МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет «Запорізька політехніка»

Факультет інформаційної безпеки та електронних комунікацій

(повне найменування інституту, назва факультету)

Кафедра інформаційної безпеки та наноелектроніки

(повна назва кафедри)

Пояснювальна записка

до дипломного проекту (роботи)

магістр

(ступінь вищої освіти)

на тему: Методи мультимодального аналізу відеоконтенту для виявлення підроблених відео

Виконав: студент 2 курсу, групи БК-814М

Спеціальності 125 Кібербезпека

(код і назва напряму підготовки, спеціальності)

Освітня програма (спеціалізація)

Безпека інформаційних та комунікаційних систем

САЮШЕВ М.А.

(ПРІЗВИЩЕ та ініціали)

Керівник КОРОТУН А.В.

(ПРІЗВИЩЕ та ініціали)

Рецензент

(ПРІЗВИЩЕ та ініціали)

2025

Форма № 25

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет «Запорізька політехніка»

Факультет інформаційної безпеки та електронних комунікації

Кафедра інформаційної безпеки та наноелектроніки

Ступінь вищої освіти: магістр

Спеціальність 125 Кібербезпека та захист інформації

Освітня програма (спеціалізація) Безпека інформаційних і комунікаційних систем

ЗАТВЕРДЖУЮ

Завідувач кафедри ІБтаН, к.ф-м.н.

_____Андрій КОРОТУН

“____” _____2025 року

З А В Д А Н Н Я
НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТА

САЮШЕВА Максима Андрійовича

(ПРІЗВИЩЕ, ім’я, по батькові)

1. Тема проєкту (роботи): Методи мультимодального аналізу відеоконтенту для виявлення підроблених відео.

Methods of multimodal video content analysis for detecting fake videos.

керівник проєкту (роботи) канд. фіз.-мат. наук, завідувач кафедри КОРОТУН Андрій Віталійович

(науковий ступінь, вчене звання, ПРІЗВИЩЕ, ім’я, по батькові)

затверджені наказом закладу вищої освіти від «26» листопада 2025 року № 530

2. Строк подання студентом проєкту (роботи): 15.12.2025

3. Вихідні дані до проєкту (роботи): розробити прототип мультимодального детектору дипфейків

4. ЗМІСТ розрахунково-пояснювальної записки (перелік питань, що їх потрібно розробити): 1. Поняття підроблених відео та їх класифікація, 2. Методи мультимодального аналізу відеоконтенту для виявлення підроблених відео, 3. Реалізація прототипу мультимодального детектора дипфейків.

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень, кількість слайдів, плакатів): Презентація у програмі Microsoft Power Point ()

6. Консультанти розділів проєкту (роботи)

Розділ	ПРІЗВИЩЕ, ініціали та посада консультанта	Підпис, дата	
		завдання видав	Прийняв виконане завдання
1-3	КОРОТУН А.В., доцент каф. ІБтаН	05.09.2025	12.12.2025
нормоконтроль	КОРОЛЬКОВ Р.Ю., доцент каф. ІБтаН		14.12.2025

7. Дата видачі завдання « 05 » вересня 2025 року.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проєкту (роботи)	Строк виконання етапів проєкту (роботи)	Примітка
1	Ознайомлення з темою та визначення ключових аспектів дипломного проєкту	1 тиждень	Виконано
2	Аналіз теоретичних відомостей щодо підробленого медіаконтенту та його класифікації	1–2 тиждень	Виконано
3	Формулювання мети, завдань, об’єкта і предмета; написання вступу	2 тиждень	Виконано
4	Огляд сучасних методів детекції дипфейків (одномодальні підходи: відео та аудіо)	3–4 тиждень	Виконано
5	Огляд мультимодальних підходів, стратегій інтеграції модальностей та метрик оцінювання	4–5 тиждень	Виконано
6	Проєктування архітектури програмного прототипу мультимодального детектора	5 тиждень	Виконано
7	Реалізація програмного прототипу: інтеграція аудіо та відеомодулів, уніфікація інтерпретації класів	5–6 тиждень	Виконано
8	Контрольне тестування прототипу, збір результатів та їх інтерпретація	6 тиждень	Виконано
9	Формування висновків, визначення обмежень і напрямів масштабування/удосконалення	7 тиждень	Виконано
10	Підготовка пояснювальної записки до подання керівнику (оформлення, перевірка)	7–8 тиждень	Виконано

Студент _____ Максим САЮШЕВ
(підпис) (Ім’я ПРІЗВИЩЕ)

Керівник проєкту (роботи) _____ Андрій КОРОТУН
(підпис) (Ім’я ПРІЗВИЩЕ)

АНОТАЦІЯ

Пояснювальна записка до магістерської роботи: 95 с., 5 табл., 7 рис., 1 дод., 17 джерел.

АУДІОВІЗУАЛЬНИЙ КОНТЕНТ, ГЛИБИННЕ НАВЧАННЯ, ДІПФЕЙК, ІНФЕРЕНС, ІНТЕГРАЦІЯ МОДАЛЬНОСТЕЙ, КІБЕРБЕЗПЕКА, МУЛЬТИМОДАЛЬНИЙ АНАЛІЗ, ТРАНСФОРМЕРИ.

Об’єкт роботи – методи та підходи до детекції підробленого аудіовізуального контенту (deepfake) у контексті кібербезпеки. Предмет роботи - мультимодальні методи аналізу відео й аудіо та алгоритмічні принципи об’єднання результатів одномодальних детекторів у єдине підсумкове рішення.

Мета роботи – ознайомлення з сучасними методами детекції підроблених відео та практична реалізація базових принципів мультимодального підходу у вигляді програмного прототипу, який поєднує виходи аудіо- та відеомодальності й формує узгоджений вердикт.

Наукова новизна полягає в розробленні та апробації інженерної схеми мультимодального конвеєра, що забезпечує уніфікацію виходів різних моделей і підтримує контрольоване зважене об’єднання оцінок для отримання кінцевого рішення.

Практичне значення роботи полягає у можливості використання створеного прототипу як демонстраційної та експериментальної основи для подальшого навчання/донавчання власних моделей під цільовий домен, калібрування правил інтеграції модальностей, а також розгортання процедури масового оцінювання якості на реалістичних наборах даних.

Магістерська робота складається зі вступу, трьох розділів, висновків, переліку джерел посилання та додатку.

ABSTRACT

Explanatory note to the master's thesis: 95 pages, 5 tables, 7 figures, 1 appendix, 17 sources.

AUDIO-VISUAL CONTENT, CYBERSECURITY, DEEP LEARNING, DEEPFAKE, INFERENCE, MODALITY FUSION, MULTIMODAL ANALYSIS, TRANSFORMERS.

The object of the thesis is methods and approaches for detecting manipulated audio-visual content (deepfakes) in the cybersecurity context. The subject of the thesis is multimodal video and audio analysis techniques and the principles of integrating unimodal detector outputs into a single final decision.

The aim of the thesis is to become familiar with modern deepfake detection methods and to provide a practical implementation of a multimodal approach in the form of a software prototype that combines audio and video outputs and produces a consistent final verdict.

Scientific novelty is defined by the design and validation of an engineering multimodal pipeline that unifies heterogeneous model outputs (label/class alignment and probability normalization) and supports a controlled weighted late-fusion strategy for final decision making.

The practical value of the work lies in using the prototype as a demonstrational and experimental basis for future scaling: replacing third-party pretrained models with domain-adapted or self-trained ones, calibrating fusion rules, and organizing large-scale evaluation on realistic datasets.

The thesis consists of an introduction, three chapters, conclusions, references, and an appendix.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

AI - Artificial Intelligence;

AUC - Area Under the Curve;

CLIP - Contrastive Language–Image Pre-training;

CNN - Convolutional Neural Network;

EER - Equal Error Rate;

FAR - False Acceptance Rate;

FPR - False Positive Rate;

FRR - False Rejection Rate;

GAN - Generative Adversarial Network;

GDPR - General Data Protection Regulation;

GMM - Gaussian Mixture Model;

NLP - Natural Language Processing;

RNN - Recurrent Neural Network;

ROC - Receiver Operating Characteristic;

ІІІ - штучний інтелект.

ВСТУП

Стрімкий розвиток генеративних моделей упродовж останнього десятиліття відкрив можливість створення медіаконтенту нового покоління. Відео та аудіосистеми синтезу, які раніше розглядалися як експериментальні або вузькоспеціалізовані рішення, сьогодні переходять до стадії масового застосування у виробництві контенту, розвагах, маркетингу та автоматизації медіапроцесів. Разом з тим виникає принципово нова проблема - забезпечення можливості надійної перевірки автентичності цифрових медіаданих та виявлення підробок, здатних вводити в оману людину й автоматизовані системи.

У класичних підходах до верифікації контенту значна частина завдань вирішується завдяки аналізу характерних технічних ознак та статистичних закономірностей сигналу, однак у випадку дипфейк-продукції пряме перенесення цих принципів є недостатнім через специфіку сучасних генеративних моделей. Реалістичні підробки можуть зменшувати або приховувати типові артефакти, а додатково на результат впливають компресія, повторне кодування, шумові спотворення й варіативність умов запису. Це ускладнює побудову універсальних детекторів, які зберігають стабільність на різних платформах поширення та різних типах контенту.

Сучасні підходи до детекції дипфейків здебільшого базуються на одномодальному аналізі - окремо відеопотоку або окремо аудіосигналу. Проте в реальних сценаріях підробки можуть бути частковими або комбінованими: змінюється лише обличчя при автентичному голосі, синтезується голос при реальному відео, або ж підробляються обидві модальності, але з різною якістю. Це створює потребу у нових архітектурах, здатних виконувати роль мультимодальних детекторів, систем, які забезпечують інтеграцію сигналів з кількох модальностей, узгодження їхніх оцінок і формування підсумкового рішення на основі контрольованого поєднання результатів.

У даній роботі розглядаються принципи побудови таких систем та пропонується підхід до мультимодальної детекції дипфейків, що демонструє можливість поєднання відео та аудіоаналізу в єдиному конвеєрі. Основний акцент зроблено на ознайомленні з методами детекції, їхніми можливостями та обмеженнями, а практична реалізація подана у вигляді прототипу, який інтегрує результати одномодальних моделей і формує інтерпретований вердикт щодо автентичності контенту. Для цього здійснено аналіз базових концепцій підробленого медіаконтенту, підходів до його виявлення, а також принципів мультимодального об’єднання, що дозволяють враховувати різнорідні джерела сигналу.

Запропонований прототипний підхід відображає основні принципи, за якими може бути побудована практично придатна система мультимодальної перевірки медіаданих, і демонструє можливі шляхи подолання типових обмежень одномодальних детекторів через інтеграцію модальностей та уніфікацію інтерпретації результатів. Робота має навчальне та прикладне значення, оскільки формує підґрунтя для подальших досліджень і розробок у галузі мультимодальної детекції deepfake та побудови більш стійких рішень для задач інформаційної і кібербезпеки.

1 ПОНЯТТЯ ПІДРОБЛЕНИХ ВІДЕО ТА ЇХ КЛАСИФІКАЦІЯ

1.1 Поняття підроблених відео та їх класифікація

Сучасні досягнення в галузі штучного інтелекту призвели до появи високореалістичних підроблених відео - відеоматеріалів, зміст яких було навмисно змінено або згенеровано з метою введення в оману глядача. Найбільш відомою технологією є так звані дипфейк-відео (від англ. deep learning - глибинне навчання, та fake - підробка), тобто фальсифікації, створені за допомогою глибоких нейронних мереж. У таких відео обличчя або голос людини замінено на обличчя/голос іншої особи таким чином, що підробку дуже важко відрізнити від справжнього запису. Поняття дипфейк виникло близько 2017 року і стало синонімом AI-генерованих аудіовізуальних фальсифікацій [1]. У ширшому сенсі до підроблених відео можна віднести також «cheapfake» або «shallowfake» - більш прості фальсифікації, створені без складних алгоритмів ШІ, а за допомогою примітивного редагування: уповільнення чи пришвидшення відеоряду, вирізання і перекомпонування фрагментів, дубляжу чи зміни контексту [1]. На відміну від глибинних фейків, cheapfake-фальсифікації не потребують спеціалізованих знань чи значних ресурсів - достатньо доступних відеоредакторів, щоб, приміром, змонтувати хибний сюжет новин або викривити висловлювання публічної особи.

Прикладами cheapfake є, зокрема, прискорення чи уповільнення відео для спотворення поведінки особи, монтаж фрагментів (вирізання або вставлення окремих кадрів, аудіофрагментів) чи неправдивий контекст (коли реальне відео подається з оманливим описом). Відомим прикладом простого фейку є відеозапис, в якому шляхом уповільнення темпу промови створили хибне враження, нібито політик говорить нетверезим голосом - жодних нейромереж для цього не потрібно, достатньо елементарного редагування, однак ефект дезінформації значний.

Дипфейк-видео можна класифікувати і за характером внесених змін. Зараз технології дозволяють: заміщувати обличчя однієї людини на іншу (face swap, або identity exchange), анімувати обличчя по відео або фото під інший голос чи міміку (reenactment, або відтворення виразу обличчя та рухів), точково змінювати риси зовнішності (наприклад, змінити колір чи стиль волосся, вік обличчя - attribute editing), а також повністю генерувати обличчя або фігуру людини, якої не існує (complete face generation). Аналогічно, в аудіосфері розрізняють підробки голосу: це синтез мовлення заданого тексту голосом певної людини, перетворення голосу з тембром, схожим на цільову особу, або імітація живої мови за допомогою аудіофрагментів (технології text-to-speech або voice conversion). Окремо виділяють комбіновані підробки, коли відео підроблено відразу в кількох модальностях - зображення і звук, або відеоряд і пов’язаний текст. Такі мультимодальні фейки найскладніші для виявлення, адже у них узгоджено і фальшиве зображення, і фальшивий аудіоряд, і навіть супровідні текстові описи.

Поняття дипфейк стрімко еволюціонувало завдяки прогресу в галузі штучного інтелекту. Ранні генеративні моделі, такі як варіаційні автокодери (VAE) та генеративні змагальні мережі (GAN) суттєво підвищили якість синтетичних медіа. У подальшому провідну роль почали відігравати дифузійні моделі, які забезпечують вищу візуальну точність, розширене керування процесом генерації та більшу стійкість до виявлення. На відміну від GAN, що базуються на змагальному навчанні, дифузійні моделі поступово уточнюють зображення шляхом редукції шуму, що зменшує ймовірність виникнення типових артефактів цифрової підробки. Така еволюція значно ускладнила розрізнення справжнього контенту від синтетичного, створюючи нові виклики як для людського сприйняття, так і для систем детекції дипфейків.

Особливу загрозу становлять фейкові відео з обличчями, що несуть істотні суспільні та етичні ризики. Хоча методи на кшталт заміни обличчя, редагування атрибутів, імітації міміки або руху губ можуть мати нейтральне або позитивне застосування у сфері розваг і створення віртуальних аватарів,

вони одночасно відкривають можливості для зловмисного використання зокрема для дезінформації, шантажу та фінансового шахрайства. У низці нещодавніх випадків було зафіксовано використання дипфейків у політичних кампаніях та корпоративних шахрайських схемах, де AI-згенеровані обличчя та голоси застосовувались з метою маніпуляції громадською думкою або обману підприємств.

Із розвитком технологій генерації дипфейків маніпуляції більше не обмежуються лише однією модальністю, наприклад, лише відео або лише аудіо. Сучасні методи фальсифікації активно використовують мультимодальні залежності, зокрема синхронізацію штучно згенерованого мовлення з рухами губ, одночасну зміну відеоряду та текстових описів, а також комплексну модифікацію аудіовізуального контенту. Така тенденція до мультимодальних дипфейків істотно підвищує ступінь їх реалістичності, роблячи неефективними традиційні одномодальні методи виявлення.

Для протидії зазначеним загрозам дослідники розробляють дедалі більш складні стратегії детекції. Перші підходи до виявлення дипфейків фокусувалися на пасивних методах, що аналізували артефакти або неузгодженості у контенті. Однак зі зростанням складності мультимодальних фальсифікацій сучасні дослідження застосовують моделі глибокого навчання, зокрема трансформери та моделі злиття візуального й мовного сигналів, з метою підвищення ефективності розпізнавання. Запроваджено низку методів мультимодального злиття ознак, які дозволяють інтегрувати візуальну, аудіальну та текстову інформацію для покращення стійкості моделей до різновидів фальсифікацій. Окрім того, дослідження детекції вийшли за межі виключно пасивного виявлення: активно розвиваються проактивні підходи, які передбачають перешкоджання несанкціонованій генерації дипфейків шляхом додавання адверсарних збурень або вбудованого маркування .

Важливо підкреслити, що не всі випадки зміни медіаконтенту несуть зловмисний характер. Існують легітимні застосування технологій заміни облич чи голосів - наприклад, у кіноіндустрії (для дубляжу або спецефектів),

в розробці віртуальних аватарів, для відновлення старих кінохронік тощо. Однак саме зловмисні дипфейки викликають найбільше занепокоєння. Глибинні фейки відкривають можливості для абсолютно нового рівня обману: вони дозволяють створити переконливе відео з людиною, яка нібито говорить те, чого ніколи не казала, чи робить дії, яких не здійснювала. Внаслідок цього підроблені відео прямо загрожують інформаційній достовірності, приватності та репутації людей. Саме поява дипфейків стала поворотною точкою, після якої експерти заговорили про «епоху постправди» у відео: відтепер бачити - не обов’язково означає вірити.

1.2 Загрози кібербезпеці та інформаційній безпеці, пов’язані з підробленими відео

Поширення підроблених відео контенту породжує серйозні ризики для кібер та інформаційної безпеки. Дипфейк-технології вже застосовуються зловмисниками для реалізації різноманітних атак і шахрайських схем. Однією з головних загроз є масштабна дезінформація: за допомогою фальшивих відеозаяв політиків або суспільних діячів можна поширювати фейкові новини, провокувати суспільний резонанс і впливати на політичні процеси. Наприклад, в Індонезії під час передвиборчої кампанії 2024 року було створено дипфейк-відео з «зверненням» генерала Сухарто на підтримку однієї з партій - цей ролик мав на меті маніпулювати емоціями виборців. Інший резонансний випадок стався у 2022 році, коли в інтернеті з’явилося підроблене відео Президента України, де він начебто закликає армію скласти зброю; ця російська інформаційна атака була швидко викрита, але сама поява такого фейку під час війни продемонструвала новий інструмент інформаційних впливів. Дипфейки, згенеровані для політичних цілей, здатні підривати демократичні процеси, підірвати довіру громадян до офіційних заяв і навіть

спровокувати заворушення. Відомий інцидент стався 2019 року в Габоні, коли після поширення відеозвернення президента, яке здавалося фальшивим, у країні розпочалася спроба військового перевороту, хоча згодом з'ясувалося, що те відео було справжнім, сам факт його підробності сприйняли настільки ймовірним, що це дестабілізувало ситуацію.

Інша загрозлива сфера - шахрайство та соціальна інженерія за допомогою дипфейку. Зловмисники можуть створити відеоконференцію або телефонний дзвінок, у якому генерований голос та обличчя керівника компанії дає співробітнику розпорядження переказати кошти на рахунок аферистів. Вже зафіксовано випадки, коли таким чином були обмануті фінансові менеджери компаній: переконливий аудіо або відеодублікат їхнього керівника наказував здійснити несанкціоновані транзакції. У Гонконзі 2020 року аферисти, використовуючи дипфейк-аватар директора, вкрали у банка понад 35 млн доларів. Такі інциденти показують новий рівень фішингу та бізнес-шахрайства, проти якого традиційні засоби кібербезпеки мало ефективні - адже співробітник бачить і чує начебто реальну людину з впізнаваною зовнішністю та голосом.

Дипфейк становить загрозу і для персональної безпеки та приватності. Одним з перших масових застосувань цієї технології стало створення порнографічних відео за участю відомих осіб без їх згоди. За оцінками, переважна більшість всіх дипфейк-відео в інтернеті - це порнографічні ролики, де обличчя акторів замінено на обличчя знаменитостей або приватних осіб без їх відома. Подібні відео грубо порушують право на приватне життя, завдають шкоди репутації та часто використовуються для шантажу. Жертвою такої підробки може стати будь-яка людина: достатньо викласти свої фото у соцмережі, щоб за наявності навичок і ресурсу зловмисник міг згенерувати фальшиве «доказове» відео компрометуючого характеру. Це створює безпрецедентні виклики для захисту персональних даних і боротьби з сексуальним насильством у цифровій сфері.

Крім прямих атак, сам факт існування технології дипфейк несе довгострокову загрозу суспільній довірі до інформації. Виник феномен, який дослідники називають «дивідендом брехуна»: якщо раніше відеодоказ вважався надійним підтвердженням події, то тепер публічні особи, викриті на небажаних діях, можуть просто заявити, що скандальне відео - це дипфейк, і частина аудиторії схильна в це повірити. Отже, глибинні фейки підривають саму можливість встановлення істини, сіють сумнів у справжності будь-яких цифрових свідчень. Як зазначається в оглядах, дипфейк-відео здатні «поширювати дезінформацію та роз'їдати суспільну довіру», даючи зловмисникам інструмент маскування правди. Це становить загрозу не лише інформаційній безпеці, але й загалом системі правосуддя.

Таким чином, підроблені відео стали новим потужним засобом атак на цілісність інформаційного простору. Вони поєднують елементи кіберзагроз та інформаційно-психологічних операцій. Кібербезпека стикається з необхідністю враховувати принципово новий вектор атак - атак на достовірність контенту. Це вимагає розвитку спеціалізованих систем виявлення підробок, про які йтиметься далі, а також комплексних заходів протидії на рівні законодавства та просвіти користувачів.

1.3 Архітектура систем аналізу відеоконтенту

Для протидії підробкам були розроблені спеціальні системи аналізу відеоконтенту - як з метою автоматичного виявлення фейків, так і для загального розуміння сцени на відео. Архітектура такої системи визначає, як саме відеодані опрацьовуються: які ознаки виділяються і яким чином робиться висновок про автентичність чи зміст відео. Сучасні підходи спираються на досягнення комп'ютерного бачення та глибинного навчання, що продемонстрували високу ефективність в аналізі зображень і відео [2].

Типова система аналізу відео, зокрема для детекції підробок, складається з кількох основних компонентів, По-перше, відбувається попередня обробка даних: відео конвертується у формат, зручний для моделі, виділяються окремі кадри або фрагменти, проводиться масштабування та нормалізація. Часто на цьому етапі здійснюють детекцію облич - визначають положення облич у кожному кадрі та вирізають область обличчя для подальшого аналізу. Наприклад, багато відомих методів спочатку вирізають обличчя з кожного кадру, приводять його до стандартного розміру, а вже потім подають у модель-класифікатор [2].

Наступний компонент - модуль ознак і класифікації. На цьому рівні в хід ідуть алгоритми глибинного навчання, що автоматично вичленовують інформативні патерни з відео. У перших поколіннях рішень широко застосовувалися CNN для аналізу окремих зображень-кадрів. Ці нейроні мережі здатні виявляти дрібномасштабні аномалії на обличчі, які можуть видати підробку: неприродні артефакти на межах накладання облич, некоректну форму очей або зубів, спотворені текстури шкіри [3]. Дослідження показали, що багато ранніх дипфейків містили артефакти, непомітні неозброєним оком, але вловимі нейромережею, - наприклад, ненормальну частоту кліпання очима. Спеціально під такі дефекти були розроблені алгоритми: зокрема, в роботі «Multimodal fake news detection» запропонували модель, що відстежує неприродно рідкісне кліпання як ознаку діпфейка . Інша відома архітектура MesoNet - це компактна згорткова нейронна мережа, оптимізована для розпізнавання підроблених облич, яка досягає високої точності на базових наборах даних за рахунок виявлення мезорівневих ознак обману [4].

Проте обмежуватися аналізом окремих кадрів недостатньо. На практиці відео - це послідовність кадрів у часі, тому узгодженість між кадрами є критичною. Багато підробок грішать дрібними несинхронностями або спотвореннями при переході між кадрами, наприклад, нестале положення тіней, ривки в русі обличчя. Щоб використати цю інформацію, в архітектуру

додають модуль аналізу часу - тобто врахування динаміки. Існують різні підходи: від рекурентних нейронних мереж - RNN, які «пам’ятають» попередній контекст кадрів, до 3D-CNN, що безпосередньо оперують блоком кадрів як об’ємним вхідним даними. Приклад - робота «Deepfakes Detection Techniques Using Deep Learning: A Survey», де застосовано комбінацію CNN для просторового аналізу кадрів та LSTM для врахування часових залежностей: така гібридна архітектура навчена помічати неузгодженості між послідовними кадрами фальшивого відео [5]. Інші дослідники пропонували використовувати оптичний потік або карти руху як окремий потік ознак: порівняння реального руху обличчя з синтезованим може виявити аномалії [6]. Таким чином, в архітектурі нерідко передбачають дві гілки аналізу: одна обробляє просторову інформацію - вигляд обличчя в кожному кадрі, інша - часову динаміку змін. Потім результати обох гілок об’єднуються або навіть взаємодіють на проміжних шарах для ухвалення рішення.

Сучасна тенденція в архітектурі систем контент аналізу - це використання трансформерів. Згорткові мережі добре виявляють локальні ознаки, але можуть пропустити глобальні контексти. Натомість архітектури на основі механізму уваги (transformers) здатні моделювати довгострокові залежності як у просторі, так і в часі. У завданні детекції фейків уже з’явилися рішення з залученням Vision Transformer та їх варіацій, які показують кращу здатність помічати тонкі невідповідності [7]. Зокрема, модель VidTR застосовує багаторівневу self-attention по кадрах, щоб відсіяти фон і сфокусуватися на ключових рисах обличчя, а cross-attention між просторовими і часовими ознаками допомагає виявити навіть приховані артефакти [8]. У результаті, трансформерні гібриди нині демонструють провідні показники на багатьох тестових наборах [9].

Отже, архітектура систем аналізу відеоконтенту для виявлення підробок зазвичай має багаторівневу структуру: від екстракції ознак на рівні кадру до інтеграції інформації між кадрами і фінальної класифікації. Такий підхід дозволяє відловлювати як статичні сліди маніпуляції, так і динамічні.

Більшість сучасних систем діють пасивно, тобто аналізують готовий контент на наявність артефактів. Проте варто згадати й про активні методи, коли в сам контент попередньо вбудовуються захисні мітки або водяні знаки, що полегшують подальшу перевірку [6]. Нині досліджуються гібридні підходи, які поєднують пасивний аналіз з активними заходами захисту.

Важливо підкреслити, що архітектура системи виявлення фейків дедалі частіше проектується з урахуванням мультимодальності - тобто можливості аналізувати не лише відеоряд, а й супутні модальності такі як звук та текст.

1.4 Мультимодальний підхід: модальності та їх взаємодія

Мультимодальний аналіз означає інтегровану обробку різнорідних інформаційних потоків, серед яких ключовими є аудіо, текст, відеоряд, графічні зображення, інфографіка та супровідні метадані. Мультимодальний підхід у аналізі відеоконтенту означає, що для розв’язання задачі виявлення підробленого відео одночасно використовуються декілька різнорідних джерел даних - так званих модальностей. У випадку відео природно виділити принаймні дві модальності: візуальну та аудіальну. Додатково можна враховувати й третю модальність - текстову. Наприклад, це можуть бути субтитри чи розпізнаний текст мовлення, а також заголовки або описи відео в соцмережах. Мультимодальний аналіз передбачає спільну обробку і зіставлення інформації з усіх цих модальностей. Такий підхід виправданий, оскільки сучасні підробки нерідко охоплюють одразу кілька аспектів контенту. Зокрема, дипфейк-відео майже завжди є аудіовізуальним фейком - правдоподібна фальсифікація обличчя супроводжується не менш правдоподібною підробкою голосу, синхронізованою з рухом губ. Якщо аналізувати лише зображення без звуку або лише звук без відеоряду, можна пропустити важливі ознаки неузгодженості, які видають фейк.

Ключова ідея мультимодального підходу - взаємодія модальностей. Справжній автентичний контент має повну узгодженість між візуальним рядом, звуком і супутнім текстом. У фальсифікаціях часто виникають помилки - невідповідності між модальностями, які покликаний виявити алгоритм. Приміром, у низькоякісних дипфейках можна помітити розсинхронізацію: рух губ актора не ідеально співпадає з вимовленими словами, міміка виглядає трохи штучною. Комп'ютерні моделі здатні вловити такі невідповідності значно краще, ніж людське око. Як зазначають дослідники, в аудіовізуальних дипфейках часто трапляються «тонкі неузгодженості між аудіо і відео» - наприклад, форма рота на відео може не точно відповідати вимовленому звуку - невідповідність між фонемою та віземою. У реальному відео артикуляція і звук синхронні, тоді як у згенерованому можливі зміщення через окреме моделювання зображення і голосу. Отже, аналіз синхронності руху губ і мовлення - один із базових аспектів мультимодального підходу. Якщо голос озвучує слова, а обличчя рухається інакше тоді є підстава підозрювати підробку.

Крім синхронізації, мультимодальні методи відстежують і більш високорівневі взаємозв'язки. Наприклад, відповідність змісту аудіо візуальному контексту: у фейковому новинному відео текст за кадром може коментувати подію, яка на відеоряді насправді не відбувається. Відомий випадок: у соцмережах поширювалося відео з дівчиною, що сміливо сперечається з озброєним солдатом, видаючи його за сюжет з війни в Україні - насправді це було архівне відео палестинської активістки з ізраїльським солдатом, тобто справжнє відео в фейковому контексті. Така невідповідність між текстовою модальністю і візуальною модальністю теж є об'єктом автоматичного аналізу. Сучасні мультимодальні моделі фактично здатні виконувати крос-модальну перевірку консистентності: співставляти, чи узгоджується те, що показано, з тим, що сказано або написано. Цей напрям особливо актуальний для виявлення фейкових новин у форматі відео, де

оманливий може бути не сам відеоряд як такий, а підпис чи озвучка, що надають неправдивий контекст.

Для реалізації мультимодального підходу застосовуються алгоритми fusion, про які згадувалося вище. Простою схемою є об'єднання ознак з різних модальностей і подавання їх на вхід моделі-класифікатора. Проте у практиці глибокого навчання доведено ефективність і складніших механізмів. Зокрема, трансформерні архітектури можуть бути спеціально налаштовані на кілька модальностей: існують, наприклад, моделі типу Vision + Language Transformer, які одночасно приймають на вхід картинку та підпис до неї і генерують спільне представлення. Для задач детекції фейків вже пропонуються архітектури, де візуальні та аудіо ознаки зливаються через шар перехресної уваги - модель навчена співвідносити патерни зображення зі звуковими патернами, що дозволяє виявити невідповідності на більш абстрактному рівні. Інший підхід - це залучення MLLM. Новітні чатботи на кшталт GPT-4 Vision можуть сприймати картинку або відеофрагмент і аналізувати його, спираючись на своє «розуміння» реального світу. В експериментах показано, що такі моделі в певних випадках здатні розпізнати штучно згенероване зображення, керуючись семантичними ознаками наприклад, помітивши нелогічні деталі. Хоча MLLM наразі не замінюють собою детектори фейків, їх інтеграція - перспективна опція для побудови асистуючих систем: скажімо, мультимодальний ШІ може пояснити, чому він вважає відео підробленим, вказавши на конкретні невідповідності між модальностями, - такий підхід підвищує пояснюваність і довіру до автоматичної детекції.

Підсумовуючи, мультимодальний підхід є необхідною відповіддю на появу складних фейків. Аналіз лише однієї модальності дає обмежену картину і може бути обманутий добре підробленим контентом. Натомість комплексний аналіз кількох модальностей одночасно значно підсилює надійність: фальсифікаторам треба однаково добре зімітувати і зображення, і звук, і контекст, аби не залишити слідів. Як показує практика досліджень, поєднання модальностей дозволяє виявити фейки, що були б непомітні при роздільному

розгляді каналів інформації. Мультимодальні системи виявлення зараз є передовою лінією оборони проти дипфейків, і в наступному розділі ми розглянемо, на яких даних вони навчаються та тестуються.

1.5 Джерела даних і відкриті набори для аналізу підроблених відео

Розвиток методів детекції фейкових відео тісно пов’язаний із наявністю якісних датасетів - великих колекцій відеозаписів із мітками «справжнє/підроблене», на яких можна навчати й перевіряти алгоритми. За останні кілька років дослідницькою спільнотою зібрано та оприлюднено низку відкритих наборів даних, що стали стандартом у цій галузі. Першим помітним кроком став датасет FaceForensics++ (FF++), представлений у 2019 році як базовий бенчмарк для задач розпізнавання підробок обличчя. Набір FaceForensics++ містив понад 1000 відео із чотирма типами маніпуляцій обличчя: FaceSwap, Face2Face, DeepFake та NeuralTextures. Кожен з цих методів генерував підробні відео на основі реальних - заміна облич або його анімація різними алгоритмами, тож FF++ надав дослідникам різноманітний матеріал для навчання моделей на кількох видах фальсифікацій.

Паралельно компанія Google у співпраці з Jigsaw випустила власний набір DeepFake Detection (DFD), який включав сотні підроблених відео, згенерованих з участю акторів-добровольців. Цей набір було передано для використання у глобальному конкурсі Facebook Deepfake Detection Challenge (DFDC), що відбувся у 2019–2020 роках. Набір DFDC став одним з найбільших на той час: він містив понад 128 тисяч відео, з яких близько 83% були штучно створеними дипфейками різних видів. Для генерації такого обсягу контенту залучалися різні алгоритми та актори; датасет DFDC суттєво підняв планку складності та різноманітності фейків, стимулювавши розвиток стійкіших методів детекції. У тому ж 2020 році з’явився ще один важливий набір - Celeb-DF, що складався з відео за участю знаменитостей. Celeb-DF вирізнявся тим,

що фальшивки в ньому були дуже високої якості і близькі до реальних умов; завдяки цьому він використовується як тестовий набір для перевірки моделей на здатність узагальнювати знання поза навчальними даними. Моделі, що показують майже 100% точності на FaceForensics++, на Celeb-DF часто мали значно гірші результати, отже цей датасет підсвітлив проблему перенавчання на вузьких ознаках і спонукав шукати більш генералізовані підходи.

У 2021–2023 роках з’явилися нові набори даних, що відображають еволюцію дипфейків та розширюють межі задачі. Зокрема, акцент змістився на мультимодальні та мульти-мовні фейки. Так, у 2024 році презентовано датасет Polyglot-Fake, який містить підроблені відео з озвучуванням різними мовами, зокрема тональними, аби дослідити вплив мовних особливостей на якість фальсифікації. Інший приклад - набір FakeAVCeleb, що поєднав фейкові відео і аудіо знаменитостей, надаючи матеріал для дослідження аудіовізуальних атак. У 2024 році з’явився бенчмарк Deepfake-Eval з колекцією фейків, виявлених у відкритому доступі протягом року. Він мультимодальний і включає приклади, де атаковано одразу кілька аспектів: обличчя, голос, текст. Також слід відзначити датасет WildDeepfake, де зібрано відео з мережі, вже помічені як фейкові, - моделі, натреновані на відносно «чистих» синтетичних даних, на таких «диких» зразках часто дають збій, тому це корисний стрес-тест.

Окремий клас даних - це спеціалізовані набори під конкретні сценарії. Наприклад, для перевірки безпеки відеоконференцій під час пандемії було створено датасет Zoom Deepfake (Zoom-DF), який моделює підробки в умовах відеодзвінків. Для задач цифрової криміналістики з акцентом на пояснюваність зроблено набір ExDF (Explainable DeepFake) з відповідною розміткою артефактів. Постійно оновлюються і згенеровані обличчя: наприклад, ThisPersonDoesNotExist та інші колекції тисяч облич людей, яких не існує, слугують допоміжними даними для відточування алгоритмів.

Важливо, що відкриті датасети супроводжуються чіткими протоколами оцінювання. На них сформовано спільні метрики - точність класифікації,

показник AUC, показник помилкових спрацьовувань тощо - які дозволяють об’єктивно порівнювати різні підходи. Наприклад, на FaceForensics++ більшість сучасних моделей досягають точності понад 90–95%, тоді як на більш складному Celeb-DF цей показник може падати до 65–75%, що свідчить про наявність простору для вдосконалення. Таким чином, прогрес виявлення підроблених відео багато в чому зумовлений доступністю різнопланових даних для навчання. Кожен новий датасет - це виклик для алгоритмів і одночасно крок до їх покращення, адже дозволяє навчити моделі помічати раніше невідомі різновиди підробок.

Таблиця 1.1 – Порівняння таблиця дата-сетів з діпфейк та реальною вибіркою даних.

Назва датасету	**Обсяг і вміст**	**Характеристики**
FaceForensics++	~1000 реальних, ~4000 підробок.	Тільки обличчя, без звуку; контрольовані маніпуляції (DeepFake, FaceSwap, Face2Face, NeuralTextures)
DFDC	~23k реальних, ~100k підробок.	Різноманітні обличчя, сцени; є звук (не фальшивий); великий розмір, різні якісті.
Celeb-DF	590 реальних, 5639 підробок.	Якісні face-swap фейки знаменитостей, складні для детекції.

Продовження таблиці 1.1

Назва датасету	Обсяг і вміст	Характеристики
WildDeepfake	410 реальних, 297 фейків.	«Дикі» відео з інтернету, різні формати, невідомі алгоритми.
DeeperForensics-1.0	5000 реальних, 10000 підробок.	Власний метод фейку, різні перекручення (шуми, обрізка) для робастності.
ASVspoof 2019/21	~20k реальних, ~20k фейкових аудіофайлів.	Різні методи синтезу і підміни голосу; для оцінки аудіодетекторів.
FakeAVCeleb	500 реальних відео, ~500 з підробленим голосом, ~500 з обличчям.	Аудіо-visual фейки: face-swap + voice clone від знаменитостей; перевірка аудіо-візуальних моделей.
LAV-DF	1,000+ відео з частковими маніпуляціями.	Фейки, де змінено локально деякі фрагменти (напр., одна фраза або емоція обличчя); для тестування локалізації .
FakeSV	36,000 коротких відеоновин.	Фейкові новини: мультимодальні (відео+текст+соцмережі); найбільший у своїй категорії

1.6 Нормативно-правові та етичні аспекти використання та виявлення підроблених відео

Стрімкий розвиток технологій створення підроблених відео викликав хвилю дискусій щодо правових і етичних наслідків їх використання. З одного боку, виникла потреба у правовому регулюванні виготовлення та розповсюдження дипфейки. З іншого - постали питання етики та законності застосування засобів для їх виявлення, адже такі засоби самі по собі можуть втручатися у приватність користувачів.

Більшість країн поки не мають спеціального законодавства, спрямованого саме на глибинні фейки, але поступово з’являються перші ініціативи. Основний акцент законодавців - боротьба з очевидно шкідливими застосуваннями: порнографічні дипфейки без згоди особи, відеопідробки, що втручаються в виборчий процес, шахрайство тощо. У низці штатів США вже ухвалено закони, які криміналізують поширення підробленої порнографії та забороняють публікацію фейкових відео кандидатів напередодні виборів. Наприклад, в Каліфорнії діє заборона на неоплачувані дипфейк-зображення кандидатів за 60 днів до виборів, а Вірджинія ввела відповідальність за створення та розповсюдження нереальної порнографії за участю реальних осіб. У 2022 році в Китаї було прийнято перші у світі правила щодо «deep synthesis» технологій, які зобов’язують провайдерів маркувати синтетичний медіаконтент спеціальними водяними знаками або попередженнями для користувачів. Також заборонено використовувати ШІ для виготовлення фейків з метою завдати шкоди національній безпеці або репутації людей - за це передбачено кримінальну відповідальність.

На рівні міжнародного права питання дипфейків тільки почало обговорюватись. Європейський Союз включив поняття синтетичного контенту до проєкту Акту про штучний інтелект – «AI Act». У фінальній редакції 2024 року там передбачено, що будь-який AI-генерований

медіаконтент, який може ввести в оману людину, повинен супроводжуватися чітким маркуванням про його штучне походження (виняток зроблено для пародій і контенту зі згодою акторів). Це фактично зобов'яже великі платформи впровадити детектори дипфейків і автоматичні мітки «це згенеровано ШІ». Також ЄС планує посилити відповідальність за шкідливе використання таких технологій у сфері дезінформації. На глобальному рівні ООН та інші міжнародні органи звертають увагу на загрозу «діджитал-обману» і обговорюють можливість координації зусиль у протидії. Проте загалом законодавство поки що не встигає за технологіями: існують значні прогалини у праві, адже традиційні норми щодо обману, шахрайства чи порушення авторських прав не повністю охоплюють специфіку дипфейків.

Етичні аспекти. Використання підроблених відео порушує базові етичні принципи достовірності інформації і поваги до особистості. Беззаперечно неетичним є створення дипфейків з метою обману, маніпуляції чи образи іншої особи. Фальшиве відео може серйозно зашкодити репутації людини, завдати психологічної травми. З погляду прав людини, несанкціоноване використання чийогось образу та голосу порушує право на приватність і самовизначення - адже в жертви забирають контроль над власним зображенням. Також постають питання згоди та авторства: чи може взагалі бути етичним створення копії людини без її відома, навіть якщо це «лише» цифровий двійник. Більшість експертів схиляються до думки, що потрібне право індивіда контролювати свої цифрові образи, включно з можливістю заборонити тренування ШІ на своїх фото або відео. На практиці реалізація цього права складна, але певні кроки вже робляться.

Етичні дилеми є і стосовно технологій виявлення дипфейків. З одного боку, їх розробка диктується благими намірами - захистити суспільство від дезінформації та людей від дискредитації. З іншого боку, масове впровадження систем моніторингу відеоконтенту породжує ризики надмірного нагляду за користувачами. Такі системи потенційно аналізують величезні обсяги особистого контенту (фото, відео), що піднімає питання

приватності: чи не стане боротьба з дипфейками приводом для масового стеження? Як зазначено в одному дослідженні, алгоритми детекції часто потребують аналізу метаданих, поведінкових патернів користувача, що викликає «занепокоєння щодо масового спостереження та можливого зловживання ШІ в онлайн-моніторингу». Також існує ризик цензури: автоматизовані фільтри можуть помилково позначати легітимний контент як фейк і видаляти його. Особливо це небезпечно в авторитарних суспільствах, де під приводом боротьби з дезінформацією влада може глушити опозиційні висловлювання або сатиру. Тому етичним імперативом є забезпечення прозорості та справедливості в роботі детекторів. Алгоритми повинні бути по можливості пояснюваними, а також неупередженими. Вимога дотримання цих принципів вже закріплюється і нормативно: наприклад, європейські регуляції (GDPR, майбутній «AI Act» містять норми про недопущення алгоритмічної дискримінації та обов'язковий людський нагляд за критичними рішеннями ШІ.

Отже, у сфері дипфейків необхідний баланс між свободою творчості і самовираження та запобіганням зловживанням. Нормативно-правові заходи мають на меті криміналізувати найбільш шкідливі форми використання підроблених відео і зобов'язати платформи маркувати синтетику, не порушуючи при цьому свободи слова. Етично важливо виробити стандарт відповідального використання: так, існує думка, що відповідальні виробники AI-контенту повинні добровільно вставляти водяні знаки в згенеровані відео, аби полегшити їх виявлення і не вводити публіку в оману. Водночас, розробники детекторів мають дотримуватися принципів пропорційності: аналізувати контент в рамках необхідного і не накопичувати зайвих даних про користувачів. Лише поєднання правових норм і етичних стандартів дозволить мінімізувати ризики від нових технологій. Наразі ця сфера активно формується - і технічна спільнота, і законодавці шукають підходи, які захистять суспільство від небезпечних фейків, не породжуючи при цьому нових загроз для прав і свобод.

2 МЕТОДИ МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ ВІДЕОКОНТЕНТУ ДЛЯ ВИЯВЛЕННЯ ПІДРОБЛЕНИХ ВІДЕО

2.1 Одномодальні підходи до виявлення підроблених відео

Перші підходи до детекції фальсифікованого відео зосереджувалися на аналізі лише зображенні. Такі одномодальні візуальні методи застосовують алгоритми комп'ютерного бачення для виявлення артефактів на обличчі та інших ознак підробки на рівні зображення [10]. Зокрема, нейронні мережі (CNN, автоенкодери тощо) навчаються розпізнавати невідповідності візуальних характеристик, що можуть виникати при накладанні облич: це морфологічні аномалії (некоректна геометрія або пропорції обличчя), артефакти згладжування на межі вставленого обличчя, неконсистентність освітлення (нереалістичні тіні, відблиски) або неприродна міміка [10]. Наприклад, глибокі фейки іноді видають себе через неприродний вираз очей чи рота, «застигле» обличчя при емоціях або невідповідність рухів губ звуку. Багато методів аналізують обличчя покадрово, намагаючись виявити сліди згладжування або шуму, внесені генеративною моделлю. Інші підходи звертаються до частотних ознак зображення - шукають аномальні періодичні структури або спектральні відбитки, що залишаються після генерації зображень нейромережею. У результаті, сучасні CNN-моделі можуть ідентифікувати дрібні візуальні невідповідності між справжнім і штучно згенерованим обличчям, які непомітні неозброєним оком.

Аналіз мікрорухів та фізіологічних сигналів. Окремий клас одномодальних візуальних методів базується на аналізі неявних рухових характеристик обличчя - так званих мікрорухів, які важко або неможливо синтезувати правдоподібно. Прикладом є частота кліпання очей: людське око кліпає із певною середньою частотою та варіаціями, які важко відтворити генератору фейків. Дослідники виявили, що в багатьох штучних відео алгоритми заміни облич не приділяли уваги правильному відтворенню

кліпання - підроблене обличчя могло майже не кліпати або робити це ненатурально рідко. Як приклад можна навести метод, що відстежує моргання: якщо частота кліпань очей істотно відхиляється від норми (наприклад, ненормально низька кількість кліпань за хвилину), це сигналізує про можливу підробку [11]. Цей підхід показав, що аномально рідке або нерегулярне кліпання є характерною ознакою багатьох ранніх дипфейк-відео, недосконало змодельованих генеративними мережами. Окрім кліпання, до мікрорухів відносять також дрібні мімічні зміни (наприклад, тремор губ, ледь помітні асиметрії рухів м’язів обличчя), а також фізіологічні сигнали - скажімо, зміна кольору шкіри від пульсації крові. Існують роботи, що намагаються зчитувати пульс з обличчя на відео (remote PPG) та перевіряти його на узгодженість: справжні відео відтворюють ледь помітний ритм зміни відтінку шкіри, тоді як у фейках такий сигнал може бути відсутній або скопійований некоректно. Щоправда, зі зростанням якості дипфейку і ці ознаки поступово враховуються зловмисниками - з’являються підробки, де і частота кліпання, і інші мікросигнали імітуються правдоподібніше [11]. Втім, аналіз мікрорухів залишається дієвим інструментом: поєднання контролю кліпання, рухів рота та інших дрібних динамічних проявів дозволяє виявляти підробки, не залежачи від суто піксельних артефактів.

Одномодальні методи виявлення фальсифікації звуку зосереджені на аналізі звукової доріжки відео. Мета цього методу - виявити спотворення голосу чи неприродні характеристики мовлення, які виникають при синтезі або підміні голосу. Сучасні аудіо фейки, такі як синтез мовлення певного диктора або конверсія голосу, досягають високої якості, але все ще можуть залишати спектральні артефакти - шуми, дискретність частотного спектру, неприродну гладкість або, навпаки, різкість звуку, відсутність характерної модуляції. Для їх виявлення застосовують аналіз спектру сигналу та ряд ознак, запозичених з класичного розпізнавання мовлення. Зокрема, Mel-frequency cepstral coefficients (MFCC) - мел-кепстральні коефіцієнти стали стандартною ознакою для представлення тембрових характеристик голосу. MFCC

компактно описують енергетичний спектр мовлення і тим самим фіксують його ключові властивості. Виявилося, що ці ознаки дуже чутливі до штучності: на їх основі алгоритми можуть надійно відрізнити справжню людську мову від згенерованої нейромережею. У практиці детекції аудіофейків аудіосигнал перетворюють на набір спектральних ознак і подають в класифікатор. Спочатку застосовувалися традиційні моделі на зразок GMM або SVM, проте швидко їх витіснили глибинні нейронні мережі. Наприклад, згорткові нейромережі можуть приймати на вхід спектрограму звуку і виявляти незвичні для людського голосу шаблони. Інший підхід - мережі, що працюють напряму з сирим аудіосигналом у часовій області (наприклад, модель RawNet), навчаючись витягати потрібні ознаки самостійно. Додатково використовуються ознаки вокальної подібності та стилю мовлення: синтезовані голоси можуть мати пласку інтонаційну криву, ненатуральні паузи чи нетипову для спікера манеру мовлення. Комплексний аналіз спектральних і просодичних характеристик дозволяє досягти високої точності: так, сучасні методи аудіо-детекції демонструють точність понад 95% на відомих наборах даних, хоча все ще слабко узагальнюються на абсолютно нові голосові моделі [12].

Відеоконтент може містити і третій модальний компонент - текст, явний чи прихований. По-перше, це текстова розшифровка мовлення (субтитри або автоматично отриманий транскрипт зі звуку). По-друге, до текстових сигналів можна віднести метадані відео (назви, описи) або навіть текстові надписи, що з’являються у кадрі. Хоча в традиційному завданні дипфейк-детекції текстові дані відіграють допоміжну роль, їх аналіз теж може підвищити виявлення підробок. Зокрема, семантичний аналіз контенту мовлення здатен виявити невідповідність між тим, що говорить персонаж на відео, і тим, що він міг би говорити насправді. Приміром, якщо в підробленому відео використано синтезований голос відомої особи, зміст мовлення (текст) може не відповідати стилю цієї особи або містити нетипові фрази. Аналіз лінгвістичних особливостей - вибору слів, швидкості мовлення, акценту може слугувати

додатковим фактором автентичності. Ще один напрям - перевірка фактичної достовірності висловлювань: у фейкових відео з дезінформацією текст промови може містити завідомо хибні або абсурдні твердження, які справжня людина навряд чи б висловила. За допомогою технологій аналізу тексту (NLP) такі невідповідності теж можуть бути виявлені. Таким чином, хоча основний фокус детекції підробленого відео зосереджений на зображенні та звуці, текстові сигнали - транскрипти і супровідний текст також можуть бути використані для підвищення впевненості у висновку про автентичність контенту.

2.2 Основні архітектури мультимодального аналізу

Перехід від одномодальних методів до мультимодального аналізу відео продиктований необхідністю подолати обмеження окремих сигналів. Як зазначалося, сучасні фейки можуть одночасно підміняти візуальний ряд і аудіодоріжку, узгоджуючи їх між собою, що різко ускладнює задачу виявлення. Якщо система перевіряє лише картинку або лише звук, такі комбіновані підробки можуть її обійти. Тому актуальні підходи використовують спільну модель кількох модальностей: аналізують відео й аудіо в єдиному архітектурному рішенні, шукаючи міжмодальні невідповідності. Об’єднання модальностей дозволяє уловити тонкі взаємозв’язки між ними - наприклад, невірну синхронізацію руху губ із мовленням, відсутність шумів довколишнього середовища, характерних для сцени, в накладеному аудіо. Завдяки цьому мульти-модальні методи можуть знаходити ознаки маніпуляції, непомітні для одномодальних детекторів [13]. Практично реалізація такого підходу потребує спеціальних архітектур глибокого навчання, здатних працювати одразу з двома (або більше) потоками даних.

Спільне представлення модальностей. Базовим завданням мультимодальної архітектури є навчитися зіставляти та об’єднувати інформацію з різних джерел - відеоряду та звуку, а інколи і тексту. Ранні нейроні мережі вирішували це через побудову окремих блоків нейронної мережі для кожної модальності, після чого ці ознаки згортались у спільний вектор і подавались на класифікатор. Таким чином досягається так зване feature fusion - злиття ознак, і мережа навчається ухвалювати рішення на основі комбінації прихованих представлень обох модальностей. Подібний архітектурний шаблон застосовано, зокрема, у роботі «A Multimodal Model for Audio-Visual Deepfake Detection», де використовувалася двоканальна конвольна мережа: один потік обробляв відеокадри, другий - аудіопослідовність, а на виході їхні ознаки об’єднувалися для спільної класифікації справжнього або підробки. Перевага підходу - можливість використати існуючі напрацювання з аналізу зображень і звуку та комбінувати їх результати. Втім, ранні реалізації мульти-моделей часто зіштовхувалися з проблемою: просте злиття ознак не враховувало складних залежностей між модальностями у часі і просторі. Наприклад, якщо ознаки об’єднуються лише на рівні остаточного вектору, модель може пропустити момент, коли звук і зображення розходяться в середині ролика. Це стимулювало появу більш продвинутих архітектур із глибшим спільним представленням, де взаємодія між модальностями відбувається на кількох рівнях моделі.

Важливим кроком у розвитку мультимодальних архітектур стало впровадження механізмів уваги (attention) та самоуваги (self-attention), запозичених з трансформерів. Механізм attention дає моделі змогу динамічно зважувати вклад кожної модальності та окремих її частин при ухваленні рішення. Інтуїтивно це схоже на те, як людина при перевірці відео може сконцентруватися на русі губ, якщо хоче звірити його із звуком, або на фоні кадру, щоб помітити невідповідний шум. Архітектури нейромереж із attention-механізмом дозволяють здійснювати такі перехресні зіставлення автоматично. Зокрема, сучасні дослідники інтегрують в моделі спеціальні блоки крос-

модальної уваги: вони отримують на вході послідовності ознак відео та аудіо і навчаються «вирішувати», які фрагменти аудіо відповідають яким кадрам, а які ні. Це дає змогу виявити, наприклад, на певній секунді у відео рот людини замкнений, тоді як аудіо містить мову - явна невідповідність. З технічної точки зору, attention-механізми часто реалізуються на базі трансформерних шарів. Трансформери зарекомендували себе як потужний інструмент для роботи з послідовностями. У завданні аналізу відео це дозволило ефективно моделювати довготривалі залежності і візуального ряду, і аудіосигналу. Низка новітніх підходів до мульти-модального дипфейк-детектування будується саме на трансформерах: модель представляється як набір токенів, що описують кадри відео і фрагменти аудіо, і за допомогою багаторівневого механізму уваги досліджуються всі можливі взаємозв’язки між цими токенами. Такі моделі навчаються спільному статистичному представленню модальностей, в якому інформація про одну модальність може впливати на іншу під час проходження по мережі. У результаті, увага дозволяє виявляти більш приховані аномалії: наприклад, якщо фальсифікація містить невірну емоцію голосу, що не відповідає виразу обличчя, крос-модальний attention-шар може зіставити ознаки емоційності аудіо і відео та сигналізувати про розбіжність.

Однією з домінуючих сучасних архітектур є трансформерні моделі, спеціально адаптовані для одночасної роботи з різнорідними даними - зображеннями, звуком, текстом. Кожну модальність зазвичай кодують у вигляді послідовності векторів, далі ці послідовності або об’єднуються в єдину або обробляються окремими трансформерними енкодерами з подальшою взаємодією. Ключова перевага трансформерів - самоувага, яка дозволяє гнучко моделювати як внутрішньомодальні залежності, так і міжмодальні. У трансформерних архітектурах для мультимодальних даних часто застосовуються спеціалізовані шари, що реалізують перехресну увагу cross-attention: вони приймають приховані представлення однієї модальності як ключі і значення, а запити формуються з іншої модальності. Таким чином,

модель може, як приклад, спроєктувати ознаки аудіо на часову лінію відео і навпаки - фактично виконати прогноз одного сигналу з іншого. Цікаву реалізацію цього підходу продемонстрували розробники модуля CRATrans (Cross-Reconstruction Attention Transformer): у навчанні цей модуль на основі ознак аудіо намагається реконструювати ознаки відеоряду і навпаки, тим самим виявляючи і підкреслюючи ті особливості, які не можуть бути відновлені через відсутність справжньої кореляції між сигналами. Експериментально продемонстровано, що такий перехресний трансформер здатен краще помічати аномалії синхронізації і інші неузгодженості, ніж традиційне об’єднання ознак без уваги [10].

Слід зазначити, що мультимодальні трансформерні моделі часто вимагають значних обсягів даних для навчання. Тому в цій галузі з’являються й підходи перенесення навчення з суміжних задач. Наприклад, моделі на кшталт CLIP, навчені на відповідність підписів до зображень, можуть бути використані для ініціалізації частини мультимодальної мережі (візуального або текстового енкодера). Далі мережу донавчають на спеціалізованих наборах з реальними та підробленими відео. Загалом, архітектури мультимодального аналізу продовжують активно розвиватися - поєднуючи окремі канали обробки для кожної модальності зі складними механізмами їх взаємодії (увагою, пам’яттю, спільними шарами), вони прагнуть максимально повно використати всю доступну інформацію з відео для виявлення навіть добре замаскованих фейків.

2.3 Методи глибинного навчання для аналізу відеоконтенту

Застосування глибинного навчання кардинально підвищило ефективність аналізу відео на наявність підробок. Сучасні моделі здатні самостійно навчатися ознак, оптимальних для детекції, з мінімальною участю

експерта. Далі розглянуто основні архітектурні рішення глибинних нейромереж, які використовуються для аналізу відеоконтенту та виявлення маніпуляцій.

3D-CNN для відео або тривимірні згорткові нейромережі (3D Convolutional Neural Networks) - це розширення звичайних 2D-CNN на вимір часу. В них згорткові фільтри охоплюють не лише площину кадру, але й кілька сусідніх кадрів, тобто мають розмірність ширина, висота, час. Таким чином, 3D-CNN одночасно екстрагують просторові та часові особливості з відеофрагментів. Для задачі дипфейк-детекції 3D-CNN корисні тим, що можуть вловлювати короткочасні артефакти між кадрами - раптові ривки або згладжування на межі склеєних сегментів, або дрібні зміни текстури обличчя від кадру до кадру. Крім того, вони здатні розпізнавати стабільні патерни руху, притаманні реальним відео, і помічати їх відсутність або спотворення у підроблених. Наприклад, справжні вирази обличчя розгортаються поступово за кілька кадрів, тоді як штучно накладена посмішка може з’явитися раптово між двома кадрами - це відхилення і фіксується 3D-конвольними шарами. Дослідження показали, що 3D-CNN, доповнені attention-механізмами, можуть особливо успішно виявляти тонкі локальні артефакти у відео і покращують результати детекції порівняно з покадровими 2D-моделями [13].

Рекурентні нейронні мережі - інший шлях моделювання часового виміру в відеопослідовностях. Усього зазначають два види рекурентних архітектур, зокрема LSTM (Long Short-Term Memory) або GRU (Gated Recurrent Unit). Ці мережі оперують послідовностями ознак, поступово «згадуючи» інформацію про попередні кадри при обробці наступних. Типовим рішенням є комбінована модель CNN+LSTM: спочатку згорткова нейромережа витягає високорівневі ознаки з кожного кадру, а потім послідовність таких ознак подається в LSTM, яка навчена виявляти аномальні динамічні патерни. LSTM може, наприклад, відстежувати чи плавно змінюються риси обличчя впродовж часу, чи не «стрибають» очі або рот між кадрами і т.д. Було виявлено, що залучення RNN для моделювання часової складової суттєво підвищує повноту виявлення

підробок у відео [10]. Рекурентні архітектури добре виявляють довготривалі залежності - наприклад, якщо в глибинному фейку поступово «пливе» положення обличчя чи деградує якість деталей до кінця відео, LSTM це помітить, тоді як окрема перевірка кадрів може не зважити на такий дрібний дрейф. Втім, RNN мають і свої обмеження: вони складніше паралеляться, а головне - можуть втрачати ефективність на дуже довгих послідовностях. Для відео тривалістю в кілька хвилин простий LSTM погано запам’ятовує початок ролика до його кінця. До того ж, якщо у фейковому відео аудіо розсинхронізоване з відео, рекурентна модель може плутатися. Практика показала, що LSTM-моделі без спеціальних модифікацій мають труднощі на великих і складних даних - з довготривалими відео, особливо якщо присутні зміщення між звуком і відеорядом [10].

У галузі комп’ютерного бачення все ширше застосування знаходять трансформери - моделі без згорток, що покладаються лише на механізм самоуваги. Спершу їх успішно застосували для статичних зображень (ViT - Vision Transformer), а згодом адаптували для відеопослідовностей. Існують різні варіанти Video Transformer: деякі просто додають час як ще один вимір самоуваги - аналізуючи одночасно просторові і часові зв’язки між пікселями та фрагментами кадрів, інші застосовують більш складні схеми, наприклад, чергують блоки просторової уваги, в межах окремого кадру та часової уваги, між кадрами, або вводять ієрархічну багаторівневу структуру, що спочатку обробляє короткі відрізки відео, а потім узагальнює на всю послідовність. Перевагою трансформерів є велика гнучкість у моделюванні залежностей: вони легко захоплюють і локальні деталі, і глобальний контекст. Для задачі виявлення підробок це означає, що модель може одночасно бачити і дрібні артефакти на обличчі, і загальну картину розвитку сцени. Наприклад, Multiscale Vision Transformer v2 (MViTv2) - модель, що використовує багаторівневі ознаки показала підвищену здатність захоплювати складні шаблони як у просторі, так і в часі. Використання подібних відео-трансформерів у детекції deepfake наразі знаходиться на етапі активних

досліджень, але перші результати обнадійливі. Такі моделі можуть перевершувати CNN або LSTM за рахунок кращої узгодженості при обробці довгих відео і кращого врахування контексту сцени. Водночас трансформери потребують дуже багато даних для навчання, тому часто їх доводиться навчати на синтетичних чи суміжних даних або використати попереднє навчання.

Глибинне навчання кардинально змінило і підхід до аналізу аудіо. Якщо раніше основну роль грали експертні ознаки, на кшталт згаданих MFCC, то зараз дедалі частіше використовуються аудіо-ембеддинги, які мережа навчається генерувати сама. Приклад - модель RawNet, що складається зі згорткових і рекурентних блоків і приймає на вхід сирий звуковий сигнал, одразу видаючи оцінку реальний або підробка. Вона навчається виокремлювати з хвильової форми такі особливості, які найкраще різнять справжній голос від штучного. Інший приклад - використання попередньо навчених моделей мовлення. Зокрема, моделі на кшталт Wav2Vec 2.0, навченої на великій кількості аудіозаписів, здатні продукувати узагальнені векторні представлення мовлення. Ці представлення можуть слугувати ознаками для класифікатора: дослідники показали, що з їх допомогою можна успішно детектувати фейки навіть без тонкої підгонки під кожен окремий голос. Сучасні архітектури часто поєднують кілька підходів: наприклад, формують мультирівневі аудіо-ознаки - спектрограми, кепстральні коефіцієнти, ознаки тону і подають їх в багатоканальну нейронну мережу, що агрегує всю цю інформацію.

Особливу категорію глибинних методів становлять моделі, що перевіряють синхронізацію руху губ з мовленням. Це можна вважати мультимодальним підходом, але його можна виділити окремо через специфічність завдання. Моделі на кшталт SyncNet навчаються на парі відеоряд обличчя - аудіо мовлення і виводять міру невідповідності між ними. Фактично, SyncNet - це дві глибинні мережі (для відео і для аудіо), об’єднані спільним шаром, що оцінює кореляцію: якщо звук і зображення узгоджені, ембеддинги будуть близькими. Для детекції фейків, в яких часто трапляються

«ліпсінк»-маніпуляції такі моделі незамінні. Вони здатні помітити не лише повний розсінхрог, а й більш тонкі розбіжності - наприклад, вимову фонеми, яка не відповідає положенню губ. В сучасних роботах системи перевірки синхронізації часто поєднуються з іншими детекторами. У 2025 році була представлена система TrueSync, яка інтегрує дві критичні ознаки: аналіз ліпсінку і моніторинг кліпання очей. Архітектурно TrueSync складається з двох модулів: CNN-LSTM, що відстежує патерн кліпань, та SyncNet, що оцінює синхронність аудіо і відео; їх результати потім об’єднані для прийняття рішення. Такий гібридний підхід виявився дуже ефективним: поєднання мікрорухів і міжмодальної узгодженості дозволило значно підвищити точність детекції складних підробок. Загалом, аналіз аудіо-візуальної узгодженості - перспективний напрям, адже підробити одночасно і автентичний голос, і ідеально підігнати під нього артикуляцію обличчя надзвичайно важко, особливо коли йдеться про довгі речення, швидку мову чи нестандартну вимову. Саме тому більшість мультимодальних дипфейк-детекторів так чи інакше включають компонент перевірки синхронності або консистентності між каналами [15].

2.4 Інтеграція модальностей: стратегії злиття ознак

Ефективність мультимодального аналізу значною мірою залежить від того, як саме об’єднані ознаки різних модальностей у моделі. Існують різні стратегії такої інтеграції - раннє злиття, пізнє, гібридні схеми, ансамблювання моделей. Кожна з них має свої переваги і недоліки. У цьому розділі буде розглянуто основні підходи до злиття ознак та прогнозів і роль вагових коефіцієнтів довіри при комбінуванні модальностей.

Раннє злиття ознак. Ця стратегія передбачає об’єднання даних різних модальностей на ранніх етапах обробки - фактично ще до винесення окремих

рішень по кожній з них. Технічно це може бути конкатенація або інше злиття векторів ознак, отриманих від двох потоків, одразу після кількох шарів мережі. Наприклад, модель може спочатку пропустити зображення і звук через відповідні згорткові шари, потім перетворити їх до співмірних векторів і склеїти в один спільний вектор, який подати далі в глибоку мережу для спільного аналізу. Таким чином, починаючи з цього об'єднаного представлення, всі наступні шари «бачать» вже змішану інформацію і можуть виявляти складні кореляції між модальностями. Плюсом раннього злиття є те, що модель максимально рано отримує повну картину і теоретично може вловити навіть ті взаємозалежності, які проявляються на низькорівневих ознаках. Наприклад, мережа може навчитися співвідносити певний шум у високочастотному спектрі аудіо з ледь помітним рябом пікселів на обличчі, якщо ці артефакти виникають разом внаслідок роботи одного генератора. Загалом раннє злиття дозволяє моделі раніше виявляти міжмодальні невідповідності. Дослідження показують, що правильно реалізоване раннє об'єднання може дати виграш у точності: наприклад, метод з раннім злиттям аудіо-візуальних ознак перевершив ряд одно-модальних підходів на тестових наборах ViolenceVD і NPDI. Втім, складність раннього злиття у можливому дисбалансі інформації: один канал може домінувати і «приглушувати» інший на спільних шарах, якщо їх відносна значущість не вирівняна.

Пізнє злиття ознак. Протилежний підхід - коли кожна модальність обробляється майже до кінця окремо, і лише на фінальному етапі їх результати комбінуються. В найпростішому випадку це може бути усереднення або голосування між двома окремими детекторами. У рамках єдиної нейромережі пізнє злиття зазвичай реалізують як конкатенацію високорівневих ознак перед фінальним класифікаційним шаром [14]. Тобто, модель до певного рівня має дві гілки - візуальну і аудіальну, які можуть навіть бути різної архітектури, наприклад, ResNet для відео і LSTM для аудіо. На передостанньому шарі ці гілки сходяться: їх вихідні вектори з'єднуються в один, який опрацьовує декілька нейронів і дає підсумковий результат справжнього або фейку. Пізнє

злиття зручне тим, що не вимагає спеціально узгоджувати просторово-часові масштаби модальностей - кожен експерт вирішує свою задачу і тільки потім їх думки зводяться. Це спрощує архітектуру і дозволяє, наприклад, легко підключати вже навчені окремо моделі як чорні скриньки. Недолік пізнього злиття - потенційна втрата деталей міжмодальних взаємозв'язків. Якщо аномалія проявляється лише у невідповідності каналів, то два роздільні детектори цього не виявлять. Тому пізнє злиття інколи доповнюють спеціальними механізмами: наприклад, вводять окремий «вузол» синхронності, або ж тренують фінальний класифікатор так, щоб він ловив статистичні залежності між виходами гілок. Гібридне злиття. Щоб отримати переваги обох підходів, дослідники все частіше вдаються до багаторівневого злиття. Гібридна стратегія означає, що модальності взаємодіють на кількох етапах: частково на ранніх шарах, потім можливо знову розходяться, і знов підключаються на пізніх шарах. Мета - ґрунтовно вивчити комплементарність ознак. На ранніх рівнях модель схоплює базові відповідності, наприклад, звук шуму руху корелює з рухом об'єкта на відео, на середніх - середньорівневі патерни, на фіналі - приймає рішення із залученням і окремих оцінок, і спільних ознак. Реалізацій такого підходу багато: від простого дублювання точок злиття до складних модулів. Приклад - «A review of deep learning based multimodal forgery detection for video and audio», які запропонували Multi-level Multimodal Hybrid Fusion (M2HF) для задачі пошуку відео за текстовим запитом. Вони спочатку виконують раннє злиття - об'єднують візуальні ознаки (з аудіо та руховими ознаками, формуючи так звані «аудіо-наведені» та «рух-наведені» візуальні представлення. Потім на фінальному етапі додається пізнє злиття, де результати з різних гілок об'єднуються для отримання остаточного прогнозу. Така багаторівнева схема дозволила одночасно врахувати і асинхронність між модальностями через роздільну обробку там, де це потрібно, і їхню додаткову інформацію через ранню інтеграцію там, де це давало ефект, що значно покращило результати. В іншій роботі для рекомендацій TikTok-відео, де присутні візуальна, звукова і текстова

модальності, запропоновано навіть динамічно вибирати схему злиття для кожного відео за допомогою мета-навчання. Хоч це інша галузь, ідея потенційно корисна і для детекції: наприклад, модель може сама визначати, коли їй достатньо аналізувати тільки відео, якщо звук відсутній чи неінформативний, а коли слід приділити більше уваги аудіо-відео взаємодії. Гібридні методи злиття є гнучкими, але їх складніше тренувати через більшу кількість параметрів, потенційно більше ризику перенавчання і також вони потребують ретельного налаштування.

Ансамблювання моделей. Окрім об’єднання ознак всередині однієї моделі, існує підхід об’єднання кількох окремих моделей - так званий ансамбль або багатокласифікаторна система. В контексті мультимодальної детекції це зазвичай виглядає так: ми маємо окремий детектор для відео і окремий детектор для аудіо; вони дають свої оцінки: ймовірності фейку, або бали аномальності, які потім комбінуються певним правилом. Це пізнє злиття на рівні рішень, яке часто називають decision-level fusion або score fusion. Такий підхід зручний тим, що дозволяє використовувати найкращі одномодальні моделі - наприклад, взяти передовий детектор обличчя тренований на FaceForensics++ і передовий детектор голосу тренований на ASVspoof, і спробувати отримати вигоду з обох. Об’єднання може відбуватися різними способами: голосуванням - коли остаточне рішення «фейк» приймається, якщо хоча б один з детекторів впевнено сигналізує про фейк, або зваженим усередненням - кожному каналові призначається вага, і підсумковий бал - це середнє з оцінок, помножених на ваги, або навіть більш складними методами на кшталт стекінгу (stacked generalization), коли поверх виходів моделей навчається метакласифікатор [14]. Такі стратегії продемонстрували, що правильна комбінація дозволяє підвищити і точність, і ефективність детекції порівняно з окремими експертами. Особливо важливо, що такі ансамблі можуть адаптивно налаштовувати пороги і враховувати консистентність між модальностями: наприклад, якщо аудіодетектор дає слабкий сигнал фейку, а відеодетектор сильний, система може прийняти

рішення на користь фейку, але з меншою впевненістю; або навпаки, якщо обидва дають середній сигнал, але узгоджено, метакласифікатор може розпізнати такий патерн як ознаку підробки. У підсумку, decision-level fusion забезпечує велику гнучкість. Його мінус - потенційно низька роздільна здатність: якщо обидві моделі-потоки помиляються у якомусь випадку, то і ансамбль не допоможе. Крім того, якщо один з детекторів сильно гірший за інший, він може лише додати шуму в рішення. Тому при ансамблюванні часто вводять вагові коефіцієнти довіри до кожної моделі. Наприклад, аудіо-модальності можна призначити меншу вагу,

Підсумовуючи, інтеграція модальностей є тонким місцем мультимодальних систем. Раннє злиття дає можливість найглибшої взаємодії ознак, але складне в реалізації і може страждати від дисбалансу. Пізнє злиття простіше і дозволяє використати наперед навчених експертів, але менш чутливе до міжмодальних невідповідностей. Гібридні схеми намагаються поєднати краще з обох, ціною ускладнення моделі. Ансамблі і взагалі виходять за межі однієї нейронної мережі, об’єднуючи декілька - це підвищує надійність, якщо правильно налаштовано ваги і правила. На практиці нерідко застосовують кілька підходів одночасно: наприклад, мережа може мати і раннє, і пізнє злиття, та ще й входити до складу ансамблю з іншими моделями. Остаточна мета - максимально використати наявні модальності, виявити перехресні ознаки маніпуляції та зменшити ризик пропустити фейк через слабку одну ознаку.

2.5 Метрики оцінювання якості виявлення підроблених відео

Для кількісної оцінки ефективності методів детекції підробок застосовується набір стандартних метрик класифікації, доповнений специфічними показниками зі сфери біометричної автентифікації. Правильне

обрання і інтерпретація метрик є важливим, оскільки дає змогу об’єктивно порівняти алгоритми та зрозуміти, наскільки добре вони працюють у різних умовах.

Точність (Precision) і повнота (Recall). У двокласовій задачі (справжнє чи підроблене відео) ці метрики відіграють ключову роль. Точність в контексті детекції фейків зазвичай визначають як частку правильно виявлених фейкових відео серед усіх відео, які класифікатор позначив як фейкові. Тобто Precision відповідає на питання: наскільки «чисті» спрацювання детектора. Висока точність означає низький рівень хибних тривог (false positives) - більшість сигналів про фейк дійсно виявляються фальсифікатами. Повнота (Recall) - це частка виявлених фейкових відео серед усіх реально фейкових у вибірці. Тобто вона характеризує здатність детектора знаходити всі підробки, відповідаючи на питання: скільки фейків пропустили. Високий Recall означає, що метод ловить майже всі підроблені відео (низький рівень пропущених, false negatives). Між точністю і повнотою часто є напруга: можна налаштувати модель на більш «строгий» режим - сигналізує тільки якщо дуже впевнена. Тоді хибних спрацювань мало, висока точність, але частину справжніх фейків вона може пропустити, знизивши повноту або навпаки, на чутливий режим - ловить максимум підозрілих випадків, високий Recall, але серед них можуть бути і помилкові тривоги, падає Precision. Тому зазвичай використовують F1-міру - гармонійне середнє точності і повноти. F1-міра дає збалансовану оцінку: вона буде високою тільки якщо і Precision, і Recall досить високі одночасно. В задачах детекції фальсифікацій, де важливо і не пропускати підробки, і марно не тривожити користувачів, F1 є одним з основних індикаторів якості моделі.

Щоб проаналізувати компроміс між чутливістю і специфічністю детектора, розглядають ROC-криву (Receiver Operating Characteristic). Вона будується наступним чином: поріг рішення класифікатора варіюють від найсуворішого до найм’якшого, і для кожного порогу обчислюють True Positive Rate (TPR) - це доля фейків, правильно виявлених, та False Positive Rate (FPR) - доля справжніх відео, помилково визначених як фейк. ROC-крива

- це графік TPR vs FPR при зміні порогу. Діагональна пряма на такому графіку відповідає випадковому вгадуванню. Чим вище над цією діагоналлю проходить ROC-крива моделі, тим краща її відокремлююча здатність. Для отримання числової оцінки обчислюють AUC (Area Under Curve) - площу під ROC-кривою. Перевага AUC-ROC метрики в тому, що вона не залежить від конкретно обраного порогу і характеризує модель загалом. В багатьох роботах з дипфейк-детекції повідомляють саме AUC: наприклад, для провідних методів на наборі FaceForensics++ AUC перевищує 0.99, тоді як для деяких старіших методів або на складніших датасетах (DFDC) AUC знаходиться в межах 0.8–0.9. ROC-аналіз також дозволяє обрати оптимальний поріг під конкретні вимоги, наприклад, якщо потрібно мінімізувати FPR - потрібно обрати поріг, при якому FPR дуже малий, і подивитись який при цьому TPR (Recall) [16].

У задачах підтвердження особи або доступу традиційно використовуються метрики False Acceptance Rate та False Reject Rate. False Acceptance Rate (FAR) - це показник того, як часто система помилково пропускає неавторизований доступ; у нашому контексті це еквівалентно частці фейкових відео, які детектор помилково визнав справжніми. Іншими словами, FAR - це альтернативна назва FPR, але розглянута з позиції системи безпеки: відсоток пропущених загроз. False Rejection Rate (FRR), навпаки, визначає як часто система відхиляє легітимного користувача; в детекції фейків це частка справжніх відео, які помилково забраковані як можливі підробки. FAR і FRR особливо важливі, коли наш детектор є частиною системи доступу або перевірки, наприклад, відеоверифікація особи при онлайн-сервісах. В таких застосуваннях треба досягти компромісу: зробити FAR достатньо низьким, щоб зловмисник з фейковим відео майже напевно не пройшов, але не зависити FRR, щоб справжніх користувачів не відхиляти без потреби. Часто наводять також метрику EER (Equal Error Rate) - це той рівень помилки, при якому FAR = FRR. Геометрично на ROC-кривій це точка, де відстань до діагоналі мінімальна. Низьке значення EER означає, що модель може

одночасно тримати низькими і помилки першого, і другого роду. Для високоякісних детекторів дипфейків EER зазвичай дуже малий, кілька відсотків або меньше. У біометричних стандартах останніх років навіть відмовляються від використання EER як основного показника, оскільки на практиці поріг роботи системи обирається нерівним FAR і FRR, але в наукових публікаціях з детекції фейків EER продовжують звітувати як зручний агрегований показник ефективності [17].

Для мультимодальних детекторів, що поєднують кілька каналів інформації, при оцінці якості варто враховувати декілька додаткових моментів. По-перше, корисно окремо оцінювати вклад кожної модальності. Наприклад, проводять експеримент «тільки відео» або «тільки аудіо» або «обидві модальності» і дивляться, наскільки об’єднання покращує показники. Це дає уявлення про інформативність кожного каналу та про те, чи несе мультимодальність реальну користь. По-друге, мультимодальні набори даних часто містять різні типи фейків: тільки візуальні, тільки аудіо, або обидва разом - як, наприклад, набір FakeAVCeleb. Тому детектор перевіряють окремо на кожному підтипі атак: модель повинна добре ловити чисто відеопідробки, не спрацьовуючи на аудіо, і навпаки - виявляти аудіофейки, навіть якщо відео справжнє. Метрики обчислюють для кожного такого сценарію, що дозволяє виявити, чи немає в системі «сліпих зон». По-третє, оцінюється стійкість до якості даних: як змінюється точність чи AUC, якщо подати стиснене відео, шумний звук, неповний фрагмент тощо. Мультимодальні моделі інколи виявляються чутливішими до цього, наприклад, через втрату синхронізації під час просмотру кліпу або через різну якість каналів. Також враховується часова та обчислювальна ефективність: мультимодальні системи складніші, і їх потрібно оцінювати не тільки по метриках точності, а й по тому, чи можуть вони працювати в реальному часі та на яких ресурсах. В такому випадку метричними показниками можуть бути середній час обробки відеофрагменту, або використання пам’яті GPU. Їх теж зазначають при характеристиці моделі.

Таким чином, якість виявлення підроблених відео описується цілим спектром показників: Accuracy, Precision, Recall, F1 дають базову оцінку класифікації; ROC-крива і AUC характеризують поведінку детектора при різних порогах; FAR і FRR важливі для сценаріїв автентифікації, де помилки різного типу мають різні наслідки; EER - зручний агрегат для порівняння. А для мультимодальних моделей додатково аналізують ефективність по кожному каналу і в різних умовах. Комплексне використання цих метрик дозволяє всебічно оцінити сильні і слабкі сторони методу, що особливо важливо з огляду на постійне удосконалення як самих детекторів, так і методів генерації фейків.

2.6 Порівняльний аналіз переваг та недоліків існуючих методів

Незважаючи на значний прогрес у галузі, жоден з існуючих методів виявлення підробленого відео не є універсальним. Кожен підхід має свої сильні сторони і слабкі місця. Проведемо порівняльний аналіз, узагальнивши відомості про стійкість до нових типів фейків, вимоги до ресурсів, залежність від даних та вразливість до атак обходу для різних класів методів.

Одномодальні та мультимодальні методи. Історично одномодальні детектори були першою лінією оборони проти дипфейків. Вони показали хороші результати у контрольованих умовах - наприклад, виявляли певний відомий тип підробки на даних, схожих на тренувальні. Перевага одномодальних підходів у відносній простоті та фокусі: модель «знає», що шукати саме артефакти на обличчі, і не відволікається на інші сигнали. Такі методи вимагають менших обчислювальних ресурсів і можуть бути натреновані на більш вузьких наборах - лише зображення або лише аудіо. Однак, їх головний недолік - низька здатність до узагальнення на нові типи фейків. З розвитком дипфейк-технологій з’ясувалося, що моделі, які колись

успішно ловили певні артефакти, часто дають збій на більш досконалих поколіннях фейків. Наприклад, якщо перші дипфейк-відео грішили відсутністю кліпання очима або спотвореною формою обличчя, то сучасні моделі генерації це виправили і детектори, націлені тільки на ці ознаки, втратили ефективність. Більше того, з'явилися фейки, що комбінують модальності, проти яких окремо взятий візуальний або аудіодетектор безсилий. Мультимодальні методи, навпаки, краще пристосовані до таких випадків: аналізуючи і картинку, і звук, вони можуть підхопити невідповідність між каналами, навіть якщо кожен окремо виглядає правдоподібно. Практика підтверджує, що мультимодальні детектори перевершують одномодальні в точності та робастності на складних датасетах, де присутні різні види фальсифікацій. Їхній мінус - складність. такі моделі важче тренувати, вони часто мають більше параметрів, вимагають більше пам'яті і часу на інференс. Таким чином, для простіших або відомих завдань одномодальні детектори можуть бути практичнішими, але для нових і витончених атак мультимодальні методи - фактично необхідність.

Якщо порівнювати типи моделей, то згорткові мережі 2D або 3D CNN традиційно добре виявляють локальні артефакти і деталі зображення. Вони успішні на високоякісних кадрах, де підробка залишає «сліди» у вигляді аномальної текстури, невірних пікселів. Але CNN можуть втрачати ефективність, коли фейк став настільки реалістичним на рівні пікселів, що артефакти мінімальні або коли відео дуже низької якості чи сильно стиснене, і ці артефати не видно. Рекурентні мережі (LSTM) добре ловлять часові невідповідності - ривки, пропуски, несправжню динаміку міміки. Вони корисні проти фейків, що створені шляхом некоректного монтажу або зміни швидкості відео. Проте LSTM-ам важко зберігати довгу пам'ять, тому якщо аномалії проявляються тільки в далекому контексті, наприклад, голос на початку відео не відповідає голосу в кінці, простий LSTM може це не спіймати. Трансформери зі своєю глобальною увагою здатні теоретично побачити довільно довгі залежності і найдрібніші узгодження, але вони

потребують дуже багато даних і обчислень. На практиці трансформерні моделі іноді страждають від перенавчання: вони можуть підлаштуватися під специфічні особливості тренувального набору але недостатньо узагальнювати. Крім того, трансформери вимагають сильного обмеження розміру вхідних даних, інакше модель стає надто громіздкою. Це теж компроміс - зменшуючи роздільність, ми ризикуємо втратити ті локальні деталі, які викривають фальсифікат. Отже, вибір архітектури повинен враховувати природу очікуваних фейків: для «грубих» підробок з явними артефактами достатньо CNN; для фейків, що проявляються переважно у часовій розбіжності варто додати LSTM; для найскладніших, де все якісно і потрібен аналіз довгих контекстів доцільно залучити трансформер, але з увагою щодо його потреб у даних.

Складність моделі безпосередньо впливає на вимоги до обчислювальних ресурсів. Легкі моделі, наприклад, невеликі CNN як MesoNet або XceptionNet можуть працювати майже в реальному часі на звичайному GPU і навіть на CPU, що важливо для практичного розгортання. Проте їх точність може бути нижчою за важчі моделі. Важкі моделі - глибокі 3D ResNet, трансформери з багатьма шарами, часто дають кращу якість на тесті, але будуть повільні. Наприклад, система, що аналізує відео одночасно і на піксельному рівні, і на рівні звуку великим трансформером, може потребувати декілька секунд для обробки секунди відео, що неприйнятно у багатьох сценаріях. Тому інженери мусять балансувати: інколи дещо спрощена модель, але придатна для розгортання, краще, ніж максимально точна, але надто повільна. Сьогодні найкращі результати часто досягаються ансамблями великих моделей, але розгорнути такий ансамбль у звичайних умовах може бути нереально. Тому один з трендів - спрощення і знесення моделей (distillation), коли замість громіздкого ансамблю тренують одну компактну модель, що імітує його поведінку.

Залежність від якості та кількості даних. Ще один критичний фактор - якість навчальних даних і їх відповідність реальним умовам. Багато методів

чудово працюють на датасетах, але втрачають ефективність на реальних відео з соцмереж. Причин кілька: по-перше, генеративні моделі швидко вдосконалюються. Якщо детектор навчено на фейках 2018 року, то проти згенерованих у 2025 році він може бути майже безсилий. Вони суттєво відрізняються за характеристиками - це так званий виклик узагальнення на нові типи атак. Дослідники намагаються розв'язати це або регуляризацією, щоб модель не заточувалась на надто специфічні деталі, або тренуванням на якомога різноманітнішому корпусі фейків. Це зробити непросто, адже отримати новітні фейки для навчання - виклик, проте спільнота рухається до цього: створюються все більші відкриті датасети (DFDC, KoDF, FaceForensics++, CelebDF та ін.), які включають багатоманітні приклади. По-друге, якість відео в реальних умовах нижча: соцмережі та месенджери застосовують сильне стиснення, багато відео знімається на смартфони при поганому освітленні. Через це ті артефакти, які детектор шукає, можуть просто не проглядатися. Наприклад, модель, що звикла до 1080p відео, може втратити точність на тих самих сценах, стиснених до 360p бо піксельні ознаки згладилися. Або аудіодетектор, навчений на чистих записах, плутатиметься, якщо в реальному кліпі є фоновий шум або зниження бітрейту. Це проблема domain gap між даними тренування і застосування. Борються з цим через штучне погіршення якості при навчанні і через спеціальні архітектурні рішення, наприклад, додавання нормалізації по частотах, використання стійких до шуму ознак. Але повністю подолати це складно. Найкраще - мати в навчальному наборі реалістичні приклади. Загалом, метод, який показує високу ефективність на одному статичному датасеті, може виявитися непрактичним на інших. Тому нині велика увага приділяється перевірці на множині наборів і дослідженню, наскільки метод залежить від даних. Ідеальним вважається детектор, що навчився вловлювати фундаментальні відмінності між справжнім і штучним, а не просто запам'ятав артефакти конкретної програми генерації.

Як оборона не стоїть на місці, так і нападники шукають способи обдурити навіть наявні детектори. Один з напрямів - антифорензика (anti-forensics): творці фейків можуть навмисно модифікувати свої відео, щоб приховати ті ознаки, які шукає детектор. Наприклад, дізнавшись, що багато моделей ловлять відсутність кліпання, зловмисники додали алгоритм, що штучно вставляє кліпання у згенероване відео. Або виявивши, що детектори орієнтуються на спектральні невідповідності, стали додавати пост-обробку аудіо (шум, еквалізацію), щоб вирівняти спектр. Це змагання триває постійно. Більш того, можливі адверсарні атаки на рівні моделі, зокрема, шляхом незначного псування вхідного відео можна змусити нейромережевий детектор помилятися. Такі adversarial examples давно відомі в комп'ютерному баченні, і deepfake-детектори теж до них вразливі. Наприклад, можна взяти фейковий кадр, до якого базовий детектор ще впевнено застосував би мітку «фейк», і додати до кожного пікселя ледве помітне коригування. В результаті мережа почне класифікувати кадр як справжній. Для аудіо аналогічно: додавання ледь чутного високочастотного шуму може збити детектор. Це серйозна проблема безпеки, адже зловмисник може цілеспрямовано підготувати свій фейк проти існуючих засобів виявлення. Академічні дослідження відзначають цей виклик і наголошують на потребі розробки захистів від навмисного обходу Пропонуються різні рішення: від розширення тренувального набору адверсарними прикладами до використання більш робастних ознак, наприклад, аналіз на рівні більш абстрактних патернів, менш чутливих до дрібних змін пікселів [14]. Втім, універсального вирішення поки немає, щоразу, коли з'являється новий потужний детектор, незабаром демонструють і спосіб його обдурити шляхом точної підгонки фейку під нього. Одне з перспективних напрямів - поєднання детектора з системою виявлення адверсарних втручань, тобто модель не тільки визначає «фейк/не фейк», а й моніторить, чи не виглядає вхід сумнівно з точки зору атак, наприклад, чи не містить статистично малоймовірних для природного відео частотних

компонентів. Це схоже на антивірус, який не лише шукає відомі віруси, а й дивиться на підозрілу активність.

Окремо варто згадати, що deepfake-технології еволюціонують не тільки в межах «обличчя-голос». З’являються нові види медіа-маніпуляцій: повна синтезація людських постатей, генерація будь-яких відеосцен за текстовим описом, як приклад, напрямок text-to-video (модель Imagen Video чи Phenaki), заміна довкілля у відео - не тільки облич, а й будь-яких об’єктів. Такі моделі ґрунтуються на дифузійних генераторах, 3D-рендерінгу та інших підходах. Відповідно, детектори облич можуть бути безсилі проти, скажімо, підроблених відео з тваринами або пейзажами. Це поки виходить за рамки класичного означення дипфейку, але межі розширюються. Методи, що вивчалися на обличчях, можуть не перенестись на інші домени. Тому у перспективах - розробка всеохопних систем цифрової медіафорензіки, які могли б працювати не тільки з конкретною ознакою, а з загальними принципами генеративної підробки. Деякі дослідження йдуть шляхом пошуку універсальних «відбитків» ШІ, наприклад, виявлено, що нейромережеві генератори можуть залишати статистичні сліди в розподілі спектральних компонент, кореляціях шуму тощо, незалежно від контенту. Такі універсальні ознаки дали б великий плюс - детектор, побудований на них, був би стійкішим до різних видів фейків. Однак і зловмисники, знаючи про ці спроби, вдосконалюються: нещодавно показано, що сучасні дипфейки можуть навіть успадковувати фізіологічні сигнали від реальних прототипів, практично нівелюючи ту перевагу, яку мали методи на кшталт аналізу серцебиття. Це підтверджує, що гонка між генерацією та детекцією триває, і методи, стійкі на сьогодні, можуть потребувати перегляду завтра. Результати узагальнення наведені у таблиці 2.1.

Таблиця 2.1 – Порівняльна характеристика основних класів методів виявлення підробленого відеоконтенту.

Клас методів	**Переваги**	**Недоліки**
Одномодальні	Простіші та швидші. Фокус на специфічних ознаках, менші вимоги до даних однієї модальності	Не виявляють міжмодальних невідповідностей. Легко обходяться, якщо підробка якісна саме в цій модальності. Погано узагальнюються на комбіновані та нові типи фейків. можуть бути хиткими при зміні умов
Мультимодальні	Виявляють комплексні маніпуляції. Фіксують розбіжності між каналами. Краща точність на складних фейках, більша стійкість до локальних поліпшень фейку в окремих модальностях	Складні у реалізації. Потребують синхронних даних для навчання. Вимогливі до обчислювальних ресурсів. Довший час інференсу. Ризик перенавчання на конкретному датасет.
CNN (2D, 3D)	Добре виявляють локальні артефакти та аномалії текстур. Відносно швидкі на GPU. Існує багато перевірених архітектур	Можуть пропускати аномалії довготривалого характеру. Вразливі до згладжування та стиснення відео. Потребують достатньо багато даних.
RNN (LSTM/GRU)	Враховують часову складову. Ефективні для виявлення розсинхронізації, стрибків, аномальної динаміки. Можуть комбінуватися з CNN для просторово-часового аналізу	Мають обмежену довготривалу пам’ять. Важче паралелізуються, що збільшує час обробки. Можуть деградувати на дуже довгих відео або за наявності значних шумів у послідовності

Продовження таблиці 2.1

Клас методів	**Переваги**	**Недоліки**
Ансамблі	Забезпечують найвищу точність. Дають змогу комбінувати різні підходи.	Найповільніші та найскладніші в реалізації. Важко розгортаються поза лабораторними умовами. Мають високі вимоги до пам’яті та CPU/GPU
Трансформери	Забезпечують глобальну увагу та врахування довільно далеких залежностей. Гнучкі у моделюванні різних типів взаємозв’язків.	Великі моделі потребують багато даних і ресурсів. Високий ризик перенавчання. Повільна робота.

Наведена таблиця ілюструє, що вибір методу детекції залежить від пріоритетів: якщо потрібна максимальна точність для певного типу фейків - можна використати потужний ансамбль, але якщо важлива універсальність і швидкість, тоді доведеться жертвувати певною часткою точності. У реальних застосунках нерідко комбінують підходи: запускають легкий детектор в режимі реального часу для первинного відсіву, а сумнівні випадки перепроверяють важчою моделлю офлайн.

Отже, сучасний ландшафт методів виявлення підробленого відеоконтенту охоплює широкий спектр підходів - від одномодальних детекторів до складних мультимодальних архітектур глибинного навчання. Одномодальні рішення, що спираються на специфічні артефакти візуального ряду, аудіосигналу або текстової складової, сформували базу для технологій детекції та й надалі залишаються корисними для відносно простих або одноканальних фальсифікацій. Водночас їхня обмеженість полягає в тому, що вони практично не «бачать» міжмодальні неузгодженості, які дедалі частіше є ключовою ознакою складних дипфейків нового покоління. У відповідь на це провідну роль поступово посідають мультимодальні підходи, де інформація з

кількох модальностей об’єднується в єдиному обчислювальному контурі, що дає змогу помічати тонкі розбіжності між зображенням, звуком і текстом, непомітні за ізольованого аналізу каналів.

Серед архітектурних рішень у цій сфері виокремлюються як класичні конвеєри на основі CNN та рекурентних мереж для аналізу просторово-часової динаміки, так і сучасні трансформерні моделі з attention-механізмами, здатні враховувати довготривалі залежності та складні кореляції між різними типами сигналів. Важливим концептуальним виміром є стратегії інтеграції модальностей: раннє злиття ознак, пізнє поєднання рішень і гібридні схеми, доповнені ансамблевими підходами та ваговими коефіцієнтами довіри до окремих каналів. Саме від того, як саме реалізовано цю інтеграцію, залежить, чи буде система здатна використати комплементарність модальностей або навпаки - втратити міжмодальні взаємозв’язки. На цьому тлі особливого значення набувають метрики оцінювання - від базових показників точності, повноти й F1-міри до ROC/AUC та спеціалізованих біометричних критеріїв FAR/FRR, які дозволяють оцінити якість детекції у сценаріях автентифікації особи та контрольованого доступу.

Узагальнюючи, нинішнє покоління методів виявлення дипфейків демонструє перехід від вузькоспеціалізованих одномодальних рішень до комплексних мультимодальних систем, що поєднують різні архітектури глибинного навчання, гнучкі стратегії злиття ознак і розгорнуті підходи до оцінювання якості. Такі системи забезпечують найвищу точність на складних випадках і дозволяють виявляти фейки, які залишалися б непомітними за аналізу лише одного каналу інформації. Водночас вони стикаються з низкою викликів - потребою в значних обчислювальних ресурсах, проблемами узагальнення на нові типи маніпуляцій та вразливістю до навмисних атак обходу. Сукупність розглянутих підходів показує, що надійна детекція підроблених відео є багатофакторною задачею, де необхідно досягти балансу між точністю, робастністю, ефективністю та здатністю адаптуватися до еволюції генеративних технологій.

3 РЕАЛІЗАЦІЯ ПРОТОТИПУ МУЛЬТИМОДАЛЬНОГО ДЕТЕКТОРА ДИПФЕЙКІВ

3.1 Загальна архітектура системи мультимодального детектування

Розроблений прототип детектування дипфейків має модульну архітектуру, що відображає розподіл задач на дві основні модальності - візуальну (зображення обличчя у відео) та аудіальну (мовний сигнал). На вхід рішення подається відеофайл, який спочатку розділяється на дві гілки обробки. Перша гілка відповідає за аналіз відеоряду: з відео вилучаються окремі кадрові зображення, над якими здійснюється візуальний аналіз облич для виявлення можливих ознак маніпуляції. Друга гілка відповідає за обробку аудіодоріжки: звуковий сигнал, що супроводжує відео, виділяється і аналізується на наявність артефактів, характерних для синтезованого чи підробленого голосу. Кінцевим етапом є мультимодальне об’єднання отриманої інформації - результати аналізу кожної модальності інтегруються в єдиний критерій для ухвалення рішення про автентичність чи фальшивість відео.

Подібна архітектура ґрунтується на ідеї, що комбінування двох джерел даних - зображення обличчя та голосу здатне забезпечити більш надійне виявлення дипфейків, ніж аналіз лише однієї з модальностей. У сучасних дипфейках можливі різні сценарії маніпуляції: може бути підроблено лише обличчя, візуально при збереженні справжнього голосу, або навпаки - підроблено голос при справжньому відеоряді, або ж змінено обидві модальності одночасно. Тому використання лише одного типу ознак може бути недостатнім. Наприклад, якщо обличчя у відео замінене на інше (фейкове), але аудіо залишається справжнім, то чисто аудіальний детектор не виявить обману; аналогічно, в разі підробки лише голосу візуальний детектор не зафіксує фальсифікації. Мультимодальний підхід дозволяє виявити невідповідність між зображенням та звуком і тим самим підвищити точність

класифікації. Дослідження підтверджують, що об’єднання ознак з різних модальностей може дати багатшу інформацію для викриття дипфейків, хоча воно і потребує ретельного підходу до злиття ознак, аби справді перевершити точність найкращих одномодальних методів. Цей прототип реалізує такий підхід шляхом незалежної обробки кожного каналу з наступним прийняттям рішення на основі обох результатів.

Архітектура рішення складається з трьох основних компонентів:

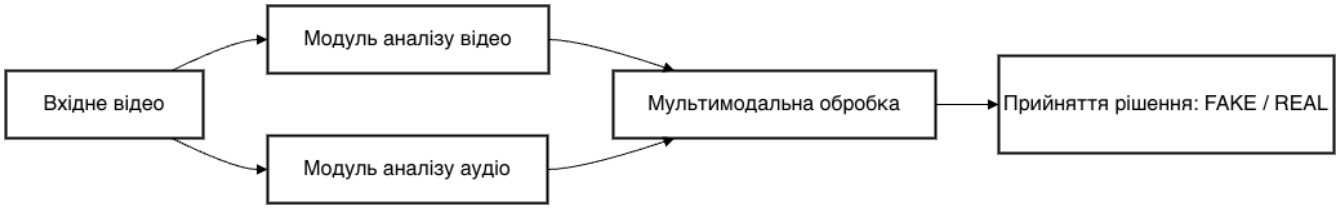

Рисунок 3.1 – Приклад архітектури прототипа мультимодального дипфейк-детектора

На рис. 3.1 кожен з цих компонентів можна уявити як окремий блок, з’єднаний потоками даних: відеокадри надходять до модуля візуального аналізу, аудіосигнал - до модуля аудіоаналізу, після чого їх виходи об’єднуються у фінальному блоці. Загалом, такий конвеєр обробки нагадує типову архітектуру, яка успішно використовується в задачах мультимедійної форензіки. Відсутність жорсткого зчеплення між гілками дозволяє гнучко модифікувати або покращувати кожен з модулів незалежно. У процесі розробки це дало змогу спочатку створити та налаштувати кожен з модулів окремо, а вже потім інтегрувати їх у єдину систему. Такий поетапний підхід до побудови мультимодальних систем відповідає кращим практикам проєктування складних моделей, оскільки спрощує налагодження і дозволяє оцінити внесок кожної компоненти в загальний результат.

3.2 Обґрунтування вибору моделей

У практичній частині роботи як базові компоненти системи було обрано дві попередньо навчені моделі з Hugging Face Hub: Gustking/wav2vec2-large-xlsr-deepfake-audio-classification для аудіоаналізу та shylhy/videomae-large-finetuned-deepfake-subset для відеоаналізу. Такий підхід зумовлений необхідністю отримати працездатний прототип мультимодальної системи без етапу повноцінного навчання моделей з нуля або масштабного донавчання, що потребує суттєвих обчислювальних ресурсів та тривалого циклу експериментів. У межах наявного середовища виконання було пріоритетизовано коректну інтеграцію, відтворюваність запуску та можливість подальшої заміни компонентів на власні моделі без перебудови загальної логіки конвеєра.

Вибір зазначених моделей є предметно й технічно обґрунтованим. По-перше, обидві моделі безпосередньо орієнтовані на задачу дипфейк-детекції у відповідній модальності та повертають інтерпретований результат у вигляді бінарної класифікації «FAKE/REAL», що суттєво спрощує узгодження виходів у мультимодальному модулі. По-друге, вони підтримуються стандартними інтерфейсами бібліотеки Transformers і коректно завантажуються через AutoModelForAudioClassification / AutoModelForVideoClassification разом із відповідними засобами препроцесингу (AutoFeatureExtractor, AutoImageProcessor). У контексті прототипування це принципово важливо, оскільки мінімізує обсяг «ручної» обробки даних і зменшує ризик помилок сумісності між форматом входів моделі та реальними файлами. По-третє, практична сумісність цих моделей у межах одного проєкту проявляється в тому, що обидва модулі формують вихід у єдиному форматі ймовірностей для класів «fake» та «real», що дозволяє застосувати єдину схему об’єднання результатів без додаткових перетворень або складної калібровки на цьому етапі.

Окремо слід підкреслити, що обрані моделі забезпечують коректний симбіоз на рівні конвеєра даних. Аудіомодуль очікує сигнал, приведений до моноформату та частоти дискретизації 16 кГц, що в прототипі досягається стандартною попередньою обробкою; відеомодуль працює із фіксованою кількістю кадрів та уніфікованим розміром зображень, що узгоджується з типовими конфігураціями відеокласифікації на основі трансформерних архітектур і реалізується через вибірку кадрів та приведення їх до потрібних параметрів.

У підсумку, підхід із використанням попередньо навчених моделей розглядається як компроміс інженерна стратегія для побудови прототипу: вона дозволяє зосередитися на правильній організації конвеєра, відтворюваності інференсу та процедурі інтеграції модальностей. При цьому архітектура прототипу зберігає модульність і допускає розвиток, наприклад, заміну використаних контрольних точок на власні донавчені моделі або підключення альтернативних компонентів без потреби переглядати загальну логіку мультимодальної обробки та прийняття рішення.

3.3 Реалізація модуля аудіоаналізу та підготовка сигналу до інференсу

У межах прототипу мультимодальної системи окремий модуль аудіоаналізу було реалізовано як незалежну гілку конвеєра, що отримує на вході аудіосигнал і повертає ймовірності належності до класів «fake» та «real». Практична доцільність такого поділу полягає в тому, що аудіослід синтетичного походження часто проявляється інакше, ніж візуальні артефакти, а отже, потребує власного спеціалізованого інструментарію. У прототипі модуль побудовано на готовій, уже навченій моделі з репозиторію Hugging Face: Gustking/wav2vec2-large-xlsr-deepfake-audio-classification, що

дозволило зосередитися на інженерній інтеграції та побудові працездатного конвеєра без витрат на ресурсоємне навчання з нуля.

Вибір саме цієї моделі для аудіогілки обґрунтовано сумісністю з поставленою задачею та придатністю до прямого використання в межах бібліотеки transformers. Модель позиціонується як fine-tuning рішення для завдання класифікації deepfake-аудіо і на сторінці моделі наведено показники якості: F1=0.95 на власних оціночних даних, а також метрики на підмножині ASVspoof2019 (Accuracy=0.9286, Precision=0.9999, Recall=0.9205, F1-Score=0.9363, EER=0.0401). Ці значення є важливими з практичної точки зору, оскільки демонструють не лише загальну збалансованість класифікації (F1), а й низький рівень помилки рівноваги (EER), що традиційно застосовується в задачах антиспуфінгу мовлення та перевірки автентичності аудіо.

З архітектурного погляду, в основі моделі лежить сімейство Wav2Vec2-XLS-R (зокрема вказано базову модель facebook/wav2vec2-xls-r-300m у дереві моделі), тобто представник самонавчальних (self-supervised) трансформерних підходів до побудови універсальних акустичних репрезентацій. Концептуально це означає, що аудіосигнал перетворюється на послідовність високорівневих ознак, після чого над ними працює класифікаційна голова, яка навчається розрізняти автентичний та синтетичний сигнал. Для практичної частини роботи це важливо тим, що модель має узгоджений інтерфейс подання даних (через AutoFeatureExtractor) і повертає логіти класів, з яких коректно отримуються ймовірності шляхом softmax, що безпосередньо відповідає вимозі прототипу повертати інтерпретовані значення «fake/real».

Реалізований у роботі аудіомодуль складається з послідовності кроків попередньої обробки та інференсу, які забезпечують узгодження «сирого» аудіофайлу з очікуваннями моделі. На першому етапі аудіодані зчитуються бібліотекою soundfile, після чого здійснюється приведення до моно (у випадку стереодоріжок) шляхом усереднення каналів. Далі забезпечується нормалізація частоти дискретизації до 16 кГц, що є цільовим значенням у конфігурації прототипу і відповідає типовим налаштуванням для мовних

моделей такого класу; для цього застосовано інтерполяційний ресемплінг до нової довжини сигналу. Після узгодження сигналу за частотою дискретизації формується тензорний пакет ознак через AutoFeatureExtractor із вказанням sampling_rate=16000 та padding=True, що забезпечує стабільність подачі даних на модель навіть для аудіо різної тривалості. На етапі інференсу модель повертає логіти двох класів, які перетворюються у ймовірності softmax, після чого формується результат у вигляді пари значень fake і real, сумісної з мультимодальним об’єднанням у наступних підрозділах.

З погляду інтеграції в загальну систему, аудіомодуль у мультимодальному сценарії може отримувати вхід не лише як окремий аудіофайл, а й як доріжку, витягнуту з відео. У прототипі цей крок реалізовано через виклик ffmpeg, який зберігає доріжку у тимчасовий WAV-файл з моно-каналом і частотою дискретизації 16 кГц. Така інженерна зв’язка забезпечує відтворюваність та однозначність експериментів: незалежно від початкового контейнера й параметрів відеофайлу на вхід класифікатора завжди потрапляє сигнал у стандартизованому форматі, що мінімізує ризик помилок, пов’язаних із кодеками, різними sampling rate або каналами. У результаті аудіогілка стає автономною і може бути замінена на іншу модель без модифікації відеогілки чи механізму мультимодального прийняття рішення, що відповідає вимогам прототипного проєктування та подальшого розвитку системи.

У підсумку реалізований аудіомодуль у складі прототипу забезпечує повний цикл попередньої обробки та класифікації звукової доріжки з метою оцінювання її автентичності, причому ключовим результатом є отримання інтерпретованих ймовірнісних оцінок належності сигналу до класів «fake» та «real». Практична цінність модуля полягає в тому, що він стандартизує вхідні дані (перетворення в моно та приведення частоти дискретизації до 16 кГц), зменшуючи вплив технічної неоднорідності вихідних файлів і забезпечуючи коректну сумісність із попередньо навченою моделлю класу wav2vec2. Вихід у форматі двох імовірностей після softmax дозволяє використовувати результат не лише як бінарне рішення, а як міру впевненості моделі, що є

важливим для подальшої інтеграції у мультимодальний конвеєр, де аудіооцінка виступає незалежним сигналом доказовості поряд із відеоаналізом. Таким чином, аудіогілка підвищує стійкість системи до випадків, коли візуальні артефакти підробки є слабко вираженими або навмисно замаскованими, тоді як ознаки синтезу/конверсії голосу можуть зберігатися в спектрально-часових характеристиках мовлення і бути використані для прийняття більш обґрунтованого підсумкового рішення.

Рисунок 3.2 – Схематичне відображення алгоритму роботи аудіо-моделі.

З метою перевірки коректності роботи розробленого аудіомодуля було виконано окремий тестовий прогін на невеликій контрольній вибірці, сформованій із двох типів аудіозаписів: прикладів із підробленим мовленням та прикладів із реальним мовленням. Як джерело тестових матеріалів було обрано відкритий датасет ASVspoof 2019, оскільки саме на даних цього класу (реальна мова проти синтезованої/підробленої) навчалися та оцінювалися багато сучасних підходів до аудіофорензіки, а також тому, що він забезпечує наявність еталонних міток для верифікації прогнозів моделі. Для експерименту було відібрано кілька коротких аудіофрагментів і приведено їх до формату, сумісного з прототипом тестового конвеєра (формат WAV із цільовою частотою дискретизації 16 кГц), після чого ці файли були подані на вхід аудіомодулю, реалізованому на основі готової попередньо навченої моделі Gustking/wav2vec2-large-xlsr-deepfake-audio-classification.

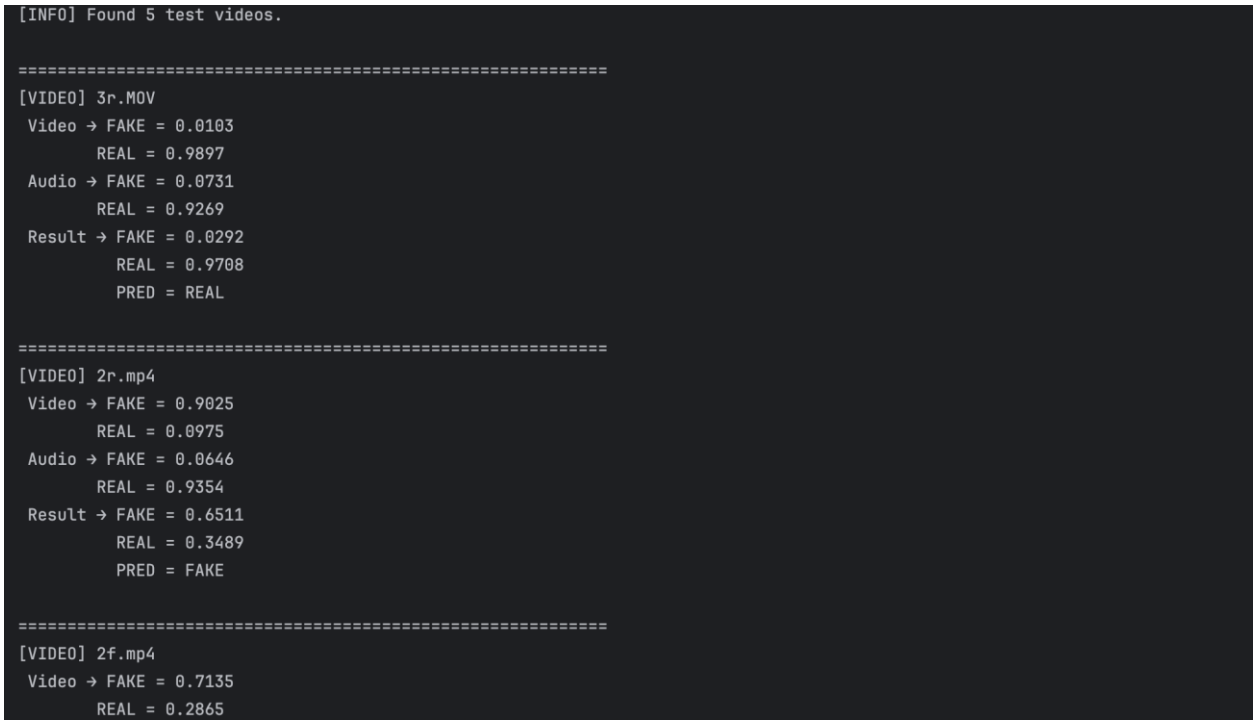

Рисунок 3.3 – Приклад часткового виводу результатів аудіо-модулю

Тестування виконувалося у режимі пакетної обробки: скрипт послідовно зчитував аудіофайли з директорії, здійснював уніфікацію сигналу (зокрема

перетворення у моно у випадку багатоканального запису та ресемплінг до 16 кГц за потреби), формував вхідні ознаки через стандартний екстрактор бібліотеки Transformers і обчислював логіти моделі без режиму навчання. Після нормалізації логітів за допомогою softmax отримувалися два значення - оцінка ймовірності класу FAKE та оцінка ймовірності класу REAL; кінцевий клас визначався за правилом максимуму. Саме ця перевірка забезпечила, що у вихідному форматі відображаються правильні значення FAKE/REAL і правильне підсумкове рішення.

Таблиця 3.1 – Результати тестування аудіо-модулю

Test audio	**FAKE**	**REAL**	**PRED**
1f.wav	0.8960	0.1040	FAKE
1r.wav	0.0674	0.9326	REAL
2f.wav	0.9019	0.0981	FAKE
2r.wav	0.0677	0.9323	REAL
3f.wav	0.9117	0.0883	FAKE
3r.wav	0.0686	0.9314	REAL

За результатами проведеного тесту було оброблено шість аудіофайлів, серед яких три відповідали підробленим прикладам (умовно позначені як «f»), а три реальним (умовно позначені як «r»). Для підроблених прикладів модель стабільно формувала високу оцінку класу FAKE та низьку оцінку класу REAL: для файлів 1f.wav, 2f.wav, 3f.wav значення FAKE становили 0.8960, 0.9019, 0.9117 відповідно, тоді як значення REAL перебували на рівні 0.1040, 0.0981, 0.0883. Для реальних прикладів спостерігалася дзеркальна картина: для файлів 1r.wav, 2r.wav, 3r.wav модель надала високі значення REAL (0.9326, 0.9323, 0.9314) при низьких значеннях FAKE (0.0674, 0.0677, 0.0686). В усіх шести

випадках підсумкове рішення PRED збігалося з очікуваним типом аудіозапису: підроблені фрагменти були класифіковані як FAKE, а реальні як REAL, без помилок на цій контрольній вибірці.

Отримані результати підтверджують працездатність прототипу аудіомодуля на практичному рівні та коректність реалізованого конвеєра обробки сигналу й інтерпретації виходу моделі.

3.4 Реалізація модуля відеоаналізу та підготовка відеокадрів до інференсу

Відеомодуль у прототипі реалізовано як окрему гілку конвеєра, що приймає на вхід відеофайл, виконує стандартизоване виділення репрезентативних кадрів та передає їх у попередньо натреновану модель відеокласифікації. Така побудова відповідає загальній логіці прототипу: за відсутності обчислювальних ресурсів для повноцінного навчання власної відеомережі на великому корпусі даних, доцільним є використання вже навченого (fine-tuned) рішення з відкритого репозиторію та його інтеграція в єдину мультимодальну схему. У межах роботи було обрано модель shylhy/videomae-large-finetuned-deepfake-subset, яка є донавченою версією MCG-NJU/videomae-large на наборі даних Deepfake Detection Challenge (DFDC) та опублікована як готовий чекпойнт для задачі відеокласифікації. Базова архітектура VideoMAE належить до сімейства трансформерних моделей для відео, у яких ознаки формуються через поділ відеопотоку на послідовність візуальних “токенів” і подальше контекстне узгодження через механізм уваги; у першоджерелі ця ідея подається як підхід самонавчання з маскуванням частини відеоспостережень, що підвищує ефективність попереднього навчання на великих масивах відеоданих.

Ключовим етапом відеомодуля є підготовка вхідних даних у форматі, сумісному з очікуваннями моделі. На практиці це означає, що “сирий”

відеофайл спочатку декодується бібліотекою imageio, після чого з нього формуються кадри у вигляді масивів пікселів (RGB). Оскільки реальні відео можуть мати різну тривалість, частоту кадрів і якість, пряме подання всіх кадрів у модель є непрактичним, а також не відповідає типовим режимам роботи відеотрансформерів, які розраховані на фіксовану довжину відеопослідовності. Тому в прототипі застосовано стратегічно просте, але методично коректне рішення: відбирається фіксована кількість кадрів, рівномірно розподілених по всій довжині відео. Технічно це реалізовано через побудову індексів за допомогою linspace від першого до останнього кадру з подальшим зчитуванням кадрів за цими індексами. Таким чином забезпечується репрезентативність відбору: у модель потрапляють не лише початкові або випадкові фрагменти, а кадри, що «покривають» відео цілісно, що є принципово важливим для задачі детекції дипфейків, де артефакти можуть проявлятися нерівномірно в часі.

Після формування набору кадрів здійснюється їх перетворення у формат тензорів, який очікує модель VideoMAE. У прототипі для цього використано AutoImageProcessor.from_pretrained(..., use_fast=False), який завантажує параметри препроцесингу, узгоджені з конкретним чекпойнтом. На цьому етапі виконуються стандартизовані операції нормалізації та приведення розміру до єдиного формату, зокрема масштабування до 224×224 пікселі, що є типовим розміром для багатьох трансформерних відео-моделей та прямо закладене в обраний прототип. Далі сформовані вхідні дані переносяться на обчислювальний пристрій, що дозволяє виконувати інференс швидше порівняно з CPU-режимом у типових сценаріях локального тестування.

Інференс відеомоделі у прототипі реалізовано у режимі без градієнтів (torch.no_grad()), що відповідає практиці експлуатації нейромереж у задачах класифікації та зменшує споживання пам’яті. На виході модель повертає logits, які інтерпретуються як «сирі» оцінки класів до нормалізації. Для отримання ймовірностей застосовується softmax, після чого формується словник результатів із ключами FAKE/REAL. У підсумку відеомодуль повертає два

числові значення: ймовірність «fake» та ймовірність «real», які надалі використовуються в мультимодальній частині як один із двох незалежних сигналів прийняття рішення.

Рисунок 3.4 – Схематичне відображення алгоритму роботи відео-моделі.

```
[INFO] Loading model: shylhy/videomae-large-finetuned-deepfake-subset

[INFO] Found 6 video files

============================================================
[FILE] r1.mp4
 FAKE = 0.1108
 REAL = 0.8892
 PRED = REAL

============================================================
[FILE] r2.mp4
 FAKE = 0.9651
 REAL = 0.0349
 PRED = FAKE
```

Рисунок 3.5 – Приклад часткового виводу результатів аудіо-модулю

Після реалізації модуля відеоаналізу було виконано контрольне тестування на невеликій вибірці з шести відеофайлів (три еталонно реальні та три фейкові). Для інференсу застосовано попередньо натреновану модель VideoMAE shylhy/videomae-large-finetuned-deepfake-subset, яка повертає розподіл імовірностей за двома класами fake та real. Обробка відео здійснювалася шляхом рівномірного семплювання 32 кадрів по всій тривалості ролика та приведення кадрів до розміру 224×224, після чого послідовність кадрів подавалася до моделі для отримання фінального прогнозу.

Таблиця 3.2 – Результати тестування відео-модулю

Test video	**FAKE**	**REAL**	**PRED**
r1.mp4	0.1108	0.8892	REAL
r2.mp4	0.9651	0.0349	FAKE
r3.mp4	0.2047	0.7953	REAL
f2.mp4	0.3057	0.6943	REAL

Продовження таблиці 3.2

Test video	FAKE	REAL	PRED
f3.mp4	0.3140	0.6860	REAL
f1.mp4	0.3813	0.6187	REAL

Отримані результати демонструють, що модель частково «бачить» відмінність між класами, однак у даному експерименті не забезпечує стабільного розділення реальних і підроблених відео. Для двох реальних файлів (r1.mp4, r3.mp4) зафіксовано коректний прогноз REAL з високими значеннями ймовірності реального класу (0.8892 та 0.7953 відповідно). Водночас один реальний файл (r2.mp4) був помилково класифікований як FAKE з дуже високою ймовірністю фейку (0.9651), що вказує на чутливість моделі до певних “шумових” характеристик відео. Для трьох фейкових файлів (f1.mp4, f2.mp4, f3.mp4) модель формально обрала клас REAL, проте важливо, що ймовірність фейку в них систематично вища (0.3057–0.3813), ніж у коректно розпізнаних реальних прикладах (≈ 0.1108–0.2047). Це означає, що вхідні фрагменти містять певні ознаки, які модель асоціює з маніпуляцією, але їх «сила» недостатня, аби переважити клас REAL у фінальному рішенні.

Причини такого ефекту зазвичай пов’язані не з помилкою реалізації конвеєра, а з обмеженнями узагальнення попередньо натренованої моделі на даних іншої природи. По-перше, модель донавчена на певному датасеті, тому при виборці медіа з іншого датасету вона може втрачати дискримінативність: фейк виглядає «занадто якісним» у термінах артефактів, на які навчалась модель, і тому оцінюється ближче до реального.

Загалом тест підтвердив працездатність модуля відеоінференсу та коректність формату вхідних даних для моделі. Водночас результати вказують на необхідність подальшого підсилення відеогілки, а також на доцільність мультимодального підходу: у випадках, коли відеодетектор дає

«прикордонні» оцінки, додавання аудіомодуля може суттєво підвищити стійкість фінального рішення за рахунок незалежних ознак підробки.

3.5 Реалізація мультимодального об’єднання

Мультимодальний модуль у розробленому прототипі виконує дві ключові функції: приводить виходи аудіо та відеодетектора до узгодженого інтерпретованого формату, формує єдине підсумкове рішення шляхом контрольованого об’єднання оцінок.

Першим практично важливим завданням інтеграції є коректна інтерпретація класів моделей. Оскільки аудіо та відеомоделі вже навчені та реалізовані окремо, вони можуть використовувати різні схеми маркування (наприклад, fake/real, deepfake/real, spoof/bonafide), у прототипі реалізовано механізм уніфікації класів через аналіз config.label2id та config.id2label. Це критично для надійності конвеєра: якщо індекси класів визначено помилково, система отримує формально «правильні» ймовірності, але з інверсією змісту (коли «real» інтерпретується як «fake»). Саме тому в модулі інтеграції застосовано нормалізацію назв міток та підтримку синонімів (fake/deepfake/spoof тощо), що знижує ризик семантичної помилки при використанні сторонніх моделей.

Другим завданням є приведення виходів до єдиної шкали. Обидві моделі повертають логіти, які перетворюються на ймовірності через softmax. У підсумку кожна гілка генерує пару значень:

$$\rho_v^{fake}, \rho_v^{real}; \rho_a^{fake}, \rho_a^{real}, \tag{3.1}$$

де сума по двох класах дорівнює 1. Для злиття використовується саме ρ^{fake}, а ρ^{real} у прототипі відновлюється як 1- ρ^{real}, що спрощує подальшу логіку і робить її прозорою у звіті.

Також у прототипі застосовано пізнє злиття на рівні рішень (decision-level / late fusion) у вигляді зваженої суми:

$$\rho^{fake} = w_v \cdot \rho_v^{fake} + w_\alpha \cdot \rho_\alpha^{fake}, \quad (3.2)$$

де $w_v + w_\alpha = 1$. У реалізації за замовчуванням встановлено $w_v = 0.7$ та $w_\alpha = 0.3$. Такий вибір є інженерно обґрунтованим для демонстраційної системи: відеомодальність у задачі дипфейк детекції часто дає більш «різку» сигналізацію у випадку маніпуляцій обличчя, тоді як аудіомодальність може бути менш стабільною через шум, різні кодеки, мікрофони, варіативність мовлення та інші доменні фактори. Водночас ваги у прототипі не є «істинно оптимальними»: вони задаються конфігураційно і мають розглядатися як початкова евристика, яку в подальшому доцільно калібрувати на валідаційній підвибірці.

Після обчислення ρ^{fake} значення обмежується на [0;1](через clip), а фінальний клас визначається правилом:

FAKE, якщо $\rho^{fake} \geq 0.5$, (3.3)

REAL, якщо $\rho^{fake} < 0.5$ (3.4)

Додатково в модулі реалізовано перевірки працездатності кожної модальності, оскільки у практичних умовах відеофайли можуть містити неповний набір медіаданих. Типовою ситуацією є відсутність або некоректність аудіодоріжки (наприклад, відео без звуку, пошкоджений контейнер, нестандартний кодек), що унеможливлює формування коректного

аудіовходу для моделі. Для запобігання некоректним оцінкам і «мовчазним» помилкам конвеєр виконує технічну валідацію після спроби екстракції аудіо: перевіряється факт створення файлу та його ненульовий розмір. У випадку негативного результату формується виняткова ситуація, що фіксує неможливість застосування аудіомодальності до конкретного зразка. Такий підхід забезпечує однозначність логіки: рішення не «домальовується» з невалідних даних, а супроводжується чіткою індикацією обмеження для даного вхідного файлу.

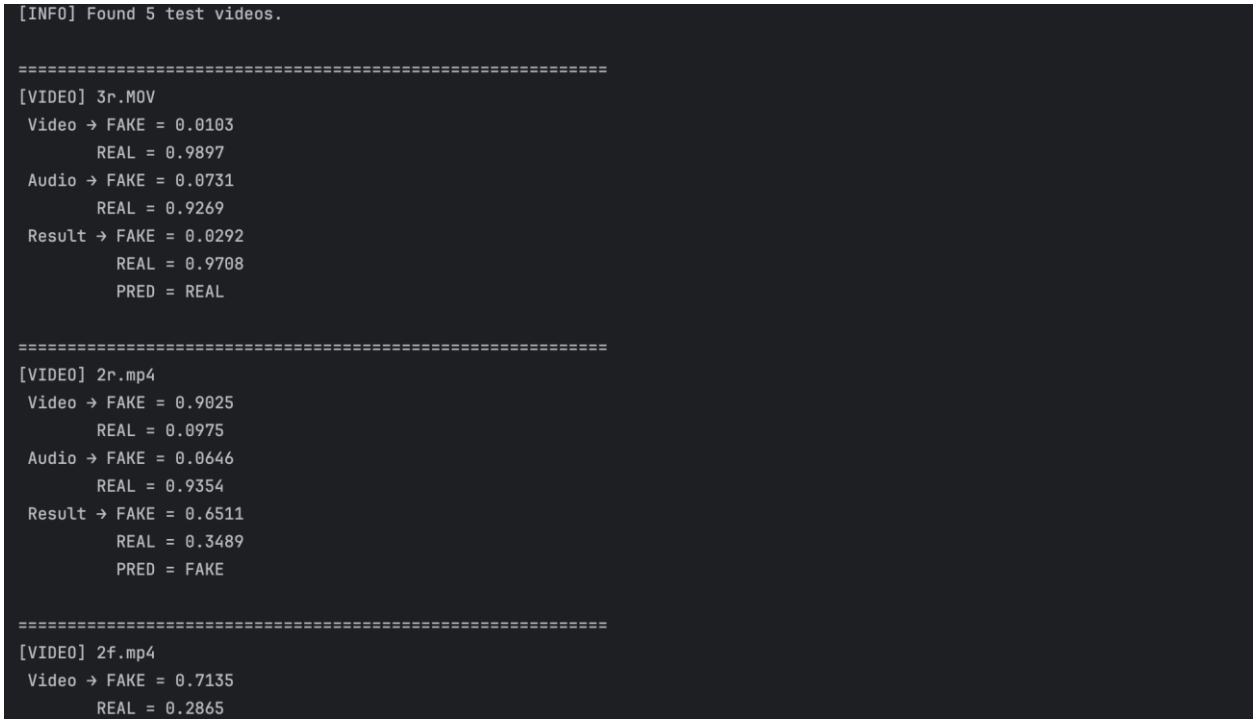

Рисунок 3.6 – Приклад часткового виводу результатів мультимодального модулю

Після реалізації мультимодального конвеєра було виконано контрольне тестування на невеликій підвибірці відеофайлів, сформованій із датасету Localized Audio Visual DeepFake Dataset (LAV-DF**)**, а також додано власний контрольний приклад як приклад реального відео.

Таблиця 3.3 – Результати тестування прототипа мультимодального детектора

Test file	**Video results (F/R)**	**Audio results (F/R)**	**Multimodal results (F/R)**	**Prediction**
3r.MOV	0.0103/0.9897	0.0731/0.9269	0.0292/0.9708	REAL
2r.mp4	0.9025/0.0975	0.0646/0.9354	0.6511/0.3489	FAKE
2f.mp4	0.7135/0.2865	0.0942/0.9058	0.5277/0.4723	FAKE
1f.mp4	0.9675/0.0325	0.2966/0.7034	0.7662/0.2338	FAKE
1r.mp4	0.9614/0.0386	0.2004/0.7996	0.7331/0.2669	FAKE

Результати тестування показали, що контрольне власне відео 3r.MOV було стабільно класифіковано як REAL як відеомоделлю, так і аудіомоделлю, а також після мультимодального об’єднання. Зокрема, для 3r.MOV отримано низькі значення ймовірності підробки в обох каналах (Video FAKE=0.0103; Audio FAKE=0.0731), що закономірно привело до інтегральної оцінки Result FAKE=0.0292 і підсумкового прогнозу REAL. Для решти тестових файлів інтегральний модуль повернув прогноз FAKE, однак спостерігалася характерна асиметрія: відеоканал демонстрував дуже високі значення FAKE (переважно 0.71–0.97), тоді як аудіоканал у кількох випадках залишався ближчим до REAL (Audio FAKE близько 0.06–0.20). За заданих ваг злиття така конфігурація приводить до домінування відеомодальності в інтегральному рішенні та, відповідно, до класифікації зразків як FAKE навіть тоді, коли аудіоканал не підтверджує високу ймовірність підробки. Практично це відображає типове обмеження прототипів на попередньо натренованих моделях без доменної адаптації: відеодетектор може бути чутливим до артефактів компресії, параметрів зйомки, монтажних вставок або специфіки підвибірки, що спричиняє завищення FAKE для частини реальних матеріалів;

водночас аудіоканал може залишатися відносно «спокійним», якщо голосовий сигнал не містить ознак синтезу. У межах цього тесту на LAV-DF та контрольному прикладі 3r.MOV отримані результати підтверджують працездатність наскрізного конвеєра (коректна екстракція, передобробка, інференс і злиття) та одночасно демонструють необхідність подальшої калібрації ваг або введення режиму «невизначеності» для випадків, коли модальності формують суперечливі сигнали.

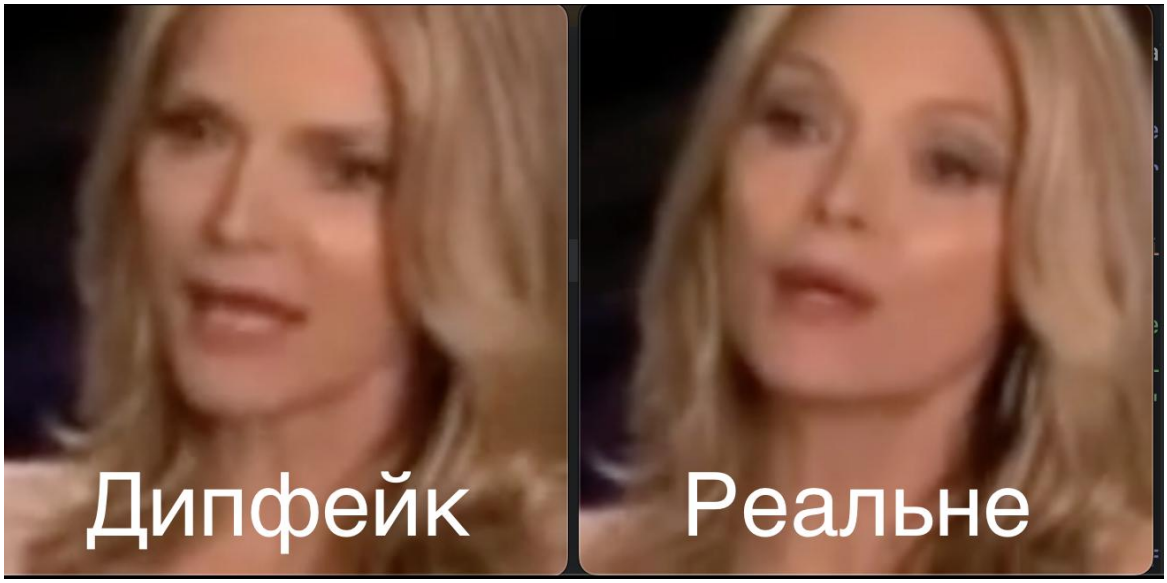

Рисунок 3.7 – Приклад протестованих відеофайлів у мультимодальному модулі

Проведене контрольне тестування мультимодального конвеєра підтвердило працездатність наскрізної реалізації: система коректно виконує завантаження попередньо натренованих моделей, витяг аудіодоріжки, формування вхідних представлень для обох модальностей, інференс та інтеграцію оцінок у спільне рішення. Разом із тим результати на решті тестових файлів виявили характерну особливість прототипу: у низці випадків відеомодель формує значно вищі значення FAKE порівняно з аудіомоделлю, а фіксовані ваги злиття приводять до домінування відеоканалу в інтегральному

рішенні. Унаслідок цього мультимодальна оцінка класифікує частину прикладів як FAKE навіть тоді, коли аудіоканал дає сигнал, ближчий до REAL.

3.6 Масштабування прототипу до практично придатної системи

Масштабування розробленого прототипу до повноцінного мультимодального детектора є концептуально прямим, оскільки базова архітектура вже реалізує ключові компоненти системи: окремий інференс аудіо та відеомодальності, уніфікацію класів та узгодження виходів, інтеграцію оцінок і формування підсумкового рішення. Отже, перехід від демонстраційного варіанту до практично придатного рішення може виконуватися шляхом поетапної заміни та розширення елементів прототипу без зміни загальної логіки конвеєра.

Шлях перетворення прототипу в реальний детектор полягає у заміщенні сторонніх попередньо натренованих моделей на власні або донавчені під цільовий домен. У поточній реалізації використано готові моделі з, що забезпечує швидкий запуск, але обмежує керованість якості й переносимість на реальні умови. Якщо одномодальні компоненти аудіо та відеомодель навчити на даних, що відповідають цільовим сценаріям застосування, система отримує суттєво вищу узгодженість з доменом, можливість контрольованої валідації та прогнозовану поведінку на “важких” прикладах (компресія, шум, різні камери/мікрофони). При цьому сам мультимодальний блок інтеграції може зберігатися практично без змін: він виступає як універсальний інтерфейс, який приймає виходи модальностей і формує кінцевий вердикт.

Також доцільно додати формалізований контур оцінювання. Реальна система має працювати не лише в режимі демонстраційного запуску, а й у режимі масового тестування, де результати накопичуються та інтерпретуються статистично. Для цього прототип можна розширити

механізмом автоматичного формування звітів: збереження виходів відео, аудіо та мультимодального каналів у структурований формат (CSV/JSON), фіксація міток датасету, підрахунок базових метрик і формування узагальнень. Такий компонент перетворює прототип на експериментальний стенд, який дозволяє порівнювати різні версії моделей, ваг злиття та порогів у стандартизований спосіб. Також до детектора варто додати механізми керованої інтеграції замість фіксованого злиття. У практичних умовах модальності можуть мати різну якість сигналу або навіть бути частково відсутніми. Тому прототип може бути доповнений адаптивною логікою: використання різних ваг залежно від впевненості кожної модальності, застосування політики «невизначено» для конфліктних випадків, а також окремий сценарій fallback, коли аудіо не витягнуто або його якість є недостатньою. Така надбудова не змінює структуру конвеєра, але суттєво підвищіть практичну коректність рішень.

Додатковим напрямом масштабування є узгодження інтерпретації виходів різних моделей на рівні ймовірностей. Навіть за однакової двокласової схеми (fake/real) різні одномодальні моделі можуть демонструвати різний рівень “впевненості”, що впливає на стабільність злиття. У практично придатній системі доцільно передбачити етап калібрування виходів (на валідаційній підвибірці) або використання узгоджених шкал оцінювання, щоб інтеграція відображала реальний інформаційний внесок кожної модальності, а не особливості конкретного розподілу softmax. Окремо варто підсилити систему механізмами контролю якості вхідних даних, оскільки мультимодальна детекція значною мірою залежить від технічних параметрів медіа. Для реального використання доцільно доповнити прототип автоматичними перевірками наявності аудіодоріжки, мінімальної тривалості фрагмента, а також базових метаданих, які можуть пояснювати відхилення у результатах (агресивна компресія, низька роздільна здатність, нестабільний FPS). Такі перевірки не змінюють модельної частини, проте дозволяють коректніше трактувати виходи детектора та зменшують ризик хибних

висновків у випадках, коли якість сигналу є явно недостатньою для надійного інференсу. Ще одним перспективним шляхом розвитку прототипу є розширення мультимодального аналізу за рахунок сигналів, що відображають узгодженість між модальностями. Поточний варіант late fusion агрегує лише підсумкові оцінки аудіо та відеодетектора, однак не використовує крос-модальні залежності, які можуть бути інформативними у випадках часткових підробок.

ВИСНОВКИ

Під час виконання магістерської роботи було комплексно розглянуто проблему виявлення підроблених відео як актуальну загрозу кібер та інформаційній безпеці, окреслено основні типи фальсифікацій від простих маніпуляцій до дипфейків, а також систематизовано підходи до аналізу відеоконтенту. У теоретичній частині узагальнено принципи побудови систем детекції, охарактеризовано одномодальні методи для відео й аудіо та обґрунтовано доцільність мультимодального підходу як засобу підвищення стійкості до часткових підробок і доменних спотворень. Окремо враховано роль джерел даних і протоколів тестування, а також нормативно-правові та етичні аспекти використання й детекції синтетичного медіаконтенту, що є критичними для практичного застосування таких рішень у реальних умовах.

У практичній частині було реалізовано прототип мультимодального детектора, що поєднує відео та аудіомодальності в єдиному конвеєрі інференсу. Прототип включає: завантаження попередньо натренованих моделей для відео та аудіо, уніфікацію інтерпретації класів через аналіз, виділення аудіодоріжки з відео через ffmpeg, отримання ймовірностей класів за кожною модальністю та пізнє злиття на рівні рішень (decision-level fusion) у вигляді зваженої суми, де внесок відеоканалу визначено як пріоритетний. Така реалізація продемонструвала працездатність архітектури інтеграції: мультимодальний блок виступає універсальним інтерфейсом, який приймає виходи одномодальних компонентів, узгоджує їх семантику та формує кінцевий вердикт у формі інтерпретованих оцінок (FAKE/REAL).

Контрольне тестування прототипу було виконано на обмеженій вибірці з датасету Localized Audio Visual DeepFake Dataset (LAV-DF) із додатковим власним контрольним відеофайлом. Результати показали, що система коректно ідентифікувала власний приклад як REAL, тоді як більшість інших тестових відео були класифіковані як FAKE. Водночас зафіксовано випадки,

коли не всі приклади, визначені прототипом як FAKE, фактично є підробками, що вказує на обмеження демонстраційного рішення: залежність від доменної відповідності попередньо навчених моделей, потенційну некаліброваність імовірностей та чутливість до характеристик запису: компресія, шум, кодеки, якість мікрофона/камери. Отримані спостереження підтверджують доцільність мультимодального принципу як такого, але водночас підкреслюють необхідність системної валідації та адаптації компонентів для досягнення практично значущої точності.

Додатково встановлено, що для прототипів, побудованих на сторонніх попередньо натренованих моделях, ключовим фактором надійності є коректна інтерпретація класів та узгодження виходів модальностей. Реалізований у роботі механізм уніфікації міток знижує ризик семантичної інверсії класів і забезпечує стабільність конвеєра при заміні моделей або оновленні їх конфігурацій. Таким чином, отримано практичний результат у вигляді універсального мультимодального «каркасу», який може бути повторно використаний для інтеграції інших одномодальних компонентів без суттєвої перебудови програмної логіки.

Перспективним напрямом подальшого розвитку визначено масштабування прототипу до повноцінного детектора шляхом заміщення сторонніх моделей на власні або донавчені під цільовий домен, введення формалізованого контуру оцінювання з накопиченням результатів і метрик, а також удосконалення логіки інтеграції. Окремо доцільним є доповнення системи інструментами експлуатаційної придатності: структурованим логуванням, конфігураційним керуванням параметрами, пакетною обробкою великих масивів відео та підготовкою звітів, що забезпечують відтворюваність експериментів. Сукупно ці кроки створюють основу для перетворення демонстраційного рішення на прикладну систему, придатну до використання у задачах інформаційної безпеки та протидії медіаманіпуляціям.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Paris B., Donovan J. Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. New York: Data & Society, 2019.

2. Nadeem M. S., Franqueira V. N. L., Zhai X., Kurugollu F. A survey of deep learning solutions for multimedia visual content analysis. IEEE Access. 2019.

3. Li Y., Lyu S. Exposing DeepFake Videos by Detecting Eye Blinking // IEEE International Workshop on Information Forensics and Security (WIFS). 2018.

4. Afchar D., Nozick V., Yamagishi J., Echizen I. MesoNet: a Compact Facial Video Forgery Detection Network // 2018 IEEE International Workshop on Information Forensics and Security (WIFS). 2018.

5. Güera D., Delp E. Deepfake Video Detection Using Recurrent Neural Networks // 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2018.

6. Nguyen T. T., Nguyen C. M., Nguyen D. T., Nguyen D. T., Nahavandi S. Deep learning for deepfakes creation and detection: A survey. arXiv preprint arXiv:1909.11573, 2019.

7. Zhao Z., Zhang H., Wu Z., Zeng A., Liu Y. ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection // IEEE Transactions on Information Forensics and Security. 2023.

8. Huang M., Liang Z., Zhang P., Li H., Zhan D., Wang S., Chan M. Spatiotemporal Attention-Based Deepfake Detection // Proceedings of the 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), 2024.

9. Zhang Y., Li X., Wang C., Yuan Y., Liu C., Zhang L., Hu H., Zhang J. VidTr: Video Transformer Without Convolutions // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021.

10. Zhang X., Chu J., Zhao J., Li H., Li S. ERF-BA-TFD+: A Multimodal Model for Audio-Visual Deepfake Detection. arXiv preprint arXiv:2508.17282, 2025.

11. Deepfakes Now Mimic Human Heartbeats, Defeating Key Detection Method. IDTechWire, 2024.

12. Zhang B., Cui H., Nguyen V., Whitty M. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. Sensors. 2025.

13. Mundra S. K., Bhat S. Detecting Multi-Modal Deepfake Videos in Latent Space. SN Applied Sciences. 2025.

14. Tan D., Yang Y., Niu C., Li S., Yang D., Tan B. A review of deep learning based multimodal forgery detection for video and audio. Discover Applied Sciences. 2025.

15. El-Taj H., Alammari F., Alkhowaiter J., Bogari L., Essa R. Deepfake Detection Based on Visual Lip-sync Match and Blink Rate. International Journal of Computational and Experimental Science and Engineering (IJCESEN), 2025.

16. Altuncu E., Franqueira V. N. L., Li S. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. arXiv preprint arXiv:2208.10913, 2022.

17. Hu X. A Comprehensive Evaluation of Deepfake Detection Methods: Approaches, Challenges and Future Prospects // ITM Web of Conferences. 2025.

ДОДАТОК А

Код програми

А.1 Аудіо модуль

```
import os
import numpy as np
import soundfile as sf
import torch
from transformers import AutoFeatureExtractor, AutoModelForAudioClassification
MODEL_ID = "Gustking/wav2vec2-large-xlsr-deepfake-audio-classification"
BASE_DIR = os.path.dirname(os.path.abspath(__file__))
AUDIO_DIR = os.path.join(BASE_DIR, "audio")
TARGET_SR = 16000
device = "mps" if torch.backends.mps.is_available() else "cpu"
def load_audio(path, target_sr=TARGET_SR):
    data, sr = sf.read(path)
    if len(data.shape) > 1:
        data = np.mean(data, axis=1)
    if sr != target_sr:
        duration = len(data) / sr
        new_len = int(duration * target_sr)
        data = np.interp(
            np.linspace(0, len(data), new_len),
            np.arange(len(data)),
            data,
        )
    return data
def _find_label_index(model, wanted_label: str):
    wanted = wanted_label.lower()
    # 1) label2id якщо є
    label2id = getattr(model.config, "label2id", None)
    if isinstance(label2id, dict):
        for k, v in label2id.items():
            if k.lower() == wanted:
```

```
                return int(v)
    # 2) id2label
    id2label = getattr(model.config, "id2label", None)
    if isinstance(id2label, dict):
        for i, lab in id2label.items():
            if str(lab).lower() == wanted:
                return int(i)
    raise ValueError(f"Cannot find label '{wanted_label}' in
model config (id2label/label2id).")
def classify_audio(path, model, extractor):
    audio_data = load_audio(path)
    inputs = extractor(
        audio_data,
        sampling_rate=TARGET_SR,
        return_tensors="pt",
        padding=True
    ).to(device)
    with torch.no_grad():
        logits = model(**inputs).logits
    probs        =        torch.softmax(logits,        dim=-
1)[0].detach().cpu().numpy()
    idx_real = _find_label_index(model, "real")
    idx_fake = _find_label_index(model, "fake")
    real_p = float(probs[idx_real])
    fake_p = float(probs[idx_fake])
    return {"fake": fake_p, "real": real_p}
     def main():
    print(f"[INFO] Loading model: {MODEL_ID}")
    model                                                    =
AutoModelForAudioClassification.from_pretrained(MODEL_ID).to
(device)
    extractor                                                =
AutoFeatureExtractor.from_pretrained(MODEL_ID)
    files = [
        f for f in os.listdir(AUDIO_DIR)
        if f.lower().endswith((".wav", ".mp3", ".flac"))
    ]
    print(f"[INFO] Found {len(files)} audio files.\n")
```

```
    for fname in sorted(files):
        full_path = os.path.join(AUDIO_DIR, fname)
        print("=" * 60)
        print(f"[FILE] {fname}")
        try:
            result    =    classify_audio(full_path,    model,
extractor)
            print(f" FAKE = {result['fake']:.4f}")
            print(f" REAL = {result['real']:.4f}")
            pred = "FAKE" if result["fake"] >= result["real"]
else "REAL"
            print(f" PRED = {pred}")
        except Exception as e:
            print(f"[ERROR] Cannot process {fname}: {e}")
    print("\n[FINISHED]")
if __name__ == "__main__":
    main()
```

A.2 Відео модуль

```
import os
import numpy as np
import torch
import imageio
from      transformers      import      AutoImageProcessor,
AutoModelForVideoClassification

MODEL_ID = "shylhy/videomae-large-finetuned-deepfake-subset"
VIDEO_DIR = "video"

device = "mps" if torch.backends.mps.is_available() else "cpu"

def load_video_frames(path, num_frames):
    reader = imageio.get_reader(path)
    try:
        try:
            total = reader.count_frames()
        except Exception:
            frames_list = [f for f in reader]
```

```
            total = len(frames_list)
            reader.close()
            reader = imageio.get_reader(path)

        if total < 1:
            raise ValueError("Video is empty or nonreadable")

        idxs = np.linspace(0, total - 1,
num_frames).astype(int)

        frames = []
        for i in idxs:
            try:
                frames.append(reader.get_data(int(i)))
            except Exception:
                pass

        if len(frames) == 0:
            raise ValueError("Can't extract frames")

        while len(frames) < num_frames:
            frames.append(frames[-1])

        return frames[:num_frames]
    finally:
        try:
            reader.close()
        except Exception:
            pass

def find_label_index(model, target_names=("fake", "real")):

    label2id = getattr(model.config, "label2id", None) or {}
    id2label = getattr(model.config, "id2label", None) or {}

    norm = {str(k).strip().lower(): int(v) for k, v in
label2id.items()}
```

```
    fake_keys = ["fake", "fa ke", "deepfake", "spoof",
"bonafide_fake", "1"]
    real_keys = ["real", "bonafide", "genuine", "0"]

    fake_idx = None
    real_idx = None

    for k in fake_keys:
        if k in norm:
            fake_idx = norm[k]
            break
    for k in real_keys:
        if k in norm:
            real_idx = norm[k]
            break
    if fake_idx is None or real_idx is None:
        inv = {int(i): str(lbl).strip().lower() for i, lbl in
id2label.items()}
        for i, lbl in inv.items():
            if fake_idx is None and ("fake" in lbl or
"deepfake" in lbl or "spoof" in lbl):
                fake_idx = i
            if real_idx is None and ("real" in lbl or
"bonafide" in lbl or "genuine" in lbl):
                real_idx = i

    return fake_idx, real_idx

def classify_video(path, model, processor):
    model.eval()

    num_frames = int(getattr(model.config, "num_frames", 16))
    image_size = int(getattr(model.config, "image_size",
224))

    frames = load_video_frames(path, num_frames=num_frames)

    inputs = processor(
```

```
        frames,
        return_tensors="pt",
        size={"height": image_size, "width": image_size},
    ).to(device)

    with torch.inference_mode():
        outputs = model(**inputs)
        probs    =    torch.softmax(outputs.logits,    dim=-
1)[0].detach().cpu().numpy()

    fake_idx, real_idx = find_label_index(model)

    id2label = getattr(model.config, "id2label", {})
    scored     =     {str(id2label.get(i,     f"label_{i}")):
float(probs[i]) for i in range(len(probs))}

    if fake_idx is None or real_idx is None:
        return {
            "fake":                  float(scored.get("fake",
scored.get("Fake", 0.0))),
            "real":                  float(scored.get("real",
scored.get("Real", 0.0))),
            "all": scored
        }

    return {
        "fake": float(probs[fake_idx]),
        "real": float(probs[real_idx]),
        "all": scored
    }

def run_all():
    print(f"[INFO] Loading model: {MODEL_ID}")

    processor = AutoImageProcessor.from_pretrained(MODEL_ID,
use_fast=False)
```

```
    model                                                    =
AutoModelForVideoClassification.from_pretrained(MODEL_ID).to
(device)

    video_files  =  [f  for  f  in  os.listdir(VIDEO_DIR)  if
f.lower().endswith((".mp4", ".mov", ".avi"))]
    print(f"\n[INFO] Found {len(video_files)} video files")

    for fname in video_files:
        path = os.path.join(VIDEO_DIR, fname)
        print("\n" + "=" * 60)
        print(f"[FILE] {fname}")

        try:
            res = classify_video(path, model, processor)

            fake_score = res["fake"]
            real_score = res["real"]
            pred  =  "FAKE"  if  fake_score  >=  real_score  else
"REAL"

            print(f" FAKE = {fake_score:.4f}")
            print(f" REAL = {real_score:.4f}")
            print(f" PRED = {pred}")

        except Exception as e:
            print(f"[ERROR] Failed to process {fname}: {e}")

    print("\n[FINISHED]")

if __name__ == "__main__":
    run_all()
```

A.3 Мультимодальний модуль

```
import os
import subprocess
import numpy as np
import soundfile as sf
```

```
import imageio.v2 as imageio
import torch

from transformers import (
    AutoFeatureExtractor,
    AutoModelForAudioClassification,
    AutoImageProcessor,
    AutoModelForVideoClassification,
)

AUDIO_MODEL_ID   =   "Gustking/wav2vec2-large-xlsr-deepfake-
audio-classification"
VIDEO_MODEL_ID  =  "shylhy/videomae-large-finetuned-deepfake-
subset"

BASE_DIR = os.path.dirname(os.path.abspath(__file__))
VIDEO_DIR = os.path.join(BASE_DIR, "testvideo")
TEMP_AUDIO_PATH = os.path.join(BASE_DIR, "temp_audio.wav")

TARGET_SR = 16000

device = "mps" if torch.backends.mps.is_available() else "cpu"
print(f"[INFO] Device: {device}")

def _norm_label(x: str) -> str:
    return               str(x).strip().lower().replace("-",
"").replace("_", "").replace(" ", "")

def resolve_fake_real_indices(model):
    label2id = getattr(model.config, "label2id", None) or {}
    id2label = getattr(model.config, "id2label", None) or {}
    norm_l2i   =   {_norm_label(k):   int(v)   for   k,   v   in
label2id.items()}

    fake_aliases = [
        "fake",     "deepfake",     "spoof",     "synthetic",
"manipulated", "tampered", "attack"
    ]
```

```
    real_aliases = [
        "real",    "bonafide",    "genuine",    "authentic",
"original", "normal", "human"
    ]

    fake_idx = None
    real_idx = None

    for k in fake_aliases:
        nk = _norm_label(k)
        if nk in norm_l2i:
            fake_idx = norm_l2i[nk]
            break

    for k in real_aliases:
        nk = _norm_label(k)
        if nk in norm_l2i:
            real_idx = norm_l2i[nk]
            break

    if fake_idx is None or real_idx is None:
        for i, lbl in id2label.items():
            nl = _norm_label(lbl)
            if fake_idx is None and any(_norm_label(a) in nl
for a in fake_aliases):
                fake_idx = int(i)
            if real_idx is None and any(_norm_label(a) in nl
for a in real_aliases):
                real_idx = int(i)

    return fake_idx, real_idx

#   AUDIO

def load_audio(path, target_sr=TARGET_SR):
    data, sr = sf.read(path)

    if len(data.shape) > 1:
```

```
        data = np.mean(data, axis=1)

    if sr != target_sr:
        duration = len(data) / sr
        new_len = int(duration * target_sr)
        data = np.interp(
            np.linspace(0, len(data), new_len),
            np.arange(len(data)),
            data,
        )

    return data


def classify_audio(path, audio_model, audio_extractor):
    audio_model.eval()

    audio_data = load_audio(path)

    inputs = audio_extractor(
        audio_data,
        sampling_rate=TARGET_SR,
        return_tensors="pt",
        padding=True
    ).to(device)

    with torch.inference_mode():
        logits = audio_model(**inputs).logits
        probs       =       torch.softmax(logits,       dim=-
1)[0].detach().cpu().numpy()

    fake_idx,                    real_idx                    =
resolve_fake_real_indices(audio_model)
    id2label = getattr(audio_model.config, "id2label", {})
    scored  =  {str(id2label.get(i,  f"label_{i}")).lower():
float(probs[i]) for i in range(len(probs))}

    if fake_idx is None or real_idx is None:
```

```
        fake            =            float(scored.get("fake",
scored.get("deepfake", 0.0)))
        real            =            float(scored.get("real",
scored.get("bonafide", 1.0 - fake)))
        return {"fake": fake, "real": real}

    return    {"fake":    float(probs[fake_idx]),    "real":
float(probs[real_idx])}


def extract_audio(video_path, out_wav=TEMP_AUDIO_PATH):
    cmd = [
        "ffmpeg", "-y",
        "-i", video_path,
        "-vn",
        "-ac", "1",
        "-ar", str(TARGET_SR),
        out_wav,
    ]
    subprocess.run(cmd,            stdout=subprocess.DEVNULL,
stderr=subprocess.DEVNULL)
    return out_wav

#   VIDEO

def load_video_frames(path, num_frames):
    reader = imageio.get_reader(path)
    try:
        try:
            total = reader.count_frames()
        except Exception:
            frames_list = [f for f in reader]
            total = len(frames_list)
            reader.close()
            reader = imageio.get_reader(path)

        if total < 1:
            raise ValueError("Video is empty or nonreadable")
```

```
idxs = np.linspace(0, total - 1, num_frames).astype(int)

        frames = []
        for i in idxs:
            try:
                frames.append(reader.get_data(int(i)))
            except Exception:
                pass

        if len(frames) == 0:
            raise ValueError("Can't extract framesы.")

        while len(frames) < num_frames:
            frames.append(frames[-1])

        return frames[:num_frames]
    finally:
        try:
            reader.close()
        except Exception:
            pass


def classify_video(path, video_model, video_processor):
    video_model.eval()

    num_frames       =       int(getattr(video_model.config,
"num_frames", 16))
    image_size       =       int(getattr(video_model.config,
"image_size", 224))

    frames = load_video_frames(path, num_frames=num_frames)

    inputs = video_processor(
        frames,
        return_tensors="pt",
        size={"height": image_size, "width": image_size},
    ).to(device)
```

```
    with torch.inference_mode():
        outputs = video_model(**inputs)
        probs    =    torch.softmax(outputs.logits,    dim=-
1)[0].detach().cpu().numpy()

    fake_idx,                    real_idx                    =
resolve_fake_real_indices(video_model)
    if fake_idx is None or real_idx is None:
        id2label  =  getattr(video_model.config,  "id2label",
{})
        scored = {str(id2label.get(i, f"label_{i}")).lower():
float(probs[i]) for i in range(len(probs))}
        fake            =            float(scored.get("fake",
scored.get("deepfake", 0.0)))
        real = float(scored.get("real", 1.0 - fake))
        return {"fake": fake, "real": real}

    return    {"fake":    float(probs[fake_idx]),    "real":
float(probs[real_idx])}

#   FUSION

def fuse(video_res, audio_res, w_video=0.7, w_audio=0.3):
    fake   =   w_video   *   video_res["fake"]   +   w_audio   *
audio_res["fake"]
    fake = float(np.clip(fake, 0.0, 1.0))
    return {"fake": fake, "real": 1.0 - fake}

def main():
    print("[INFO] Loading models...")

    audio_model                                                =
AutoModelForAudioClassification.from_pretrained(AUDIO_MODEL_
ID).to(device)
    audio_extractor                                            =
AutoFeatureExtractor.from_pretrained(AUDIO_MODEL_ID)
```

```
    video_processor                                          =
AutoImageProcessor.from_pretrained(VIDEO_MODEL_ID,
use_fast=False)
    video_model                                              =
AutoModelForVideoClassification.from_pretrained(VIDEO_MODEL_
ID).to(device)

    print("[INFO] Models loaded.\n")

    video_files = [
        os.path.join(VIDEO_DIR, f)
        for f in os.listdir(VIDEO_DIR)
        if f.lower().endswith((".mp4", ".mov", ".avi"))
    ]

    print(f"\n[INFO] Found {len(video_files)} test videos.")

    for vf in video_files:
        print("\n" + "=" * 60)
        print(f"[VIDEO] {os.path.basename(vf)}")
        try:
            # video branch
            video_res   =   classify_video(vf,   video_model,
video_processor)
            pred_v   =   "FAKE"   if   video_res["fake"]   >=
video_res["real"] else "REAL"
            print(f" Video → FAKE = {video_res['fake']:.4f}")
            print(f"        REAL = {video_res['real']:.4f}")

            # audio branch
            audio_path = extract_audio(vf)
            if     not     os.path.exists(audio_path)     or
os.path.getsize(audio_path) == 0:
                raise RuntimeError("Can't extract audio.")

            audio_res      =      classify_audio(audio_path,
audio_model, audio_extractor)
```

```
            pred_a = "FAKE" if audio_res["fake"] >=
audio_res["real"] else "REAL"
            print(f" Audio → FAKE = {audio_res['fake']:.4f}")
            print(f"        REAL = {audio_res['real']:.4f}")

            # fusion
            fused = fuse(video_res, audio_res, w_video=0.7,
w_audio=0.3)
            final_pred = "FAKE" if fused["fake"] >=
fused["real"] else "REAL"
            print(f" Result → FAKE = {fused['fake']:.4f}")
            print(f"          REAL = {fused['real']:.4f}")
            print(f"          PRED = {final_pred}")

        except Exception as e:
            print(f"[ERROR] {os.path.basename(vf)}: {e}")

    print("\n[DONE]")

if name == "__main__":
    main()
```