виходить далеко за межі простого

завантаження зображень у нейронну мережу. У даній роботі реалізовано комплексний підхід, що включає попередню обробку, аугментацію даних, налаштування архітектури мережі та ітеративний цикл навчання з використанням сучасних методів оптимізації. Весь процес можна умовно розділити на підготовчий етап, етап безпосереднього тренування та етап валідації. Візуалізацію даного процесу можна побачити на рисунку 3.3.

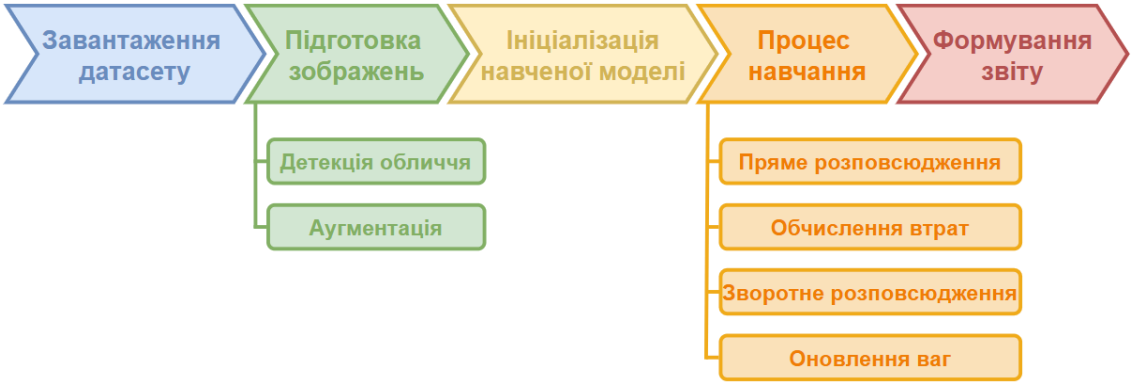


Рис. 3.3. Етапи навчання моделі

На самому початку система працює з сирими вхідними даними. Оскільки *deepfake*-маніпуляції зазвичай стосуються лише ділянки обличчя, а не всього кадру, критично важливим першим кроком є локалізація та виокремлення обличчя. Для цього використовується спеціалізований детектор (у системі реалізовано підтримку алгоритму *MTCNN*), який сканує зображення та визначає координати обличчя. Важливим нюансом є те, що система не просто вирізає обличчя по контуру, а додає певний відступ (*margin*) навколо нього. Це робиться для того, щоб зберегти контекст (контур голови, волосся, вуха), оскільки артефакти генерації часто з'являються саме на межах накладання маски. Після виокремлення зображення обличчя масштабується до фіксованого розміру (наприклад, 224 × 224 пікселі), що є необхідною умовою для подачі даних на вхід нейронної мережі.

Наступним етапом є аугментація даних. Оскільки моделі глибокого навчання схильні до перенавчання (запам'ятовування конкретних прикладів замість вивчення закономірностей), можна штучно розширити та урізноманітнити навчальну вибірку. У розробленій системі застосовується два типи трансформацій: геометричні та піксельні. Геометричні перетворення включають випадкові

повороти, горизонтальні відображення та зміни масштабу. Це вчить модель розпізнавати обличчя незалежно від його положення в кадрі. Піксельні перетворення є більш специфічними для задачі *deepfake*. Сюди входить додавання цифрового шуму, розмиття (*blur*), зміни яскравості та контрасту, а також, що найважливіше, імітація артефактів стиснення *JPEG*. Останнє є ключовим, оскільки більшість відео в інтернеті проходять через алгоритми стиснення, і модель повинна вміти відрізняти справжні артефакти стиснення від артефактів, залишених генеративними мережами. Окрім стандартних методів, під час навчання використовується техніка *Mixup*. Вона полягає у змішуванні двох різних зображень (наприклад, реального та фейкового) та їхніх міток у певній пропорції. Це змушує нейромережу поводитися більш лінійно та стабільно при обробці неоднозначних прикладів, суттєво покращуючи її узагальнюючу здатність.

Після підготовки даних починається ініціалізація моделі. Використовується підхід *Transfer Learning* (перенос навчання). Замість того, щоб тренувати мережу "з нуля" з випадковими ваговими коефіцієнтами, за основу береться архітектура (*EfficientNet-B0*), яка вже була попередньо навчена на гігантському наборі даних *ImageNet*. Це дозволяє моделі вже на старті "розуміти", як виглядають базові текстури, форми та об'єкти. Замінюється лише остання частина мережі (класифікатор) на нові шари, призначені для потрібної бінарної задачі: відрізнити "реальне" від "підробленого". Також в архітектуру додається блок уваги (*Attention Block*), який допомагає моделі фокусуватися на найбільш інформативних зонах обличчя (очі, рот), ігноруючи фон.

Безпосереднє навчання відбувається ітеративно, епохами. Одна епоха – це повний прохід через весь набір навчальних даних. Дані подаються в модель не по одному, а пакетами (батчами). Процес всередині епохи складається з кількох кроків. Спочатку відбувається пряме розповсюдження (*forward pass*): модель отримує зображення і робить передбачення. Далі обчислюється функція втрат (*Loss Function*). У системі використовується *Cross Entropy Loss* з технікою *Label Smoothing*. Це означає, що від моделі не вимагається абсолютної впевненості (100%

real або 100% *fake*), а дозволено невелику похибку (наприклад, 90% впевненості), що запобігає надмірній самовпевненості моделі на складних прикладах.

Після обчислення помилки відбувається зворотне розповсюдження (*backward pass*). Система обчислює градієнти – напрямки, в яких потрібно змінити ваги нейронів, щоб зменшити помилку. Тут вступає в дію оптимізатор *AdamW*, який коригує ваги моделі. Для підвищення ефективності та швидкості навчання використовується технологія змішаної точності (*Automatic Mixed Precision – AMP*). Вона дозволяє виконувати частину обчислень у форматі половинної точності (*FP16*) замість стандартної (*FP32*), що економить пам'ять відеокарти та прискорює процес без втрати якості навчання. Також застосовується "обрізання градієнтів" (*Gradient Clipping*), щоб запобігти ситуаціям, коли ваги змінюються занадто різко, що може зруйнувати процес навчання.

Важливим елементом є керування швидкістю навчання (*Learning Rate*). Використовується планувальник (*Scheduler*) типу *Cosine Annealing* або *ReduceLROnPlateau*. На початку навчання швидкість вища, щоб модель швидко наблизилася до оптимального рішення, а згодом вона плавно знижується, дозволяючи мережі "тонко налаштуватися" і знайти глобальний мінімум помилки.

Після завершення кожної епохи навчання настає фаза валідації. Модель перемикається в режим оцінки і проганяється через окремий, відкладений набір даних, який вона ніколи не бачила під час тренування. Це дозволяє об'єктивно оцінити, наскільки добре система працює на нових даних. Обчислюються метрики точності, *F1-score* та *AUC*. Якщо поточна модель показує кращий результат на валідації, ніж попередня найкраща, вона зберігається як контрольна точка. Також реалізовано механізм ранньої зупинки (*Early Stopping*): якщо протягом певної кількості епох результати на валідації не покращуються, навчання припиняється автоматично. Це економить обчислювальні ресурси та запобігає перенавчанню моделі, коли вона починає просто запам'ятовувати тренувальні дані, втрачаючи здатність до узагальнення.

Таким чином, розроблений процес навчання є замкненим циклом, який поєднує в собі просунуту підготовку даних, сучасну архітектуру нейромережі та

ефективні методи оптимізації, що в сукупності дозволяє досягти високої точності детекції *deepfakes*.

Для перевірки працездатності розробленої системи та оцінки її ефективності було проведено експериментальне навчання моделі на основі архітектури *EfficientNet-B0*.

Процес навчання відбувався на збалансованому наборі даних, який складався з 140002 зображень для тренування (порівну реальних та фейкових зразків – по 70001 кожного класу) та 39428 зображень для валідації. Такий обсяг даних дозволив мінімізувати ризик зміщення результатів у бік одного з класів. Навчання проводилося протягом двох епох із розміром пакету (*batch size*) 32 зображення. Варто зазначити, що через технічні обмеження обчислення виконувалися на *CPU*, що суттєво вплинуло на тривалість процесу – кожна епоха займала понад 8,5 годин, проте це підтвердило можливість роботи системи навіть без спеціалізованих графічних прискорювачів.

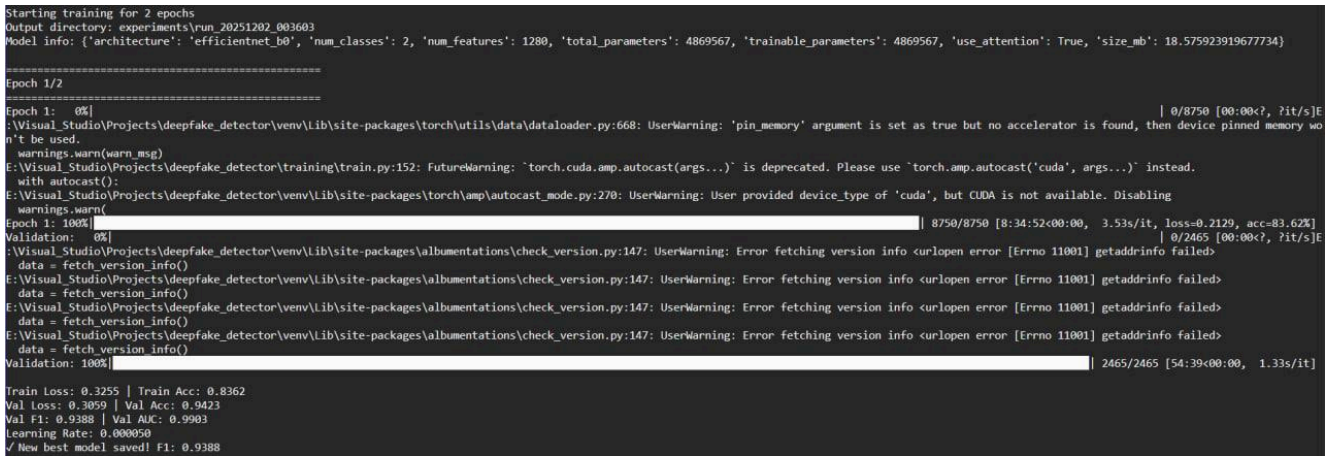


Рис. 3.4. Результати першої епохи навчання

Аналіз результатів першої епохи (рис. 3.4) демонструє високу ефективність обраного методу трансферного навчання (*Transfer Learning*). Вже після першого проходу по всьому навчальному набору даних модель показала точність на валідації (*Validation Accuracy*) на рівні 94,23%. При цьому точність на тренувальному наборі склала 83,62%. Той факт, що точність на валідації виявилася вищою, ніж на тренуванні, є характерною ознакою використання сильної аугментації даних (*Mixup*, спотворення кольорів, геометричні трансформації) під

час навчання. Моделі було "важко" вчитися на сильно змінених зображеннях, але вона сформувала стійкі ознаки, які дозволили їй легко класифікувати "чисті" зображення з валідаційного набору.

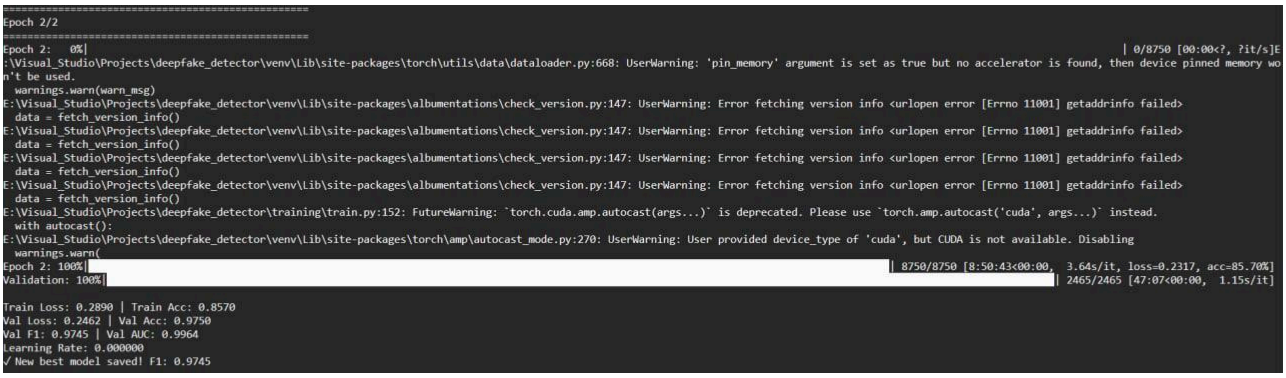


Рис. 3.5. Результати другої епохи навчання

Друга епоха (рис. 3.5) дозволила покращити результати та "доналаштувати" ваги нейронної мережі. Спостерігалося подальше зниження помилки класифікації, це свідчить про те, що модель стала більш "впевненою" у своїх прогнозах. Фінальна точність на валідаційному наборі даних досягла вражаючого показника 97,50%. Це означає, що система правильно класифікувала майже 98 зі 100 зображень, які вона ніколи не бачила раніше. Точність на тренувальному етапі також зросла, що підтверджує стабільний процес засвоєння закономірностей без ознак перенавчання.

Таблиця 3.1

Числове порівняння результатів епох навчання

Метрика	Епоха 1	Епоха 2
Train Loss	0.3255	0.2890
Train Acc	83.62%	85.70%
Val Loss	0.3059	0.2462
Val Acc	94.23%	97.50%
Val F1	0.9388	0.9745
Val AUC	0.9903	0.9964

Особливу увагу слід звернути на комплексні метрики якості. Метрика *Accuracy* (точність) є найбільш зрозумілою – вона показує загальний відсоток правильних відповідей, тобто як часто модель не помилялася (у даному випадку це 97,5%). Показник *Loss* (функція втрат) можна порівняти зі "штрафними балами": чим нижче це число, тим менше модель вагається і тим точніші її передбачення. Метрика *F1-score* є особливою важливою для задач безпеки, оскільки вона контролює баланс: система повинна не лише знаходити всі діпфейки, але й не плутати їх зі справжніми відео (щоб не було хибних тривог). Нарешті, *ROC-AUC* (0.9964) оцінює загальну "професійність" моделі: наскільки якісно вона вміє розрізняти два класи між собою. Значення, близьке до 1.0, свідчить про майже ідеальну здатність відділяти підробку від оригіналу.

Фінальний показник *F1-score* склав 0.9745. Це критично важливо для задачі виявлення діпфейків, оскільки високий *F1* свідчить про баланс: система однаково добре уникає як пропуску підробок (*False Negatives*), так і помилкових звинувачень реальних фото (*False Positives*). Метрика *ROC-AUC*, що оцінює здатність моделі розрізняти класи при різних порогах чутливості, досягла значення 0.9964 (максимум – 1.0). Такий результат наближається до ідеального і вказує на те, що ймовірність того, що модель оцінить випадковий фейк вище, ніж випадкове реальне зображення, складає понад 99%.

За результатами експерименту найкращою версією моделі було визнано стан мережі після другої епохи, оскільки було досягнуто максимального значення цільової метрики *F1* (0.9745). Отримані дані підтверджують, що обрана архітектура у поєднанні з розробленим алгоритмом попередньої обробки забезпечує високу надійність детекції навіть при відносно короткому циклі навчання.

Після завершення процесу навчання важливо провести ретельне тестування моделі на незалежних даних, які вона не бачила під час тренування. Для цього був розроблений спеціальний діагностичний скрипт *diagnose_model.py*, який виконує оцінку продуктивності моделі.

```
 Знайдено зображень:
   Real: 19787
   Fake: 19641

 Тестування на 50 зображеннях кожного класу...

============================================================
 РЕЗУЛЬТАТИ
============================================================

 REAL зображення (очікуємо predictions < 0.5):
   Mean prediction: 0.0747
   Std prediction:  0.0309
   Min: 0.0517, Max: 0.2303
   Правильно класифіковано: 50/50 (100.0%)

 FAKE зображення (очікуємо predictions > 0.5):
   Mean prediction: 0.8749
   Std prediction:  0.1486
   Min: 0.0824, Max: 0.9368
   Правильно класифіковано: 47/50 (94.0%)

 Загальна точність: 97/100 (97.0%)
```

Рис. 3.6. Результати тестування навченої моделі

Результати тестування моделі (рис. 3.6) на 50 зображеннях кожного класу з *validation* датасету. Модель продемонструвала відмінні результати: для *Real* зображень середнє значення *prediction* становить лише 0.0747, що означає що модель впевнено класифікує справжні фото як справжні (усі 50 з 50 правильно, тобто 100% точність). Для *Fake* зображень середнє *prediction* дорівнює 0.8749, що також дуже добре, оскільки модель розпізнає підроблені зображення з високою впевненістю (47 з 50 правильно, 94% точність). Загальна точність моделі склала 97%, що є дуже високим показником і свідчить про якісне навчання.

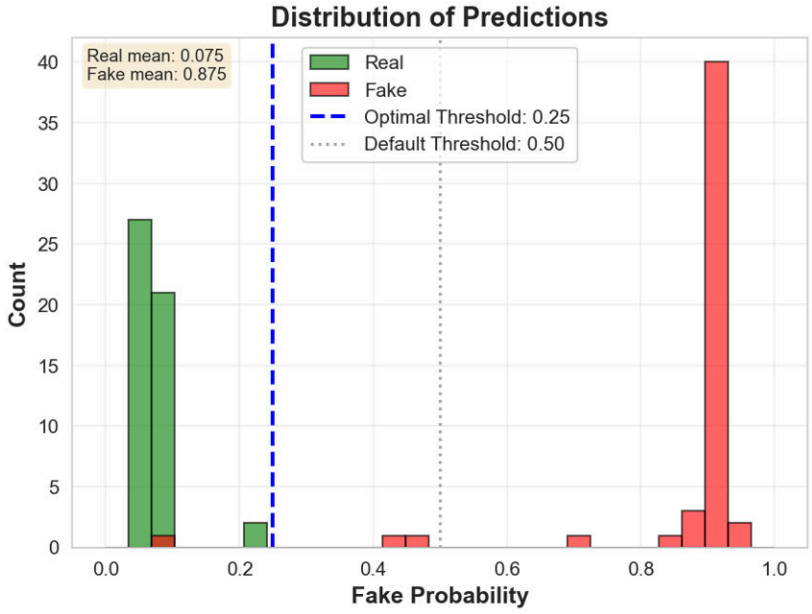


Рис. 3.7. Гістограма розподілу *predictions*

Гістограму розподілу *predictions* (рис. 3.7), яка наочно демонструє чому модель працює так добре. Зелені стовпці (*Real* зображення) сконцентровані в лівій частині графіка біля нуля, а червоні стовпці (*Fake* зображення) зосереджені в правій частині біля одиниці. Це означає що модель чітко розділяє два класи з мінімальним *overlap*. Оптимальний поріг класифікації був визначений як 0.25 (синя пунктирна лінія), що нижче стандартного значення 0.5, і це пояснюється тим що *Real predictions* мають дуже низькі значення. При такому порозі практично всі зображення класифікуються правильно.

Отже, модель навчилася надзвичайно добре і демонструє чудову здатність розрізняти справжні та підроблені зображення. Велика відстань між середніми значеннями *predictions* для двох класів (0.075 для *Real* vs 0.875 для *Fake*) показує що модель має сильну дискримінаційну здатність і не плутає класи між собою. Високі показники точності (100% для *Real*, 94% для *Fake*) і загальна точність 97% підтверджують що архітектура моделі, датасет і процес навчання були підібрані оптимально, а модель готова до практичного використання для детекції *deepfake* зображень.

3.5. Процес виявдення *deepfakes* у відео інтелектуальною системою

Процес детекції *deepfake* відеофайлу в розробленій системі являє собою чітку послідовність етапів, починаючи від завантаження файлу користувачем і закінчуючи формуванням комплексного звіту про його автентичність. Цей алгоритм (рис. 3.8) поєднує в собі глибоке навчання на рівні окремих кадрів та просунутий статистичний аналіз часових закономірностей.

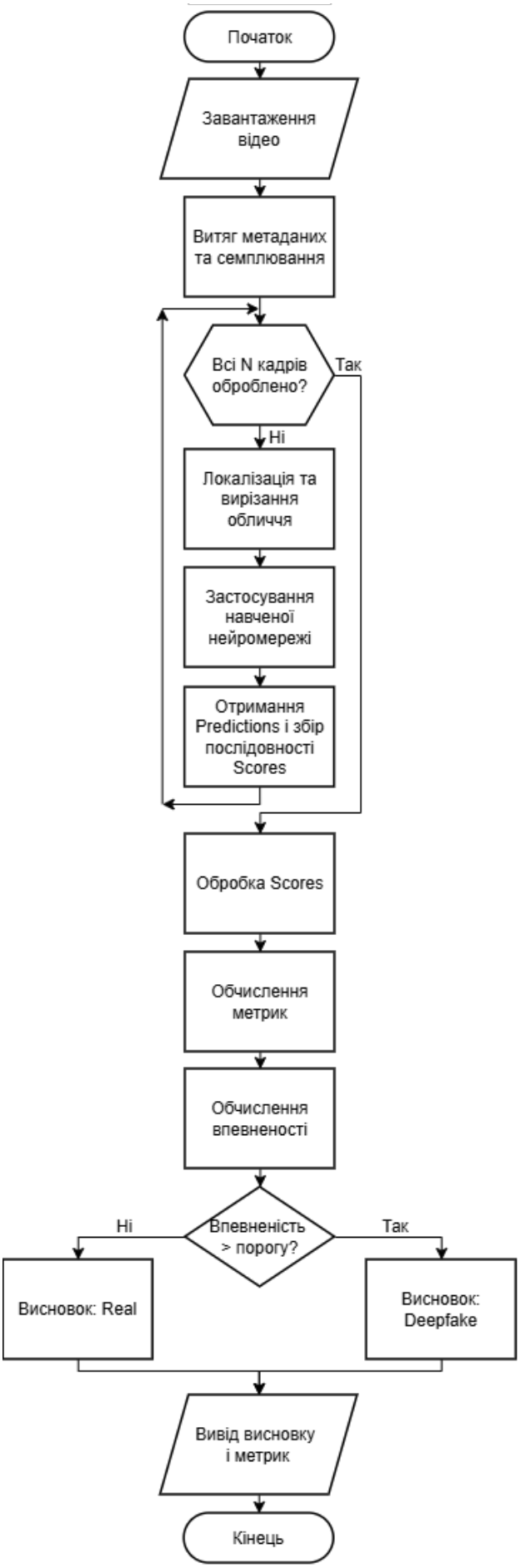


Рис. 3.8. Схема алгоритму *flowchart* процесу аналізу відео

Етап 1: Попередня обробка та вибірка кадрів

Коли користувач завантажує відеофайл для аналізу, система розпочинає роботу з відеопроцесора. Насамперед, вона витягує метадані (роздільна здатність, частота кадрів, тривалість), щоб переконатися, що файл коректний. Оскільки відеофайли можуть містити тисячі кадрів, а аналіз кожного з них є надзвичайно ресурсомістким, система застосовує стратегію вибіркового аналізу (*Sampling*). Замість того, щоб обробляти всі кадри, вона обирає фіксовану, репрезентативну кількість кадрів (наприклад, 32 або 64), рівномірно розподіляючи їх по всій довжині відео. Це забезпечує як високу швидкість обробки, так і покриття всіх ключових сцен відео.

Для кожного вибраного кадру система виконує локалізацію обличчя. Оскільки *deepfakes* зазвичай створюються шляхом накладання згенерованого обличчя на тіло, лише ділянка обличчя є інформативною. Використовуючи потужний детектор (наприклад, *MTCNN*), система знаходить точні координати обличчя, вирізає його із невеликим відступом (це важливо для захоплення артефактів на контурах) і масштабує до стандартного розміру, необхідного для нейромережі (наприклад, 224 × 224 пікселі). Якщо в кадрі виявлено кілька обличь, аналізується або найбільш помітне, або всі одночасно, щоб потім агрегувати результати.

Етап 2: Класифікація кадрів та отримання "сирих" оцінок

На цьому етапі до роботи підключається вже навчена нейронна мережа – класифікатор *deepfake*. Кожне підготовлене зображення обличчя подається на вхід моделі. На виході модель генерує не просто відповідь "фейк" чи "реальне", а числову ймовірність (*Prediction Score*), яке ми трактуємо як впевненість у тому, що обличчя є фейковим (від 0.0 до 1.0). Наприклад, *Score* 0.1 означає, що модель на 10% впевнена, що це фейк (і на 90% – що реальне), а *Score* 0.9 означає високу впевненість у підробці.

В результаті цього етапу ми отримуємо послідовність числових оцінок – один ряд чисел для всього відео: $S_1, S_2, S_3, \ldots, S_n$, де S_i – оцінка і-го кадру. Це і є "сирі" дані, на основі яких буде проводитися фінальний аналіз.

Етап 3: Темпоральний та Статистичний аналіз

Це найбільш важливий та інноваційний етап, який дозволяє відрізнити справжнє відео від глибокої підробки. В основі лежить принцип: справжнє відео є консистентним, а *deepfake*, згенерований незалежно для кожного кадру, зазвичай має порушення плавності та стабільності між сусідніми кадрами. Спочатку проводиться Статистичний аналіз. Обчислюємо кілька ключових показників для всієї послідовності оцінок [$S_1, S_2, S_3, \ldots, S_n$]:

– Дисперсія (*Variance*): Показує, наскільки сильно "розкидані" оцінки. Якщо модель для справжнього відео постійно дає 0.1, дисперсія низька. Якщо ж для фейку оцінки стрибають між 0.2, 0.8, 0.3, дисперсія буде високою.

– Частота Стрибків (*Jump Rate*): Підраховується, як часто відбувається різка зміна оцінки між сусідніми кадрами (наприклад, зміна на 0.5 і більше). *Deepfakes* мають набагато вищий *Jump Rate*.

– Ентропія (*Entropy*): Міра невизначеності. Висока ентропія означає, що модель постійно вагається між "реальним" і "фейком", що є характерною ознакою нестабільної генерації.

Далі долучається Темпоральний аналізатор. Він шукає патерни у часі, зокрема:

– Згладжування (*Smoothing*): Послідовність оцінок трохи згладжується, щоб виявити загальний тренд (наприклад, модель поступово стає більш впевненою у підробці).

– Виявлення Точок Зміни (*Change Point Detection*): Знаходяться кадри, після яких оцінки різко і надовго змінюються. Наприклад, якщо в середині відео обличчя починає "мерехтіти" або з'являється виражений артефакт.

Етап 4: Фінальне рішення та формування звіту

На основі всіх отриманих метрик система обчислює Комплексну Оцінку Стабільності (*Stability Score*). Ця оцінка агрегує Дисперсію, *Jump Rate* та Ентропію в єдине число, що відображає ступінь консистентності відео.

Оцінка стабільності використовується для коригування середньої впевненості моделі. Якщо середня "сира" оцінка показує, що відео, ймовірно, є

фейком (наприклад, 0.7), але *Stability Score* дуже низький, це підтверджує, що нестабільність є артефактом підробки, і фінальна впевненість у фейку зростає. І навпаки, якщо середня оцінка висока, але стабільність ідеальна (як у справжньому відео), система може знизити фінальну впевненість у фейку, виключаючи "хибні тривоги". Цей механізм призводить до отримання Фінальної Скоригованої Впевненості (*Adjusted Confidence*).

На основі цього числа виконується остаточний висновок:

– Якщо *Adjusted Confidence* нижче певного порогу (наприклад, 0.5) – відео оголошується Справжнім.

– Якщо *Adjusted Confidence* вище порогу – відео оголошується *Deepfake*.

Нарешті, система формує Візуальний Звіт. Користувач бачить фінальний висновок, а також детальну інформацію: графік зміни оцінок моделі по кадрах (де видно стрибки), ключові статистичні метрики та причини, які призвели до висновку (наприклад, "Високий *Jump Rate* 25%" або "Низький *Stability Score* 0.3"). Таким чином, система не лише дає відповідь, але й пояснює її, підвищуючи довіру до результатів.

3.6. Інструменти, обрані для розробки інтелектуальної системи

Для реалізації інтелектуальної системи виявлення *deepfakes* було обрано стек технологій, який забезпечує оптимальний баланс між високою продуктивністю обчислень, гнучкістю розробки алгоритмів глибокого навчання та зручністю фінальної візуалізації результатів. Ядром усієї системи є мова програмування *Python*, яка є загальновизнаним стандартом у галузі машинного навчання завдяки її величезній екосистемі бібліотек, що значно прискорюють розробку.

Ядро глибокого навчання та моделювання: *PyTorch*

Основним фреймворком для створення, навчання та розгортання нейронної мережі був обраний *PyTorch*.

Переваги *PyTorch* для проєкту:

– Динамічний граф обчислень: На відміну від статичних фреймворків, *PyTorch* використовує динамічний граф. Це значно полегшує налагодження, експериментування та реалізацію складних структур, таких як інтеграція додаткових шарів *Attention Block* (шар уваги), що був доданий до класифікатора для підвищення його фокусу на критичних ділянках обличчя (очі, рот, контури).

– Ефективність навчання: *PyTorch* надає передові інструменти для оптимізації процесу тренування, включаючи технологію *AMP*, яка дозволяє використовувати обчислення з меншою точністю (*FP16*), суттєво прискорюючи навчання та зменшуючи споживання відеопам’яті без втрати точності.

– Екосистема *TIMM*: Для швидкого доступу до сучасних архітектур було використано бібліотеку *timm* (*PyTorch Image Models*). Саме завдяки їй було легко інтегровано архітектуру *EfficientNet-B0*. Цей вибір був обумовлений його оптимальним співвідношенням між точністю та обчислювальною швидкістю, що важливо для розгортання в реальних умовах.

Передобробка даних, аугментація та детекція обличчя

Якість вхідних даних має вирішальне значення, особливо у випадку відеоаналізу. Тому було інтегровано спеціалізовані бібліотеки для роботи з мультимедіа та зображеннями.

OpenCV (*cv2*): Ця потужна бібліотека програмного забезпечення з відкритим кодом для комп'ютерного зору та машинного навчання. Була створена для забезпечення спільної інфраструктури для програм комп'ютерного зору та пришвидшення використання машинного сприйняття в комерційних продуктах. Будучи ліцензованим продуктом *Apache* 2, *OpenCV* спрощує використання та модифікацію коду для підприємств.

Бібліотека містить понад 2500 оптимізованих алгоритмів, включаючи повний набір як класичних, так і найсучасніших алгоритмів комп'ютерного зору та машинного навчання. Ці алгоритми можна використовувати для виявлення та розпізнавання облич, ідентифікації об'єктів, класифікації дій людини у відео, відстеження рухів камери, відстеження рухомих об'єктів, вилучення *3D*-моделей об'єктів, створення *3D*-хмар точок зі стереокамер, зшивання зображень для

створення зображення високої роздільної здатності цілої сцени, пошуку подібних зображень з бази даних зображень, видалення ефекту червоних очей із зображень, зроблених за допомогою спалаху, відстеження рухів очей, розпізнавання пейзажів та встановлення маркерів для накладання на них доповненої реальності тощо. *OpenCV* має спільноту користувачів, що налічує сотні тисяч, а оціночна кількість щомісячних завантажень перевищує 40 мільйонів. Бібліотека широко використовується компаніями, дослідницькими групами, аматорами та урядовими органами, такими як *NASA* [24].

Детектор обличчя *MTCNN*: Детектор обличчя *MTCNN*: Для гарантованого аналізу обличчя, а не фону, було інтегровано детектор *MTCNN (Multi-task Cascaded Convolutional Networks)*. Це метод розпізнавання та вирівнювання облич на основі глибокого навчання, який використовує каскадну серію згорткових нейронних мереж (*CNN*) для виявлення та локалізації облич на цифрових зображеннях або відео [25]. Алгоритм здатний виявляти обличчя різного масштабу та орієнтації, а також є стійким до змін умов освітлення, виразів обличчя та перекриттів.

Алгоритм *MTCNN* складається з трьох основних етапів: мережа пропозицій (*P-Net*), мережа уточнення (*R-Net*) та вихідна мережа (*O-Net*).

Мережа пропозицій (*P-Net*): Першим етапом алгоритму *MTCNN* є *P-Net*, яка генерує набір кандидатів-обмежувальних рамок, що можуть містити обличчя. *P-Net* бере вхідне зображення та застосовує серію згорткових фільтрів для генерації набору карт ознак. Ці карти ознак потім обробляються набором повністю зв'язаних шарів для прогнозування ймовірності присутності обличчя в кожній області зображення. *P-Net* також регресує координати обмежувальної рамки навколо виявленого обличчя.

Мережа уточнення (*R-Net*): Другим етапом алгоритму *MTCNN* є *R-Net*, яка уточнює кандидати-обмежувальні рамки, згенеровані *P-Net*. *R-Net* бере кандидати-обмежувальні рамки та обрізає відповідні області вхідного зображення. Ці обрізані області потім змінюються на фіксований розмір і пропускаються через серію згорткових та повністю зв'язаних шарів, щоб класифікувати кожну обмежувальну рамку як обличчя або не-обличчя.

Вихідна мережа (*O*-мережа): Заключним етапом алгоритму *MTCNN* є *O*-мережа, яка додатково уточнює обмежувальні рамки та витягує орієнтири обличчя. *O*-мережа бере уточнені обмежувальні рамки з *R*-мережі та обрізає відповідні області вхідного зображення. Ці обрізані області потім змінюються на фіксований розмір і пропускаються через серію згорткових та повністю зв'язаних шарів, щоб класифікувати кожну обмежувальну рамку як обличчя або не-обличчя.

Albumentations: Для реалізації просунутої аугментації даних на етапі навчання була обрана бібліотека *Albumentations*. Це швидка та гнучка бібліотека для покращення зображень. Незалежно від того, чи працюєте ви над класифікацією, сегментацією, виявленням об'єктів чи іншими завданнями комп'ютерного зору, Albumentations надає комплексний набір перетворень та потужну платформу конвеєра [26]. Її перевага полягає у спеціалізованих трансформаціях, необхідних саме для *deepfake*-детекції, зокрема, імітація артефактів стиснення *JPEG* та різні види шуму та розмиття. Це вчить модель ігнорувати неважливі деталі та фокусуватися на справжніх артефактах генерації, роблячи її стійкою до різних умов зйомки та поширення відео в мережі.

Аналіз часової послідовності та прийняття рішення

Унікальність цієї системи полягає у багатокадровому статистичному аналізі, який вимагав використання наукових бібліотек для роботи з числовими даними та сигналами.

NumPy: Як основа для всіх числових обчислень, *NumPy* використовується для ефективного зберігання та маніпулювання масивами даних, включаючи послідовності *Prediction Scores* (оцінок впевненості моделі), які отримуються для кожного кадру.

SciPy: Ця бібліотека наукових обчислень була необхідна для реалізації складних статистичних та темпоральних функцій, зокрема для: обчислення Ентропії Шеннона (*Shannon Entropy*) та Дисперсії (*Variance*) – ключових показників хаотичності та нестабільності; реалізації алгоритмів згладжування (наприклад, *Savitzky-Golay filter*), які використовуються в модулі *Temporal Analyzer* для виявлення *Change Points* (точок різких змін) і неприродних патернів у часі.

Scikit-learn: Цей стандартний набір інструментів машинного навчання був використаний у модулі metrics для обчислення фінальних показників якості (*F1-score, ROC-AUC, Precision, Recall*) та для визначення оптимального порогу класифікації.

Вебзастосунок та візуалізація: *Streamlit*

Для забезпечення зручного користувацького інтерфейсу та демонстрації роботи системи було обрано фреймворк *Streamlit*.

Переваги *Streamlit* для проєкту:

– Швидке прототипування: *Streamlit* дозволяє швидко перетворити *Python*-скрипти на інтерактивний вебзастосунок без необхідності знання веброзробки (*HTML, CSS, JavaScript*). Це значно прискорило етап розгортання та тестування.

– Інтерактивна візуалізація: Фреймворк легко інтегрує графіки та діаграми, що дозволило візуалізувати складні результати аналізу: графік *Temporal Confidence Profile* (Часовий профіль впевненості), який показує динаміку впевненості моделі по кадрах; індикатори стабільності (*Stability Gauge*) та гістограми стрибків (*Jump Analysis*).

– Прозорість результатів: Завдяки *Streamlit*, користувач отримує не просто фінальний вердикт, а детальний, аргументований звіт з усіма ключовими метриками та порівнянням з еталонними значеннями, що підвищує довіру до інтелектуальної системи.

Таким чином, комбінація *PyTorch* для обчислювального ядра, *OpenCV* та *Albumentations* для ефективної обробки даних, *SciPy/NumPy* для наукового аналізу часової динаміки та *Streamlit* для користувацького інтерфейсу забезпечила створення комплексної, продуктивної та прозорої системи для виявлення *deepfake* відео.

3.7. Висновки до розділу

У третьому розділі детально розглянуто процес розроблення системи для виявлення *deepfake*-відео. Основну увагу приділено вибору підходу до навчання моделі, розробці архітектури системи, структурі даних для навчання, алгоритму навчання, а також реалізації процесу виявлення фейкових відео.

Для навчання моделі було обрано підхід *Transfer Learning* із використанням попередньо навченої архітектури *EfficientNet-B0*. Це рішення дозволило суттєво скоротити час навчання, ефективно використовувати обмежену кількість даних і досягти високих показників точності. Додатково застосовано стратегію *Fine-tuning*, яка адаптувала модель до специфічних особливостей *deepfake*-зображень. Використання прийомів аугментації, таких як *Mixup*, зміна яскравості, імітація артефактів *JPEG*, допомогло збільшити стійкість моделі до різноманітних умов.

Розроблена система побудована за принципами модульної багатошарової архітектури. Кожен модуль відповідає за конкретну функцію, що значно спрощує тестування, підтримку та масштабування системи. Такий підхід дозволяє легко додавати нові компоненти, модифікувати існуючі алгоритми та забезпечує високу гнучкість у розробці.

Для навчання використовувався збалансований датасет із реальними та синтетичними зображеннями. Він був структурований у навчальну, валідаційну та тестову вибірки, що дозволило забезпечити об’єктивність навчання та оцінки моделі. Завдяки правильній структурі даних та попередній обробці (виділення та нормалізація обличь) модель змогла зосередитися на аналізі ключових характеристик *deepfake*.

Навчання моделі включало багатоетапний процес, починаючи з підготовки даних і закінчуючи використанням сучасних методів оптимізації, таких як *AdamW*, автоматична змішана точність (*AMP*) і планування швидкості навчання. Модель досягла високих результатів, зокрема точності 97,5% на валідаційній вибірці, що підтверджує ефективність розробленого підходу.

Процес виявлення *deepfake*-відео реалізовано як послідовність етапів: вибірка кадрів, локалізація облич, класифікація кожного кадру та статистичний аналіз часових закономірностей. Використання метрик, таких як дисперсія, ентропія та частота стрибків, дозволяє системі не лише визначати, чи є відео фейковим, а й пояснювати своє рішення.

Таким чином, у цьому розділі створено повноцінну систему для детекції *deepfake*-відео, яка поєднує сучасні методи машинного навчання, ефективну архітектуру та надійний алгоритм аналізу. Отримані результати підтверджують здатність системи виявляти підробки з високою точністю та надійністю.

РОЗДІЛ 4
ПРОТОТИП РОЗРОБЛЕНОЇ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ВИЯВДЕННЯ *DEEPFAKES* У ВІДЕО

4.1. Огляд прототипу

Робота з системою починається з головного екрана вебзастосунку «*Deepfake Detector*» (рис. 4.1). Користувач потрапляє в інтерфейс, розділений на дві частини: ліворуч розташована панель налаштувань, а по центру – основна робоча зона. На бічній панелі користувач може перевірити параметри системи, зокрема використовуваний пристрій (у даному випадку відображається «*CPU*»), налаштувати кількість кадрів для аналізу, обрати метод вибірки та задати поріг чутливості до стрибків.

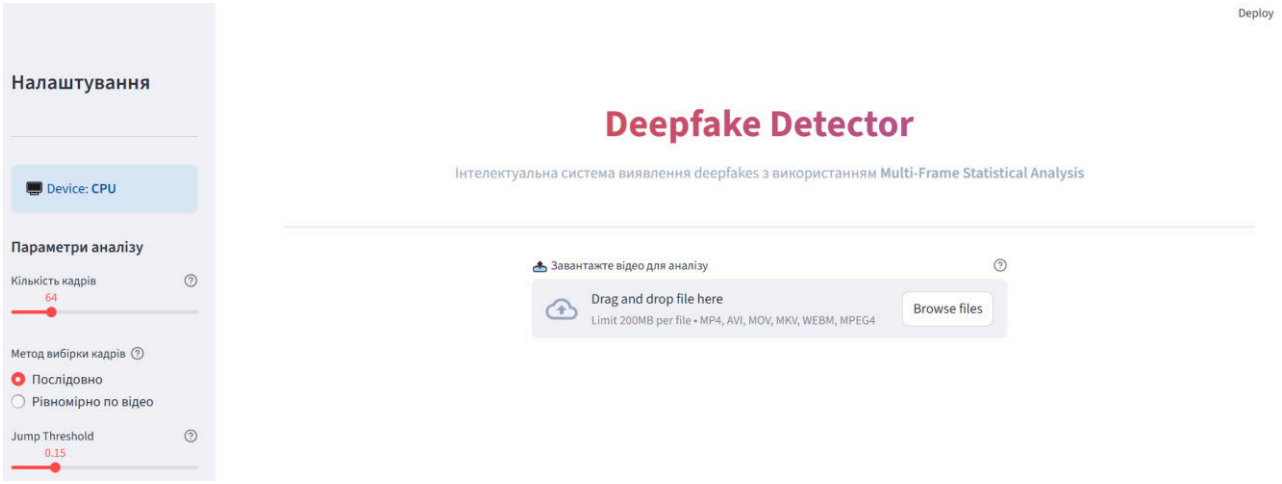


Рис. 4.1. Головний екран вебзастосунку «*Deepfake Detector*»

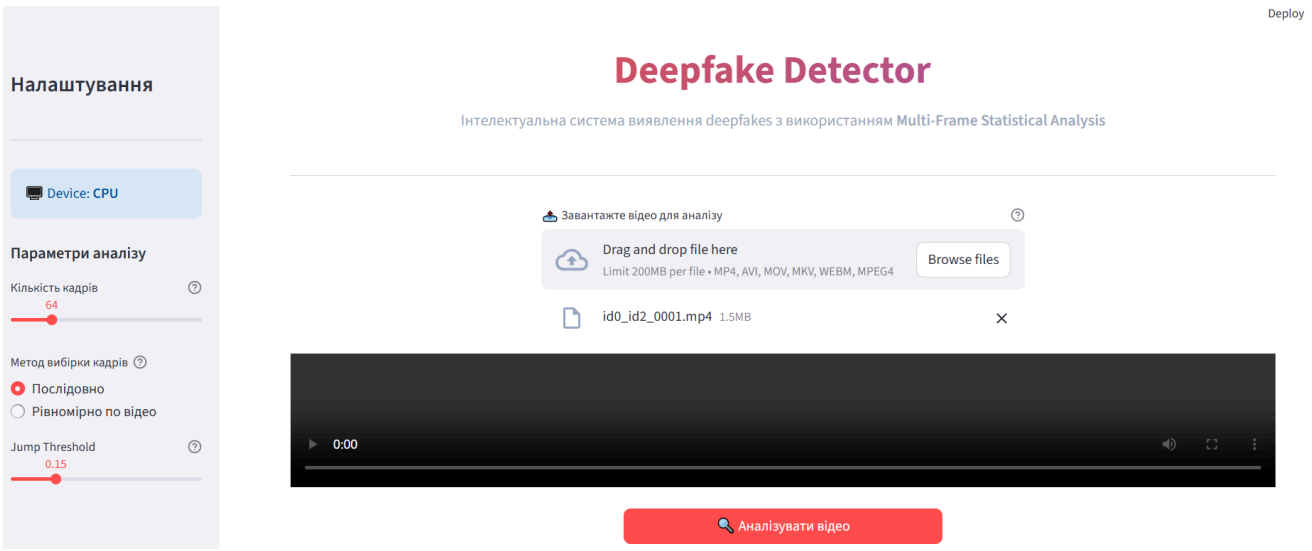


Рис. 4.2. Завантаження відео

Процес аналізу ініціюється завантаженням підозрілого медіафайлу. Користувач перетягує відеофайл у спеціальне поле «*Drag and drop file here*» або обирає його через провідник. Після завантаження на екрані з'являється відеоплеєр, що дозволяє переглянути файл перед перевіркою, та велика червона кнопка «Аналізувати відео» (рис 4.2).

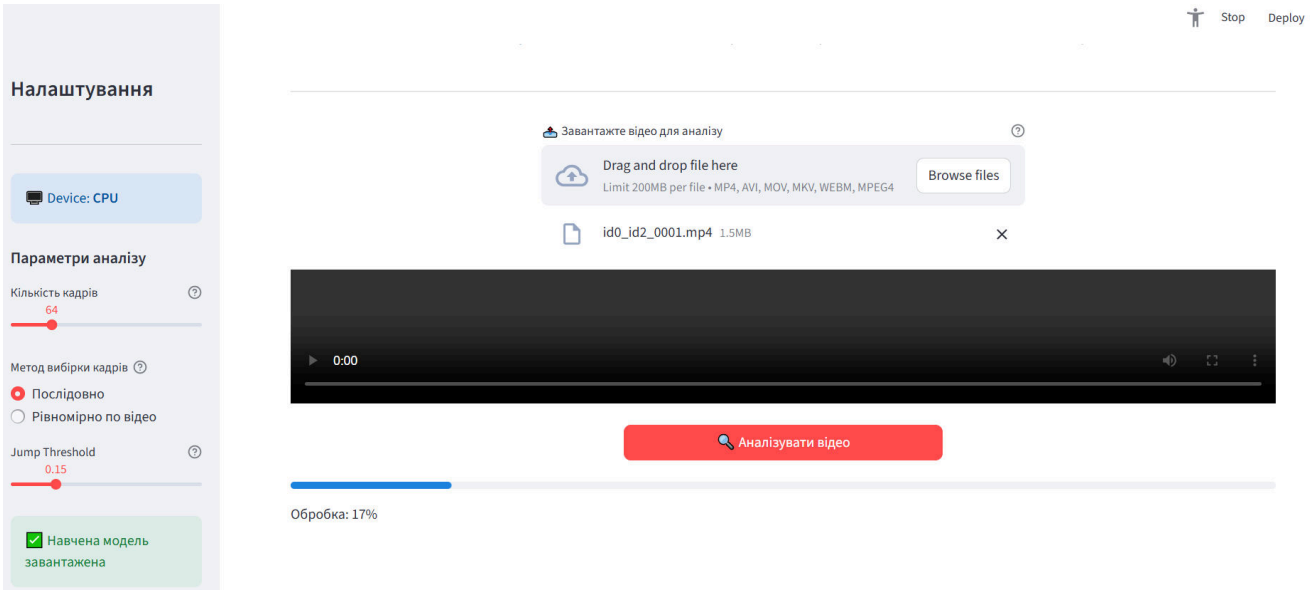


Рис. 4.3. Процес аналізу завантаженого відеоматеріалу

Після початку аналізу у бічній панелі з’являється інформація чи вдало завантажена модель. І користувач спостерігає за прогресом через індикатор завантаження, який інформує про стан обробки у відсотках (рис. 4.3).

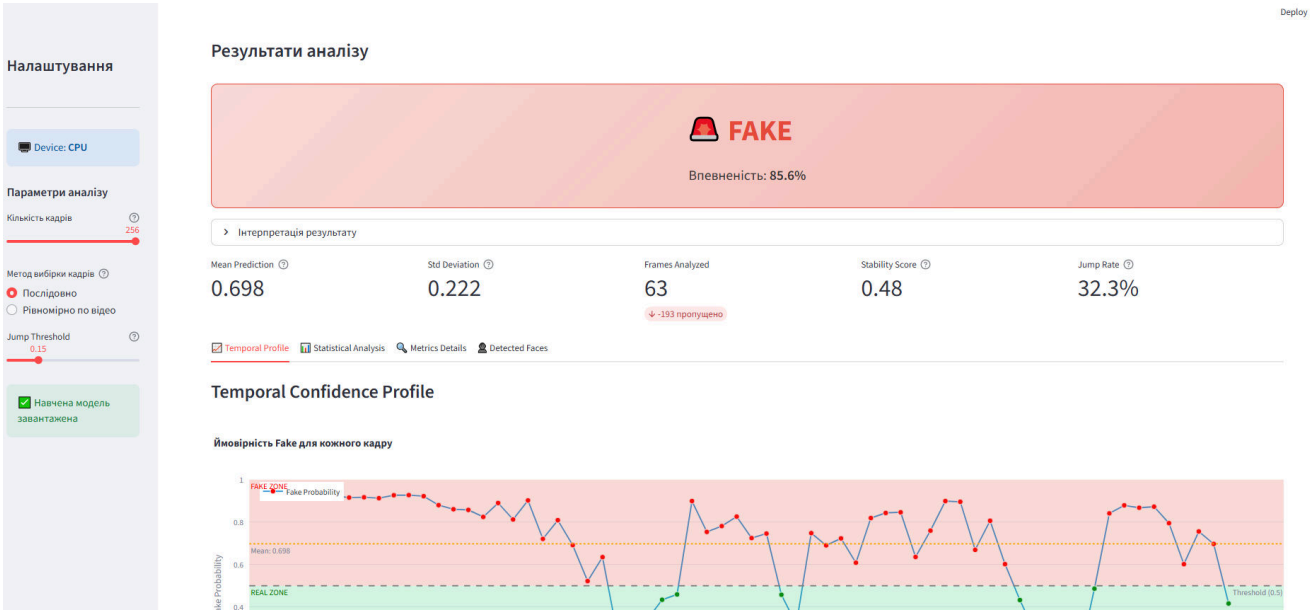


Рис. 4.4. Результат аналізу *deepfake*-відео

Після завершення аналізу інтерфейс миттєво оновлюється, демонструючи фінальний вердикт у вигляді великого кольорового блоку (рис. 4.4). У даному випадку користувач бачить яскраво-червоний банер із написом «*FAKE*» та рівнем впевненості 85.6%. Це означає, що система виявила сукупність ознак, які з дуже високою ймовірністю вказують на штучне походження обличчя у відео. Усі подальші дані надані для того, щоб користувач міг зрозуміти, які аспекти відеоматеріалу повпливали на результат.

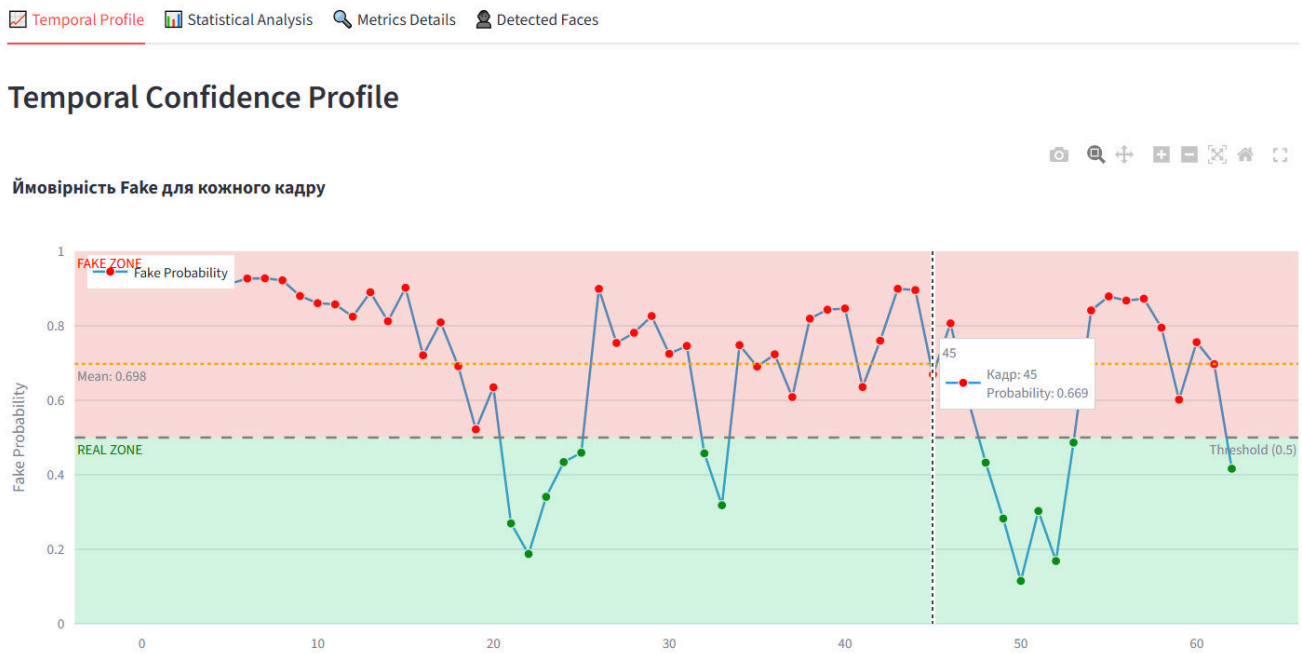


Рис. 4.5. Графік часової зміни впевненості моделі під час обробки *deepfake*-відео

Наступним елементом, який бачить користувач, є інтерактивний графік «*Temporal Confidence Profile*» (Часовий профіль впевненості), який візуалізує динаміку аналізу в часі (рис. 4.5).

Цей графік дозволяє користувачеві зазирнути "всередину" процесу прийняття рішення. По горизонтальній осі відкладено номер кадру відео, а по вертикальній – ймовірність того, що цей кадр є фейковим (від 0 до 1).

На наведеному прикладі видно, що графік є «ламаним» і нестабільним: значення різко стрибають між червоною та зеленою зонами. Така поведінка називається «мерехтінням» (*flickering*) і є характерною ознакою *deepfake*-генерації, оскільки нейромережа не може стабільно відмальовувати обличчя на кожному кадрі.

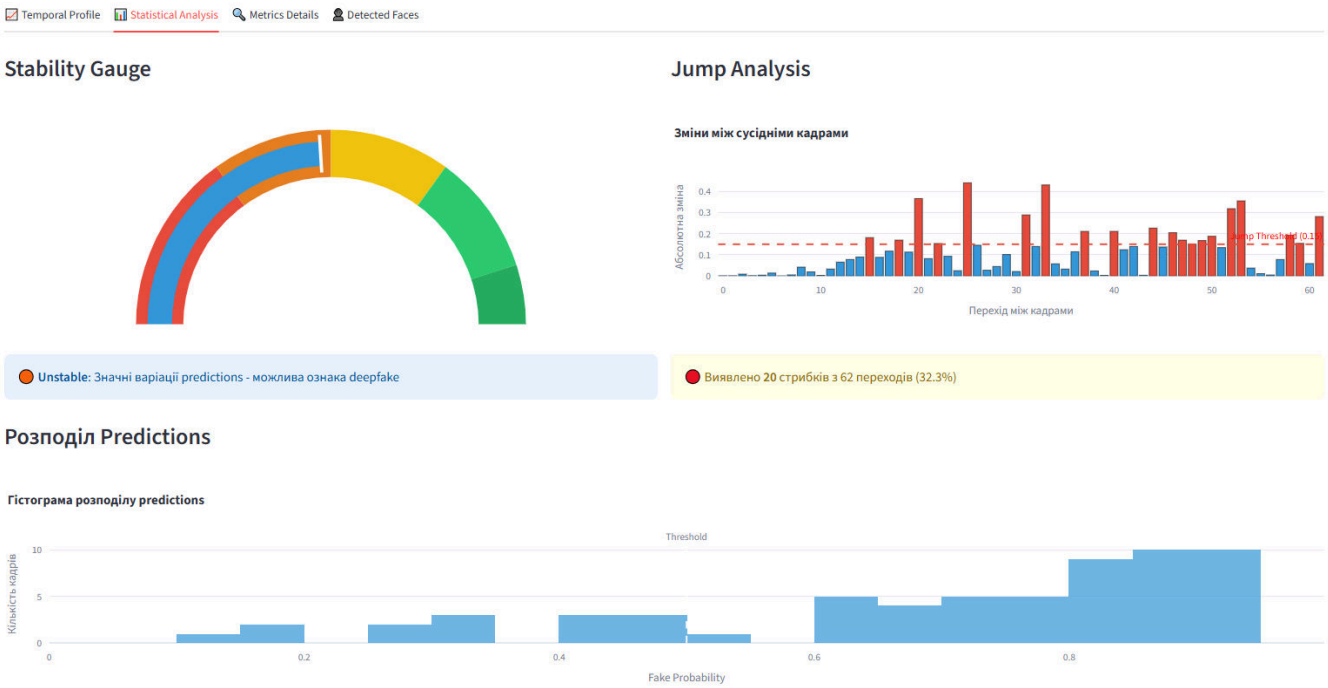


Рис. 4.6. Індикатори стабільності та гістограма стрибків

У вкладці «*Statistical Analysis*» користувач може переглянути інструменти візуалізації статистичних аномалій: «*Stability Gauge*» (Вимірювач стабільності), «*Jump Analysis*» (Аналіз стрибків) та «*Розподіл Predictions*» (рис. 4.6).

Індикатор стабільності виконаний у вигляді спідометра. У даному випадку стрілка вказує на помаранчевий сектор «*Unstable*» (Нестабільно). Це простий візуальний сигнал для користувача, що відео має неприродну динаміку.

Поруч розташована гістограма «*Jump Analysis*». Стовпчики показують силу зміни обличчя між сусідніми кадрами. Червона пунктирна лінія – це поріг допустимих змін. Користувач може побачити, що багато стовпчиків перетинають цю червону лінію. Система повідомляє: «Виявлено 20 стрибків», що складає 32.3% від усіх переходів. Це означає, що у третині випадків обличчя змінювалося надто різко, що фізично неможливо для живої людини.

На сторінці є гістограма «Розподіл *Predictions*», яка показує частоту появи різних оцінок впевненості моделі протягом усього відео. По горизонтальній осі відкладено ймовірність підробки (від 0 до 1), а по вертикальній – кількість кадрів, які отримали таку оцінку. На наведеному прикладі чітко видно концентрацію синіх стовпчиків у правій частині графіка (діапазон 0.6 – 1.0). Це візуально підтверджує, що для переважної більшості проаналізованих кадрів нейромережа визначила високу ймовірність фейку, тоді як низькі значення (зліва) зустрічаються значно рідше.

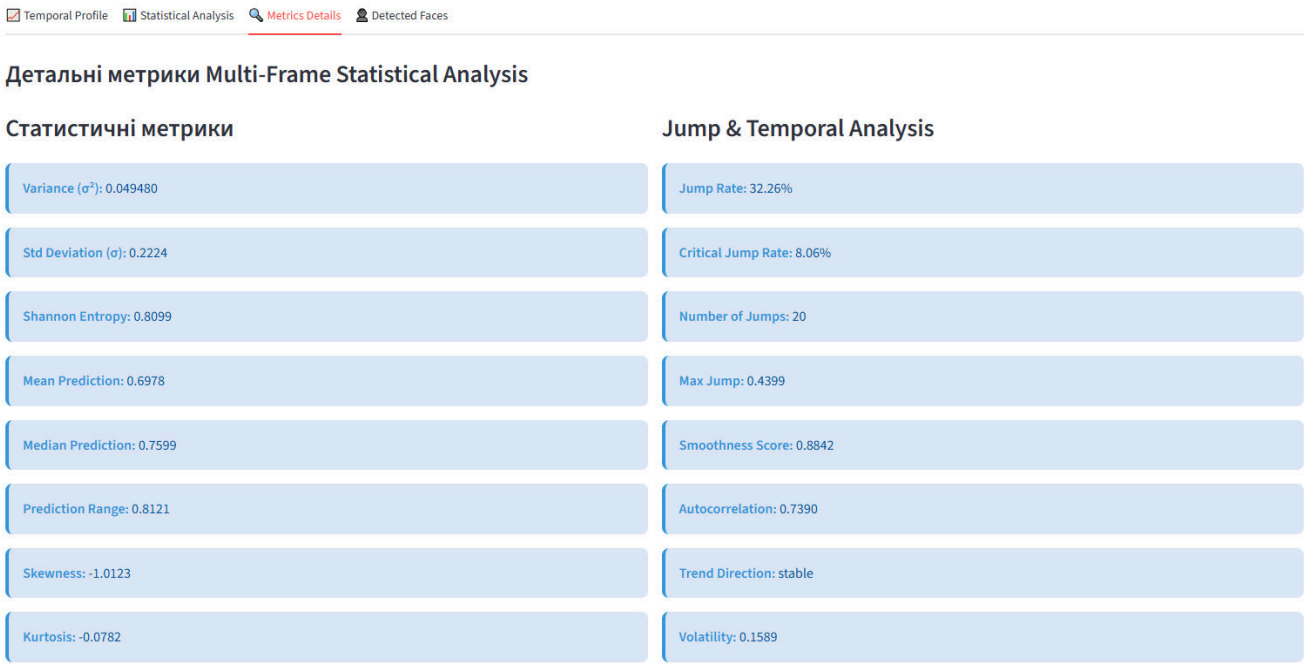


Рис. 4.7. Детальні метрики статистичного та темпорального аналізу

Для фахівців та детального аудиту результатів система надає розгорнутий блок «Детальні метрики» (рис. 4.7). У цьому розділі зібрані точні математичні показники, розраховані модулями статистичного аналізу. Зліва у блоці

«Статистичні метрики» наведено базові параметри розподілу оцінок: дисперсію (*Variance*), стандартне відхилення (*Std Deviation*) та ентропію Шеннона (*Shannon Entropy*), яка в даному випадку становить 0.8099, вказуючи на високий рівень хаотичності даних. Праворуч, у блоці «*Jump & Temporal Analysis*», відображаються специфічні показники динаміки: відсоток стрибків (*Jump Rate* – 32.26%), кількість критичних змін та оцінка плавності відео. Ці цифри слугують «сирими» даними, які обґрунтовують виявлені аномалії.

Інтерпретація метрик

Метрика	Real (справжнє)	Fake (deepfake)	Поточне значення
Variance	< 0.01	> 0.05	0.0495
Std Deviation	< 0.1	> 0.2	0.2224
Entropy	< 0.5	> 0.7	0.8099
Jump Rate	< 10%	> 25%	32.3%
Smoothness	> 0.9	< 0.7	0.8842
Stability Score	> 0.7	< 0.5	0.4808

Рис. 4.8. Таблиця порівняння та інтерпретації метрик

Щоб зробити технічні дані зрозумілими для кінцевого користувача, у нижній частині сторінки розміщено таблицю «Інтерпретація метрик» (рис. 4.8).

Ця таблиця виступає ключовим інструментом для верифікації рішення системи. Вона порівнює отримані для поточного відео показники («Поточне значення») з еталонними діапазонами, характерними для справжніх відео («*Real*») та діпфейків («*Fake*»). Наприклад, користувач може наочно побачити, що поточне значення *Jump Rate* (32.3%) значно перевищує норму для реального відео (< 10%) і потрапляє в категорію фейків (> 25%). Аналогічно, низький *Stability Score* (0.4808) не відповідає критеріям справжнього відео (> 0.7). Таке зіставлення дозволяє

переконатися в об'єктивності фінального вердикту, спираючись на чіткі порогові значення.

Є також інші можливі варіанти результату обробки відео. Наприклад, аналіз справжнього відео показує (рис. 4.9), що на деяких кадрах система сумнівається, але переважна їх кількість визначається реальною. Але при цьому впевненість складає лише 68%, бо інші показники не є достатньо однозначними. Тому користувач може переглянути числові метрики та графіки, щоб зрозуміти, які моменти є причиною невизначеності.

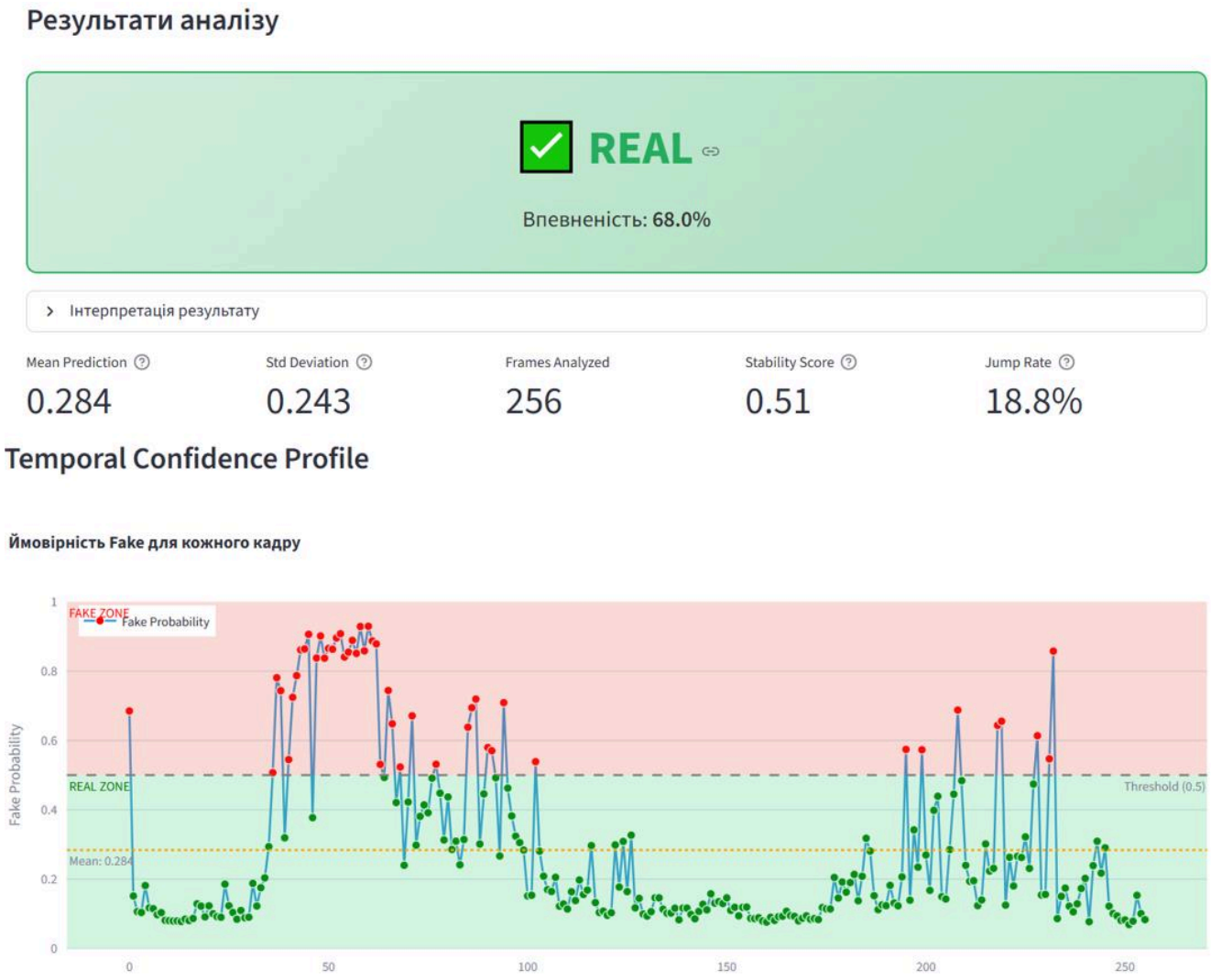
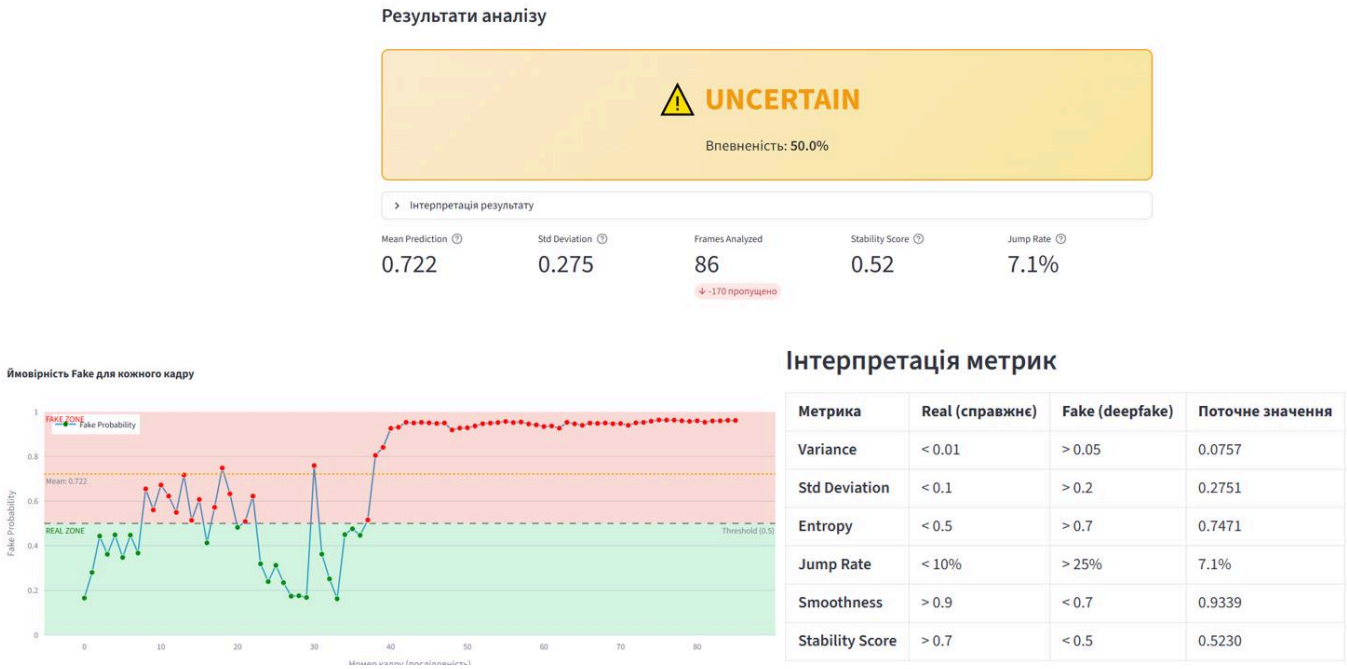


Рис. 4.9. Результат аналізу справжнього відео

Також можливий варіант, коли система не може визначити точно, чи є наданий відеоматеріал реальним. Наприклад, варіант обробки реального відео, але

дуже поганою якістю (рис. 4.10). Відео дуже непослідовне, тому і стабільність середня. По метрикам в такому випадку теж складно оцінити надіслане відео.



Метрика	Real (справжнє)	Fake (deepfake)	Поточне значення
Variance	< 0.01	> 0.05	0.0757
Std Deviation	< 0.1	> 0.2	0.2751
Entropy	< 0.5	> 0.7	0.7471
Jump Rate	< 10%	> 25%	7.1%
Smoothness	> 0.9	< 0.7	0.9339
Stability Score	> 0.7	< 0.5	0.5230

Рис. 4.10. Результат аналізу справжнього відео з поганою якістю

У випадку надсилання відео без наявного на ньому обличчя, виводиться помилка (рис. 4.11).

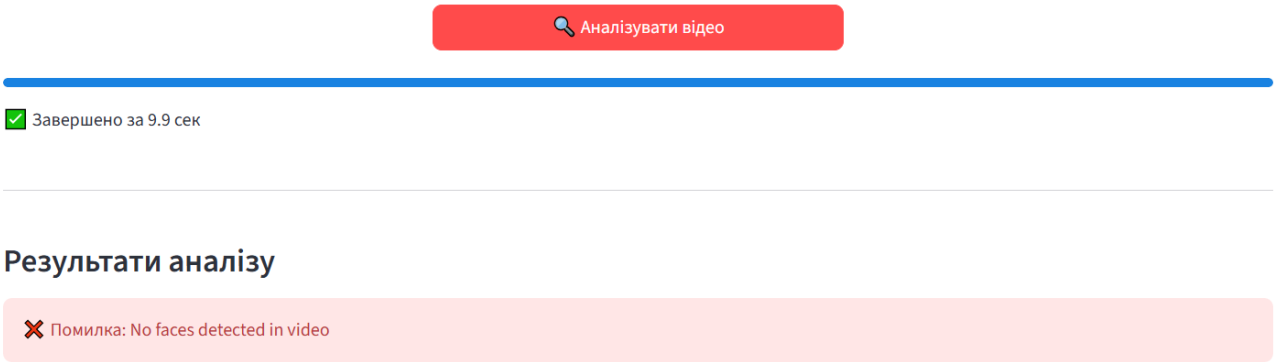


Рис. 4.11. Виведення помилки, якщо розпізнати обличчя на відео не вдалося

Завдяки такому поданню інформації користувач отримує не просто сухий висновок після аналізу надісланого ним відео, а може переглянути повне, аргументоване пояснення причин такого рішення на основі об'єктивних даних.

4.2. Подальші вдосконалення

Розроблена система для автоматизованого виявлення *deepfake*-відео демонструє високу ефективність, точність та зручність використання. Однак, як і будь-яке програмне рішення, вона має потенціал для подальшого вдосконалення. Нижче наведено основні напрями, які можуть бути реалізовані для покращення системи в майбутньому.

1. Підвищення точності моделі

Хоча система вже досягає високих показників точності, існує можливість подальшого вдосконалення нейронної мережі шляхом:

– Розширення навчального датасету: використання більшого та різноманітнішого набору даних, який включатиме нові типи *deepfake*, створені за допомогою сучасних генеративних моделей (наприклад, *StyleGAN3*, *DALL-E* 3 тощо). Це дозволить моделі краще розпізнавати новітні підходи до створення підробок.

– Додавання мультимодальних даних: можливе використання не лише зображень обличь, але й звукових доріжок для аналізу синтетично згенерованих голосів, що часто супроводжують *deepfake*-відео.

– Використання сучасніших архітектур: перехід на більш потужні моделі, зокрема *EfficientNetV2* або *Vision Transformers* (*ViT*), які мають покращені властивості генералізації та здатність працювати з великою кількістю даних.

2. Оптимізація швидкості аналізу

Скорочення часу обробки є важливим аспектом, особливо для аналізу великих відеофайлів. Для цього можна:

– Реалізувати обробку на графічних процесорах (*GPU*): запровадження підтримки сучасних *GPU* для значного прискорення обчислень, особливо для покадрової класифікації.

– Паралельна обробка кадрів: впровадження механізму розподіленої або багатопотокової обробки кадрів для прискорення аналізу відео.

– Динамічна вибірка кадрів: замість рівномірного вибору кадрів по всьому відео, можна розробити алгоритм, який виділяє для аналізу найбільш інформативні частини відео (наприклад, сцени з більшою кількістю обличь або ключові моменти).

3. Покращення інтерфейсу користувача

Хоча система вже має зручний вебінтерфейс, її можна вдосконалити для зручнішої роботи користувачів:

– Деталізовані звіти: додавання інтерактивних елементів, які дозволять користувачу переглядати результати аналізу для кожного кадру, зокрема теплові карти (*heatmaps*), що показують області кадру, які найбільше вплинули на рішення моделі.

– Мобільна версія: розроблення мобільного додатку або адаптація інтерфейсу для зручності використання на смартфонах.

– Мовна локалізація: додавання підтримки кількох мов для розширення аудиторії користувачів.

4. Розширення функціональності

Система може бути розширена для вирішення додаткових задач:

– Детекція інших типів фейкових медіа: додавання підтримки для аналізу синтетично створених фотографій, аудіофайлів чи текстів, створених за допомогою великих мовних моделей.

– Розпізнавання джерела підробки: впровадження функції визначення методу, який використовувався для створення *deepfake* (наприклад, *StyleGAN*, *DeepFaceLab* тощо).

– Аналіз метаданих відео: додавання інструментів для виявлення аномалій у метаданих файлу, які можуть вказувати на маніпуляції.

5. Інтеграція з іншими платформами

Система може бути інтегрована в сторонні сервіси для більш широкого використання:

– Соціальні мережі: інтеграція з платформами, такими як *Facebook*, *Instagram* чи *TikTok*, для автоматичного аналізу контенту, що завантажується користувачами.

– Системи безпеки: використання системи для перевірки відео в корпоративних чи урядових структурах для боротьби з дезінформацією.

8. Зменшення вимог до ресурсів

Наразі система може працювати на *CPU*, але для зменшення обчислювальних вимог варто:

– Оптимізувати модель: використання методів, зокрема *pruning* (підрізання нейронів) або *knowledge distillation* (передача знань від великої моделі до меншої), дозволить зменшити розмір моделі без втрати продуктивності.

– Хмарні сервіси: переміщення обчислювальних процесів у хмару дозволить зменшити вимоги до локальних пристроїв користувача.

Подальші вдосконалення системи спрямовані на підвищення її точності, швидкості, функціональності та зручності використання. Реалізація запропонованих змін дозволить зробити систему ще більш ефективною, адаптивною до нових викликів та доступною для ширшого кола користувачів. Це сприятиме боротьбі з поширенням синтетичних відеоматеріалів та забезпеченню інформаційної безпеки.

4.3. Висновки до розділу

У цьому розділі розглянуто практичну роботу вебзастосунку «*Deepfake Detector*» та перспективи його розвитку. Система поєднує потужність нейронних мереж із зручним інтерфейсом, забезпечуючи комплексний підхід до виявлення підроблених відеоматеріалів.

Робота системи побудована так, щоб користувач розумів, на основі яких даних прийнято рішення. Після завантаження відеофайлу система обробляє його покадрово, виявляючи ознаки штучного походження обличь. Результати

представлені у вигляді зрозумілого висновку з відсотком впевненості та детальною візуалізацією.

Система надає часовий профіль впевненості, який показує динаміку оцінок для кожного кадру та дозволяє виявити характерні аномалії, такі як мерехтіння або різкі стрибки. Статистичні інструменти, включаючи індикатор стабільності, аналіз стрибків та гістограми, допомагають оцінити природність динаміки відео. Детальні метрики надають фахівцям можливість глибокого аудиту, а таблиця інтерпретації пояснює технічні показники доступною мовою.

Система коректно обробляє різні сценарії: впевнено визначає діпфейки та справжні відео, а у неоднозначних випадках чесно повідомляє про невизначеність. Передбачено також обробку помилок, зокрема при відсутності обличчя на відео.

Визначено конкретні напрями вдосконалення системи. Підвищення точності досягається через розширення навчальних даних, впровадження аналізу звукових доріжок та використання сучасніших архітектур. Швидкість роботи покращується через обробку на графічних процесорах та паралелізацію обчислень. Функціональність може бути розширена на інші типи синтетичних медіа та визначення методів створення підробок. Важливим напрямком є інтеграція з соціальними мережами та корпоративними платформами. Реалізація запропонованих удосконалень забезпечить ефективний інструмент для боротьби з дезінформацією в цифровому просторі.

ВИСНОВКИ

У даній кваліфікаційній роботі розроблено інтелектуальну систему для автоматизованого виявлення *deepfake*-відео, яка поєднує сучасні методи глибокого навчання з оригінальним підходом до аналізу відеоматеріалів. Проблема виявлення синтетичних медіа є надзвичайно актуальною в умовах стрімкого розвитку генеративних технологій та їх потенційного використання для поширення дезінформації, маніпуляції громадською думкою та порушення приватності громадян.

У процесі дослідження проведено комплексний аналіз існуючих підходів до детекції *deepfakes*, який виявив, що більшість систем зосереджується виключно на покадровому аналізі візуальних артефактів. Такий підхід має суттєві обмеження, оскільки сучасні генеративні моделі створюють дедалі якісніші підробки з мінімальними візуальними недоліками. Тому запропоновано інноваційне рішення, що поєднує класичну класифікацію окремих кадрів із статистичним аналізом темпоральної послідовності оцінок моделі. Ключова ідея полягає в тому, що навіть високоякісні *deepfakes* демонструють нестабільність характеристик між послідовними кадрами, оскільки генеративні алгоритми обробляють кожен кадр незалежно, не здатні підтримувати абсолютну консистентність протягом всього відео.

Для реалізації системи обрано підхід трансферного навчання з використанням архітектури *EfficientNet-B0*, попередньо навченої на датасеті *ImageNet*. Це рішення дозволило значно скоротити час навчання та ефективно використовувати обмежені обчислювальні ресурси, водночас досягаючи високих показників точності. Модель була навчена на збалансованому датасеті з реальних та синтетичних зображень облич, де застосовувалися сучасні методи аугментації даних, включаючи *Mixup*, імітацію артефактів стиснення та зміни освітлення. Результати навчання продемонстрували точність класифікації на рівні 97,5% на валідаційній вибірці, що підтверджує ефективність обраної архітектури та методології.

Розроблена система має модульну багатошарову архітектуру, яка забезпечує чітке розділення відповідальності між компонентами. Основні функціональні модулі включають компоненти для завантаження та обробки відео, детекції та класифікації обличь, статистичного та темпорального аналізу результатів, а також візуалізації даних. Така організація коду значно спрощує тестування, підтримку та можливість подальшого розширення функціональності системи. Вебінтерфейс реалізовано за допомогою фреймворку *Streamlit*, що дозволило створити зручний та інтуїтивно зрозумілий додаток, доступний через браузер без необхідності встановлення додаткового програмного забезпечення.

Процес виявлення *deepfakes* у системі організовано як послідовність чітко визначених етапів. Спочатку відбувається завантаження відеофайлу та вибірка репрезентативних кадрів, що дозволяє суттєво прискорити обробку без втрати якості аналізу. На кожному кадрі система виконує детекцію обличчя, після чого виділена область подається на вхід нейронної мережі для класифікації. Отримана послідовність оцінок проходить через модулі статистичного та темпорального аналізу, які обчислюють такі показники як дисперсія, ентропія розподілу, частота різких стрибків між сусідніми кадрами та коефіцієнт стабільності. Ці метрики дозволяють системі не лише визначити, чи є відео підробленим, а й пояснити своє рішення, вказавши конкретні аномалії у поведінці послідовності оцінок.

Вебзастосунок надає користувачу повну та зрозумілу інформацію про результати аналізу. Фінальний вердикт представлено у вигляді кольорового індикатора з відсотком впевненості системи, а детальні панелі містять графік часової зміни оцінок, індикатор стабільності, гістограми розподілу та таблицю порівняння метрик з еталонними значеннями для реальних та підроблених відео. Така організація інформації дозволяє як звичайним користувачам швидко зрозуміти висновок системи, так і фахівцям провести детальний аудит результатів для прийняття обґрунтованих рішень.

Тестування системи підтвердило її високу ефективність у виявленні різних типів *deepfakes*. Система впевнено розпізнає як очевидні підробки з яскраво вираженими артефактами, так і складні випадки, де візуальна якість окремих кадрів

є високою, але темпоральна нестабільність видає синтетичне походження матеріалу. У випадках невизначеності, наприклад при аналізі відео дуже низької якості, система чесно повідомляє про неможливість впевненої класифікації та надає всі метрики для самостійної оцінки користувачем.

Визначено перспективні напрями подальшого вдосконалення системи, включаючи розширення навчальних даних для покращення генералізації на нові типи *deepfakes*, додавання аналізу звукових доріжок для виявлення синтетично згенерованих голосів, оптимізацію швидкості обробки через використання *GPU* та паралелізацію обчислень, а також інтеграцію з популярними платформами соціальних мереж для автоматичного моніторингу контенту. Реалізація цих удосконалень дозволить системі ефективніше протидіяти еволюції технологій створення *deepfakes* та забезпечить надійний інструмент для боротьби з цифровою дезінформацією.

Таким чином, розроблена інтелектуальна система виявлення *deepfake*-відео успішно вирішує поставлені завдання, демонструючи високу точність детекції, зручність використання та можливість пояснення прийнятих рішень. Система готова до практичного застосування та може бути корисною для медіаорганізацій, правоохоронних органів, освітніх установ та всіх, хто потребує перевірки автентичності відеоматеріалів у сучасному цифровому середовищі.

СПИСОК БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бойченко С.В., Іванченко О.В. Положення про дипломні роботи (проекти) випускників Національного авіаційного університету. Київ : НАУ, 2024. 63 с.

2. ДСТУ 3008:2015. Звіти у сфері науки і техніки. Структура і правила оформлення. Київ: ДП «УкрНДНЦ», 2015. 24 с.

3. *Stepney E. S., Lally C. Disinformation: sources, spread and impact. UK Parliamet POST*. 25.04.2024. *URL*: *https://researchbriefings.files.parliament.uk/documents/POST-PN-0719/POST-PN-0719.pdf* (дата звернення: 20.09.2025).

4. *The psychological drivers of misinformation belief and its resistance to correction / U. Ecker ma ін. Nature*. 12.01.2022. *URL: https://www.nature.com/articles/s44159-021-00006-y* (дата звернення: 20.09.2025).

5. *Deepfake disruption: A cybersecurity-scale challenge and its far-reaching consequences / M. Steinhart ma ін. Deloitte*. 19.11.2024. *URL: https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/gen-ai-trust-standards.html* (дата звернення: 22.09.2025).

6. *Altuncu E., Franqueira V. N. L., Li S. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. Frontiers.* 04.09.2024. *URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1400024/full* (дата звернення: 22.09.2025).

7. *Deepfakes in Society: Risks and Realities / A. Chaturvedi ma ін. Prerpints.org.* 21.04.2025. *URL: https://www.preprints.org/manuscript/202504.1776/v1?utm_source=chatgpt.com#sec4-preprints-15653* (дата звернення: 25.09.2025).

8. *Afshari N., Mohammadi A. The Legal Implications of Deepfake Technology: Privacy, Defamation, and the Challenge of Regulating Synthetic Media. Legal Studies in Digital Age.* 01.04.2023. *URL: https://jlsda.com/index.php/lsda/article/view/13* (дата звернення: 25.09.2025).

9. *Dabas P. The Impact Of Deepfake Technology: Legal Risks And Regulatory Solutions. Mondaq.* 28.11.2024. *URL: https://www.mondaq.com/india/new-technology/1550822/the-impact-of-deepfake-technology-legal-risks-and-regulatory-solutions* (дата звернення: 25.09.2025).

10. *Deepfake Generation and Detection: A Benchmark and Survey / G. Pei ma ін. Arxiv.* 20.04.2024. *URL: https://arxiv.org/html/2403.17881v3* (дата звернення: 30.09.2025).

11. *Diffusion Models: A Comprehensive Survey of Methods and Applications / L. Yang ma ін. Arxiv.* 02.12.2024. *URL: https://arxiv.org/abs/2209.00796* (дата звернення: 30.09.2025).

12. *Deepfake Detection: A Comprehensive Survey from the Reliability Perspective / T. Wang ma ін. Arxiv.* 13.09.2024. *URL: https://arxiv.org/html/2211.10881v3* (дата звернення: 04.10.2025).

13. Каганець І. *Intel* представила технологію розпізнавання фейкових відео, згенерованих штучним інтелектом. Народний оглядач. 20.11.2022. *URL: https://www.ar25.org/article/intel-predstavyla-tehnologiyu-rozpiznavannya-feykovyh-video-zgenerovanyh-shtuchnym* (дата звернення: 06.10.2025).

14. *Burt T., Horvitz E. New steps to combat disinformation. Microsoft.* 01.09.2020. *URL: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/* (дата звернення: 14.10.2025).

15. *Intel* представила технологію *FakeCatcher* для розпізнавання відео з діпфейками. *Overclockers.ua.* 15.11.2022. *URL: https://www.overclockers.ua/ua/news/hardware/2022-11-15/131768/* (дата звернення: 14.10.2025).

16. *Sensity AI. Detectortools. URL: https://detectortools.ai/tool/sensity-deepfake-detection/* (дата звернення: 15.10.2025).

17. *The Complete Guide on How to Train an AI Model from Scratch. Smartosc.* 22.08.2021. *URL: https://www.smartosc.com/the-complete-guide-on-how-to-train-an-ai-model-from-scratch/* (дата звернення: 20.10.2025).

18. *Murel J., Kavlakoglu E. What is transfer learning?. IBM. URL: https://www.ibm.com/think/topics/transfer-learning* (дата звернення: 21.10.2025).

19. *Bergmann D. What is fine-tuning?. IBM. URL: https://www.ibm.com/think/topics/fine-tuning* (дата звернення: 24.10.2025).

20. *Kundu R., Skelton J., Mukherjee S. Everything you need to know about Few-Shot Learning. DigitalOcean.* 01.08.2025. *URL: https://www.digitalocean.com/community/tutorials/few-shot-learning (дата звернення:* 22.10.2025).

21. *Bergmann D. What is self-supervised learning?. IBM. URL: https://www.ibm.com/think/topics/self-supervised-learning* (дата звернення: 22.10.2025).

22. *efficientnetb0. MathWorks. URL: https://www.mathworks.com/help/deeplearning/ref/efficientnetb0.html* (дата звернення: 24.10.2025).

23. *Karki M. deepfake and real images. Kaggle. URL: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images* (дата звернення: 24.10.2025).

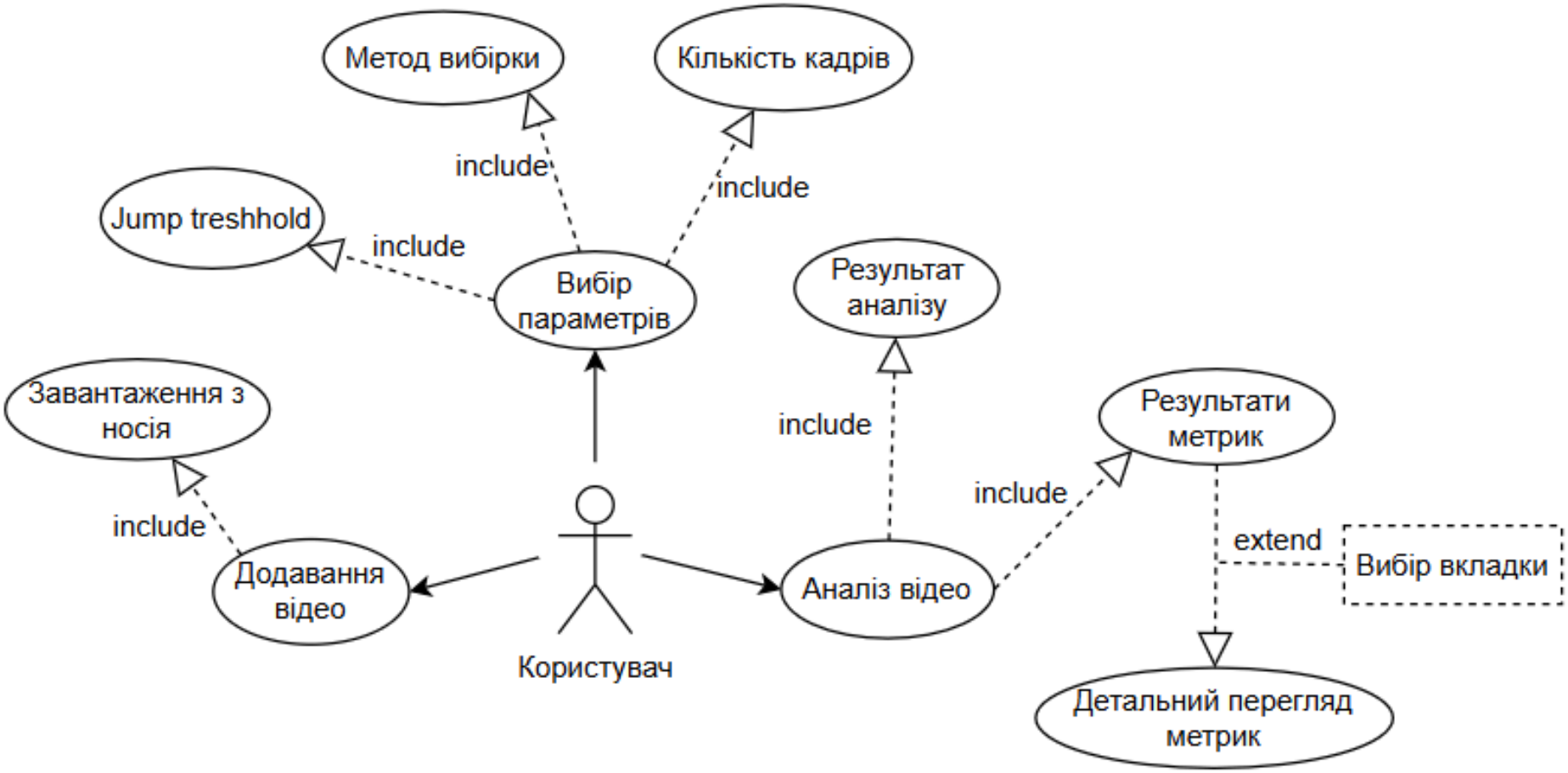
24. *OpenCV is the world's biggest computer vision library.. OpenCV. URL: https://opencv.org/about/* (дата звернення: 24.11.2025).

25. *Jawabreh A. Explore the most advanced deep learning algorithm for face detection. Medium.* 01.04.2023. *URL: https://medium.com/the-modern-scientist/multi-task-cascaded-convolutional-neural-network-mtcnn-a31d88f501c8* (дата звернення: 21.11.2025).

26. *Welcome to Albumentations Documentation!. Albumentations. URL: https://albumentations.ai/docs/* (дата звернення: 28.11.2025).

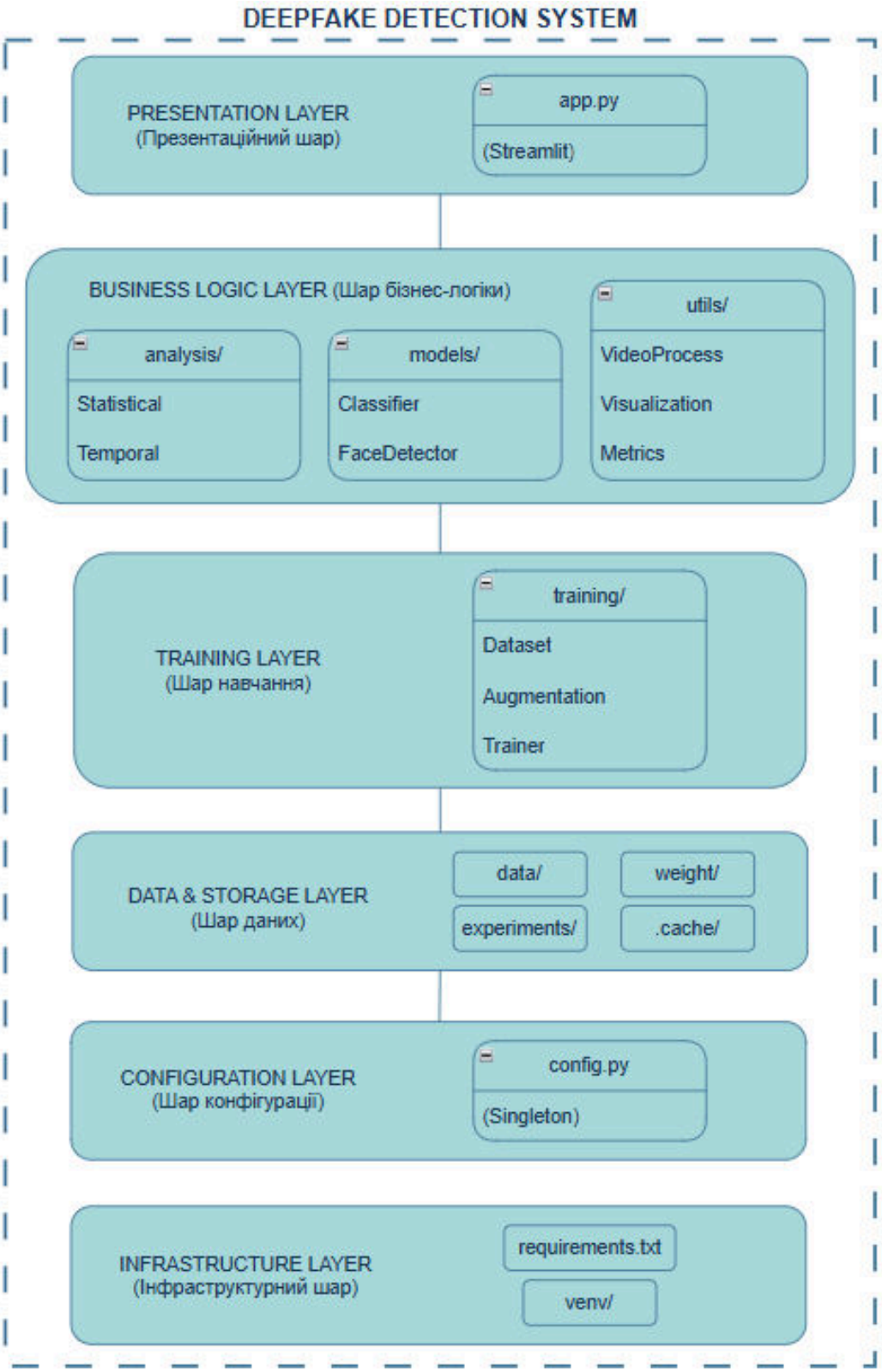
27. Старенька К. О., Супрун О. М. Аналіз проблематики виявлення *deepfake*-відео. Сучасні тенденції розвитку системного програмування, Київ, Україна, 20–21 листоп. 2025. С. 81–82.

Діаграма прецедентів системи виявлення *deepfake*

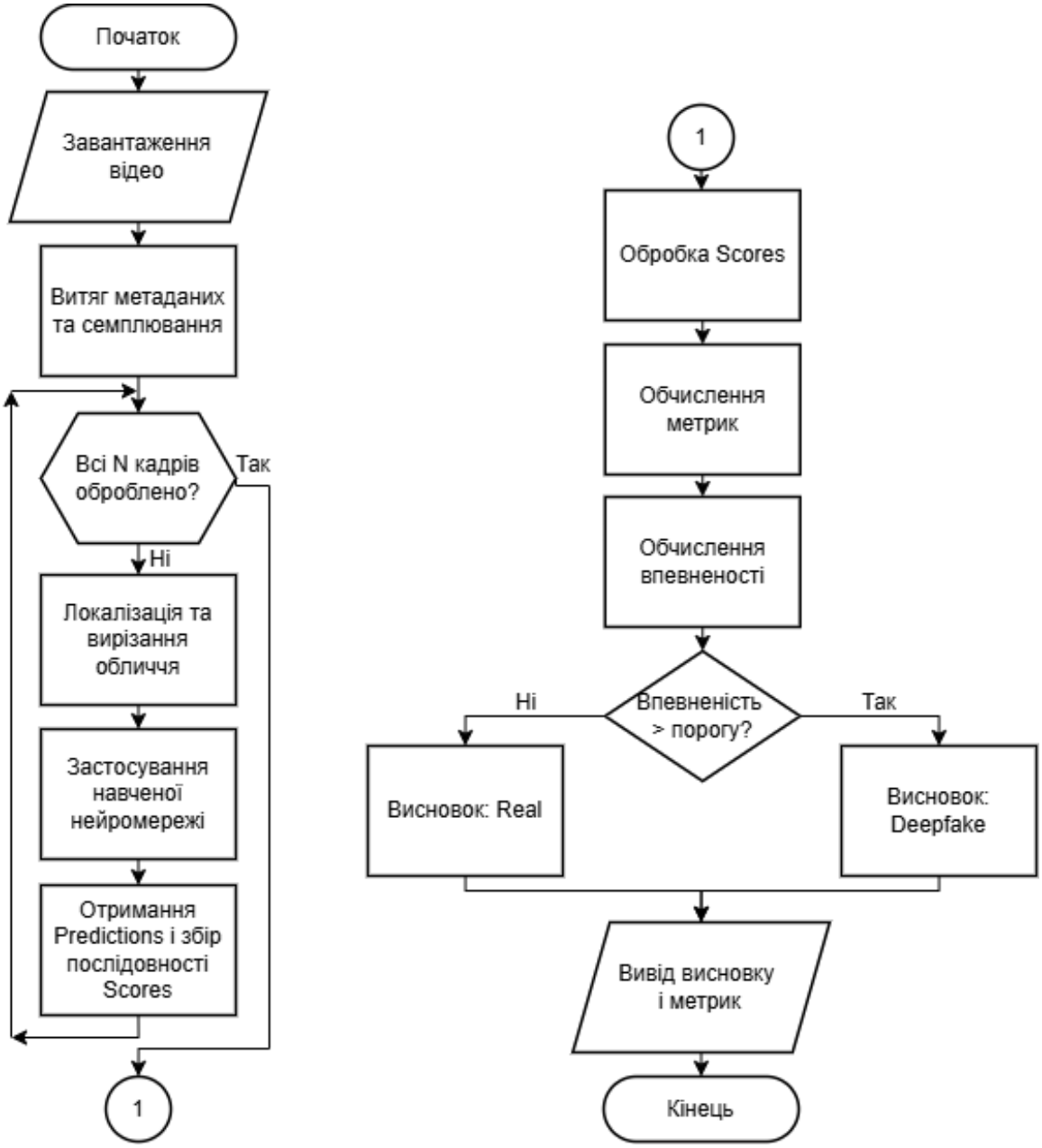


					КАІ 25 13 87 001 ПЛ		
					Діаграма прецедентів системи виявлення deepfake	Літ.	Маса	Масштаб
Змн.	Арк.	№ документа	Підпис	Дата				
Виконала		Старенька К.О.				Арк. 1	Аркушів 1	
Керівник		Супрун О.М.						
Реценз.								
Н. Контр.		Тупота Є.В				М-126-24-1-ІТ		
Зав. каф.		Нечипорук О.П.						

Архітектура інтелектуальної системи

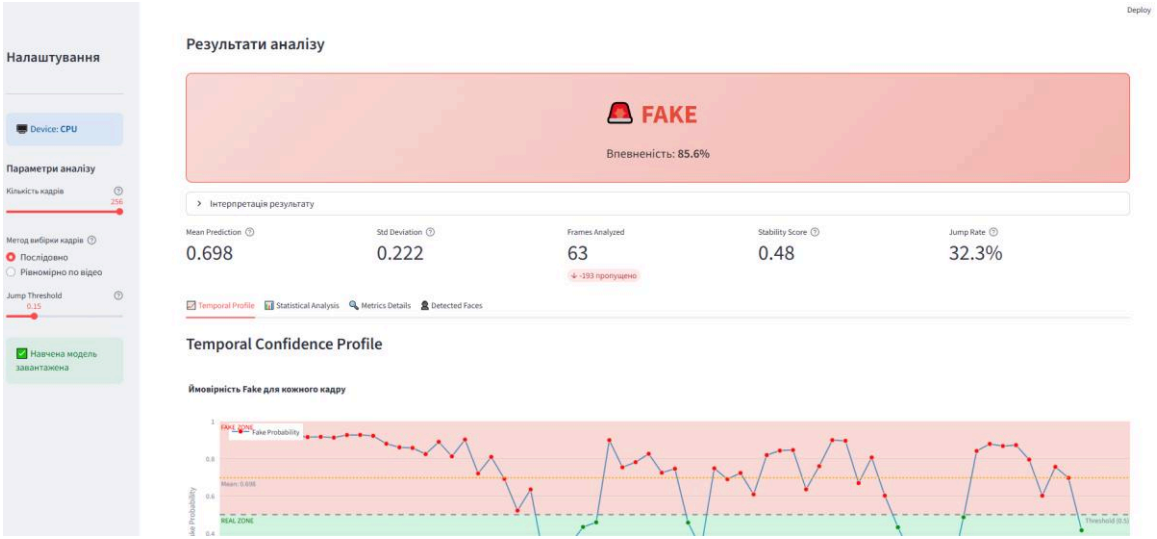


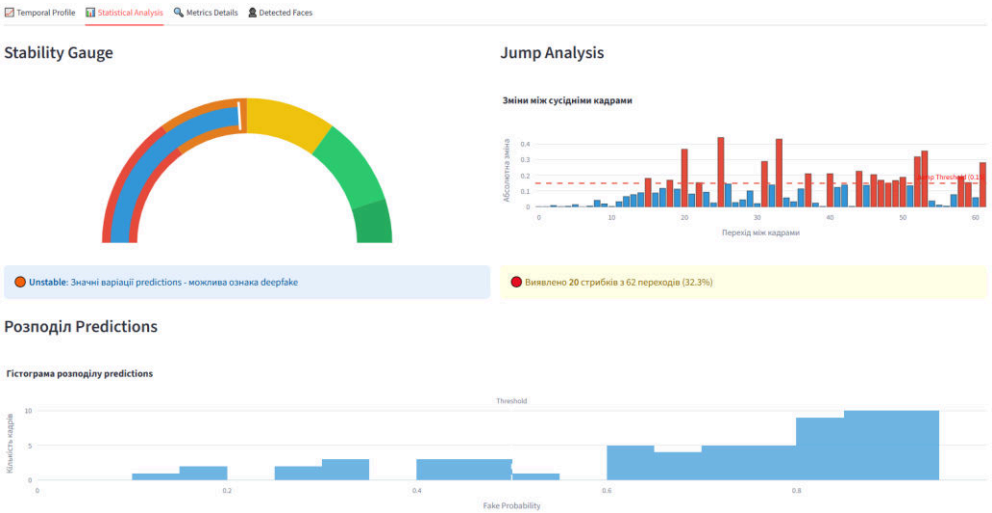
					КАІ 25 13 87 002 ПЛ		
Змн.	Арк.	№ документа	Підпис	Дата	Архітектура інтелектуальної системи	Літ.	Маса	Масштаб
Виконала		Старенька К.О.						
Керівник		Супрун О.М.				Арк. 1	Аркушів 1	
Реценз.								
Н. Контр.		Тупота Є.В				М-126-24-1-ІТ		
Зав. каф.		Нечипорук О.П.						



					КАІ 25 13 87 003 ПМ	
Змн.	Арк.	№ документа	Підпис	Дата	Алгоритм flowchart процесу аналізу відео (схема алгоритму)	Літ. / Маса / Масштаб
Виконала		Старенька К.О.				
Керівник		Супрун О.М.				Арк. 1 / Аркушів 1
Реценз.						
Н. Контр.		Тупота Є.В				М-126-24-1-ІТ
Зав. каф.		Нечипорук О.П.				

Інтерфейс розробленого вебзастосунку з результатами аналізу





Інтерпретація метрик

Метрика	Real (справжнє)	Fake (deepfake)	Поточне значення
Variance	< 0.01	> 0.05	0.0495
Std Deviation	< 0.1	> 0.2	0.2224
Entropy	< 0.5	> 0.7	0.8099
Jump Rate	< 10%	> 25%	32.3%
Smoothness	> 0.9	< 0.7	0.8842
Stability Score	> 0.7	< 0.5	0.4808

					КАІ 25 13 87 004 ПЛ		
					Інтерфейс розробленого вебзастосунку з результатами аналізу	Літ.	Маса	Масштаб
Змн.	Арк.	№ документа	Підпис	Дата				
Виконала		Старенька К.О.				Арк. 1	Аркушів 1	
Керівник		Супрун О.М.						
Реценз.								
Н. Контр.		Тупота Є.В				М-126-24-1-ІТ		
Зав. каф.		Нечипорук О.П.						