МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

ЛУЧКА Святослав Ігорович

Багатоступенева модель для виявлення діпфейкових відео на основі глибинного навчання / A Multistage Model for Deepfake Video Detection Based on Deep Learning

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КНм-11
Лучка С.І.
_____

Науковий керівник:
к.т.н., професор, Кочан В.В.
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
В.о. завідувача кафедри
_____ Н.М. Васильків

ТЕРНОПІЛЬ - 2024

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
ЛУЧЦІ Святославу Ігоровичу
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Багатоступенева модель для виявлення діпфейкових відео на основі глибинного навчання / A Multistage Model for Deepfake Video Detection Based on Deep Learning
керівник роботи к.т.н., професор, Кочан В.В.
затверджені наказом по університету від 12 грудня 2023 року № 753.

2. Строк подання студентом закінченої кваліфікаційної роботи 4 грудня 2024 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити
- Провести огляд проблеми класифікації емоцій та визначити основні виклики.
- Проаналізувати існуючі підходи до класифікації емоцій у різних контекстах.
- Розробити модель IG-CNN для розпізнавання емоцій із розділенням інтерпретованих та автономних репрезентацій.
- Впровадити обмеження інтерпретованості для підвищення ефективності навчання моделі.
- Використати групову CNN для забезпечення некорельованості репрезентацій у процесі класифікації.
- Дослідити методи розпізнавання емоцій на основі мовлення та інтегрувати їх у модель.
- Провести експериментальне тестування та оцінити ефективність моделі в різних ситуаціях і умовах.

5. Перелік графічного матеріалу у роботі
- Структура запропонованої моделі.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 12 грудня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2024 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2024 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 20.05.2024 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 28.10. 2024 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 11.11.2024 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 25.11.2024 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту.	до 30.11.2024 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 04.12.2024 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.12. 2024 р.	

Студент _____ Лучка С.І.
підпис

Керівник роботи _____ к.т.н., професор, Кочан В.В.
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему «Багатоступенева модель для виявлення діпфейкових відео на основі глибинного навчання» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 87 сторінки і містить 20 ілюстрацій, 11 таблиць, 1 додаток та 61 використаних джерела.

Метою даної кваліфікаційної роботи є розробка багатоступеневої моделі для виявлення діпфейкових відео з використанням вибору кадрів за золотим перетином та капсульних мереж для підвищення точності класифікації.

Методи досліджень: глибинне навчання, трансферне навчання, вибір ключових кадрів, капсульні нейронні мережі.

Результати дослідження: запропоновано модель для виявлення діпфейкових відео, яка забезпечує високу точність на наборах даних Celeb-DF і DFDC-P завдяки поєднанню капсульних мереж та попередньо натренованих моделей VGG19 і EfficientNet.

Результати роботи можуть застосовуватися для автоматизованого моніторингу відеоконтенту та боротьби з дезінформацією.

Ключові слова: ДІПФЕЙКИ, ГЛИБИННЕ НАВЧАННЯ, КАПСУЛЬНІ МЕРЕЖІ, ТРАНСФЕРНЕ НАВЧАННЯ.

ABSTRACT

The qualification thesis titled "A Multistage Model for Deepfake Video Detection Based on Deep Learning" for obtaining a Master's degree in the specialty 122 "Computer Science" of the educational program "Computer Science" comprises 87 pages and includes 20 illustrations, 11 tables, 1 appendices, and 61 references.

The purpose of this qualification work is to develop a multistage model for detecting deepfake videos using golden ratio frame selection and capsule networks to enhance classification accuracy.

Research methods: deep learning, transfer learning, key frame selection, capsule neural networks.

Research results: A model for detecting deepfake videos has been proposed, which ensures high accuracy on datasets such as Celeb-DF and DFDC-P through the combination of capsule networks and pre-trained models VGG19 and EfficientNet.

The results of the work can be applied for automated video content monitoring and combating disinformation.

Keywords: DEEPFAKES, DEEP LEARNING, CAPSULE NETWORKS, TRANSFER LEARNING.

ЗМІСТ

Вступ 7

1. Аналіз предметної області і постановка задачі дослідження 11

1.1 Аналіз сучасних технологій створення діпфейків та методів їх виявлення 11

1.2 Огляд існуючих підходів 19

1.3 Вибір перспективного шляху і постановка задачі дослідження 28

Висновки до 1 розділу 30

2. Багатоступенева модель для виявлення діпфейкових відео 31

2.1 Багатоступенева модель для виявлення діпфейкових відео 31

2.2 Попередня обробка 32

2.3 Вибір кадрів за зіницями 35

2.4 Вибір золотого кадру 37

2.5 Витягнення ознак 39

2.6 Класифікаційна мережа 41

2.7 Об'єднання оцінок класифікації 42

2.8 Критерії оцінювання 44

Висновки до 2 розділу. 46

3. Програмно-технологічне забезпечення 48

3.1 Аналіз еталонних наборів даних та експериментальні конфігурації для виявлення діпфейкових відео 48

3.2 Оцінка продуктивності 50

3.3 Порівняння з попередніми роботами 60

Висновки до 3 розділу. 65

Висновки 67

Список використаних джерел 69

Додаток А Апробація отриманих результатів 77

ВСТУП

Актуальність цієї роботи визначається стрімким зростанням використання технологій глибинного навчання для створення діпфейкових відео, що призводить до значних соціальних, політичних та економічних наслідків. Діпфейки дозволяють створювати реалістичні, але фейкові зображення та відео, які важко відрізнити від справжніх, що породжує ризики для приватного життя, кібербезпеки та суспільної довіри до цифрового контенту. Поява діпфейків уже викликала численні інциденти маніпуляцій громадською думкою, порушення конфіденційності та навіть шантажу, що підкреслює необхідність розвитку ефективних методів їх виявлення.

Крім того, завдяки поширенню програмного забезпечення для створення діпфейків, таких як FaceApp і ZAO, технологія стала доступною не лише для професіоналів, але й для широкого кола користувачів. Це підвищує ймовірність зловживання діпфейками в різних сферах, включаючи фальсифікацію політичних виступів, підробку документів та поширення дезінформації. Важливість дослідження полягає в тому, що на даному етапі розвитку технології вже неможливо покладатися виключно на людський фактор для ідентифікації підробок. Необхідні автоматизовані інструменти та алгоритми, які зможуть ефективно протистояти цим загрозам.

Сучасні методи виявлення діпфейків постійно вдосконалюються, проте вони мають певні обмеження. Більшість з них сфокусовані на окремих типах підробок або не забезпечують достатнього рівня точності для нових високоякісних діпфейкових відео. Тому розробка універсальних, масштабованих і ефективних методів, які здатні працювати з різними типами фейкових матеріалів, залишається актуальним завданням у сфері досліджень.

У цьому контексті запропонований підхід, заснований на поєднанні різних архітектур нейронних мереж та використанні передових методів трансферного навчання, дозволяє досягти значних покращень у точності та швидкості

виявлення діпфейків. Дослідження спрямоване на вирішення основних проблем сучасних моделей, таких як висока обчислювальна складність та невисока здатність до узагальнення на нові набори даних. Це робить роботу важливим внеском у розвиток технологій кібербезпеки та боротьби з дезінформацією.

Мета дослідження полягає в розробці багатоступеневої моделі для виявлення діпфейкових відео на основі методів глибинного навчання, що забезпечить високу точність і ефективність виявлення маніпульованих медіа.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Провести огляд проблеми класифікації емоцій та визначити основні виклики.
2. Проаналізувати існуючі підходи до класифікації емоцій у різних контекстах.
3. Розробити модель IG-CNN для розпізнавання емоцій із розділенням інтерпретованих та автономних репрезентацій.
4. Впровадити обмеження інтерпретованості для підвищення ефективності навчання моделі.
5. Використати групову CNN для забезпечення некорельованості репрезентацій у процесі класифікації.
6. Дослідити методи розпізнавання емоцій на основі мовлення та інтегрувати їх у модель.
7. Провести експериментальне тестування та оцінити ефективність моделі в різних ситуаціях і умовах.

Об’єктом дослідження є процес виявлення діпфейкових відео за допомогою алгоритмів штучного інтелекту та глибинного навчання.

Предметом дослідження є методи та моделі глибинного навчання, що використовуються для виявлення фейкових відео та маніпуляцій з обличчям у відеоматеріалах.

Методи дослідження включають використання математичного моделювання та експериментального аналізу з використанням алгоритмів глибинного навчання. Основними методами є побудова та навчання згорткових нейронних мереж (CNN), капсульних мереж (CapsNet) та інших сучасних архітектур глибинного навчання для класифікації відео. Застосовуються методи попередньої обробки даних, вибору ознак, а також експериментального тестування на еталонних наборах даних, таких як Celeb-DF та DFDC-P, для оцінки ефективності запропонованої моделі.

Наукова новизна отриманих результатів полягає в розробці багатоступеневої моделі для виявлення діпфейкових відео на основі глибинного навчання, яка перевершує існуючі рішення за точністю та ефективністю за рахунок використання капсульних мереж та вдосконаленого підходу до вибору ключових кадрів.

Практичне значення отриманих результатів полягає в можливості застосування розробленої багатоступеневої моделі для автоматизованого виявлення діпфейкових відео в реальних умовах, таких як соціальні мережі, платформи для обміну відео та медіа-контентом. Модель може бути інтегрована в системи моніторингу та безпеки для швидкого й ефективного виявлення підробленого контенту, що сприятиме зменшенню шкоди від його використання у шкідливих або неправомірних цілях, таких як політичні маніпуляції, кіберзлочини або порушення приватності.

Структура та обсяг роботи. Кваліфікаційна робота складається зі вступу, трьох основних розділів, висновків, списку літератури та додатків.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися: V Міжнародній науковій конференції «Стратегічні напрямки розвитку науки: фактори впливу та взаємодії», яка відбулася 11 жовтня 2024 року у місті Луцьк, Україна; VI Міжнародній мультидисциплінарній студентський науковій

конференції «Тренди та перспективи розвитку мультидисциплінарних досліджень», яка відбулася 11 жовтня 2024 року утмісті Кременчук, Україна.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Аналіз сучасних технологій створення діпфейків та методів їх виявлення

Термін "діпфейк" відноситься до фейкових матеріалів, створених методами глибинного навчання. Ці фейкові матеріали можуть бути використані в багатьох сферах. Розробка діпфейків — це трудомісткий і тривалий процес, що вимагає спеціалізованих інструментів і професійних знань. Зазвичай для створення таких матеріалів використовуються генеративні змагальні мережі (GAN) [1] і автоенкодери [2], що дозволяє створювати високоякісні та реалістичні матеріали.

Технологія діпфейків має багато корисних застосувань, таких як дубляж фільмів, відтворення акторів, що вже померли, у фільмах, відтворення історичних персонажів для ілюстрацій, демонстрація результатів у косметичній індустрії та деякі корисні застосування для онлайн-зустрічей.

Незважаючи на всі ці корисні застосування, діпфейкові відео та зображення стали серйозною загрозою, здатною завдавати шкоди, наприклад, маніпуляції політичною думкою або порушення приватного життя. У 2017 році користувач форуму Reddit з іменем "deepfakes" поділився відео, де обличчя деяких знаменитостей було замінено обличчями з порнографічних відео. Це відео мало значний вплив на спільноту, і після цього термін "діпфейк" став асоціюватися з відео. У 2019 році вийшов мобільний додаток під назвою "DeepNude", що робив людей на фотографіях роздягнутими. Незабаром з'явилися iOS додаток "ZAO" і Android додаток "FaceApp". За допомогою цих додатків користувачі можуть одним кліком змінювати завантажені зображення, такі як зміна обличчя, додавання чи видалення волосся і бороди, зміна статі та

віку. У 2021 році фейкові відео актора Тома Круза викликали стурбованість багатьох людей (Рисунок 1.1) [3].

Рисунок 1.1 - Оригінальний Майлз Фішер (ліворуч) та підроблена версія Тома Круза (праворуч)

З появою відкритого програмного забезпечення і різноманітних мобільних додатків виробництво фейкових зображень і відео досягло рівня, де майже кожен може легко створювати такі матеріали, незалежно від наявності технічних знань. Фейкові зображення та відео виглядають настільки реалістично, що людям важко зрозуміти, що це підробка.

Створення діпфейків для маніпуляцій обличчям включає складне використання штучного інтелекту (AI) і алгоритмів глибинного навчання для переконливого змінення або заміни рис обличчя на відео чи зображеннях. Цей процес зазвичай починається з збирання великих наборів даних, що містять вирази обличчя, різні кути та умови освітлення. Далі використовуються просунуті нейронні мережі, такі як GAN або автокодери, для навчання й імітації деталей людських облич та їхніх рухів. Рисунок 1.2 висвітлює принцип роботи автокодерів та GAN. Під час навчання модель отримує здатність генерувати дуже реалістичні, але часто оманливі маніпуляції обличчями.

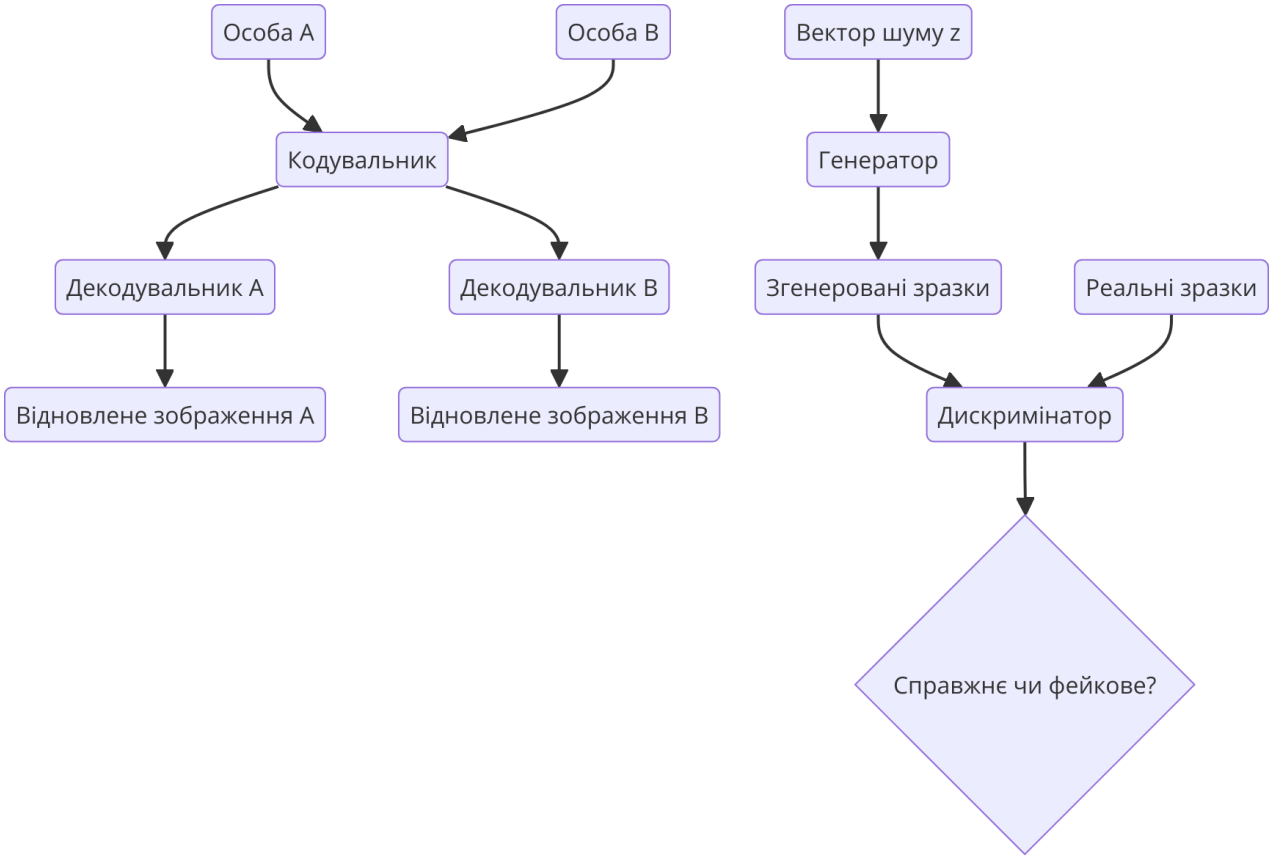


Рисунок 1.2 - (a) Принцип роботи автокодерів (b) Принцип роботи GAN [4]

Маніпуляції обличчями можна розділити на кілька категорій:

1. Повний синтез обличчя
2. Заміна особи
3. Перехід між обличчями
4. Маніпуляція атрибутами
5. Заміна виразів обличчя.

Повний синтез обличчя: у цьому методі створюється неіснуюче обличчя з нуля. Приклади зображень, створених цим методом, показані на рисунку 1.3. Зазвичай використовуються архітектури на базі GAN, такі як ProGAN і StyleGAN. Зображення, створені таким чином, настільки реалістичні та високоякісні, що людське око не може їх розпізнати як підробку. Однак ці

зображення можуть містити "відбиток GAN", схожий на "відбиток пристрою" на реальних зображеннях.

Рисунок 1.3 - Фейкові обличчя, взяті з сайту thispersondoesnotexist.com

Заміна особи: Ця маніпуляція полягає в заміні обличчя людини у відео обличчям іншої людини. Це одна з найпопулярніших сфер маніпуляції обличчями через недавні приклади, які викликали велике занепокоєння в суспільстві. Вона реалізується шляхом заміни зображення обличчя джерела обличчям цільової особи. Приклади цього методу наведені на рисунку 1.4. Метод реалізується двома різними підходами: на основі комп'ютерної графіки та на основі глибинного навчання. Використання цієї техніки є особливо корисним у кіноіндустрії, але вона також використовується зловмисно для створення фейкових відео.

Рисунок 1.4 - Приклади методу заміни ідентичності. Обличчя у вихідному відео змінюються відповідно до обличчя в цільовому відео

Перехід обличчя: створення нового штучного біометричного обличчя з використанням біометричних характеристик зображення обличчя двох або більше людей, як показано на рисунку 1.5. Цей метод становить загрозу для систем розпізнавання облич, оскільки нове фейкове обличчя містить риси облич, які використовувались при його створенні. На відміну від методу зміни ідентичності, замість створення фейкового контенту на рівні відео, намагаються створити фейковий зразок на рівні зображення.

Рисунок 1.5 - Нове фейкове обличчя, створене на основі біометричних даних особи 1 та особи 2

Маніпуляція атрибутами: техніка зміни таких характеристик, як колір волосся або шкіри, стать, вік, використання окулярів тощо.

Рисунок 1.6 показує приклади цього методу. Виробництво здійснюється за допомогою технік на базі GAN.

Використовуючи мобільний додаток FaceApp, фейковий контент можна створити одним натисканням. Ця техніка є корисною в косметичній індустрії, оскільки дозволяє побачити результат без використання продукту. Крім того, можливе зловмисне використання.

Рисунок 1.6 - Приклади методу маніпуляції атрибутами

Заміна виразу обличчя: техніка зміни виразу обличчя людини на відео або зображенні. Вирази обличчя людини у вихідному відео переносяться на вирази обличчя людини в цільовому відео, як показано на рисунку 1.7. Використовуючи техніки на основі GAN, фейковий контент створюється на рівні зображення, а за допомогою методів, таких як Face2Face та NeuralTextures, — на рівні відео. За допомогою цього методу можна зробити так, що людина здається, ніби говорить слово, яке насправді не вимовляла.

Рисунок 1.7 - Приклади методу заміни виразів. Вираз обличчя людини на вихідному відео змінюється відповідно до виразу особи на цільовому відео

У літературі існує багато методів для виявлення маніпуляцій обличчям у діпфейках. Як видно на рисунку 1.8, кількість досліджень щодо діпфейків постійно зростає. На рисунку 1.9, методи виявлення діпфейків поділені на п'ять підгруп: на основі відбитків камери, часової послідовності, візуальних артефактів, біологічних сигналів та загальних нейронних мереж.

Часова послідовність є унікальною ознакою для відео. Під час створення діпфейків можуть виникати невідповідності між кадрами. У методах, що базуються на часовій послідовності, ці невідповідності намагаються вловити. Також можуть виникати аномалії між створеним обличчям та фоном обличчя. Ці аномалії є основою методів, що базуються на візуальних артефактах.

У методах, що базуються на відбитках камери, використовуються унікальні "відбитки" камери, звані PRNU, для виявлення діпфейків. Біологічні сигнали, такі як моргання очей або частота серцевих скорочень, також є вирішальними ознаками для розрізнення реального та фейкового. Ця відмінність є основою методів на основі біологічних сигналів. Великий розвиток штучних нейронних мереж і їхній успіх у класифікаційних задачах також привів до їхнього використання у завданнях з виявлення діпфейків. У методах на основі загальних нейронних мереж використовуються різні штучні нейронні мережі для класифікації діпфейкових або реальних відео. Більше деталей про методи виявлення діпфейків буде наведено у розділі "Аналіз літератури".

Рисунок 1.8 - Цифри, взяті з webofscience.com, для статей, що містять слово deepfake у темі

Існує багато досліджень, розроблених на основі методів, згаданих вище, для виявлення діпфейкових відео. Ці дослідження мають певні проблеми. Ось ці проблеми та рішення для них:

— Проблема поколінь наборів даних: Набори даних для діпфейків поділяються на перше та друге покоління залежно від якості зразків, які вони містять. Набори даних першого покоління легше піддаються виявленню, ніж другого покоління. Більшість досліджень оцінюють тільки набори даних першого покоління. Розширені тести на наборах даних другого покоління, які є складнішими, краще продемонструють ефективність запропонованої моделі. Тому прагнемо досягти успішних результатів виявлення на двох наборах даних другого покоління, Celeb-DF та DFDC-P, які надають складні приклади для виявлення діпфейків.

— Проблема покращення тільки одного показника (ACC або AUC): Дослідження з виявлення діпфейків зазвичай спрямовані на покращення тільки ACC або AUC. Проте покращення обох показників одночасно є більш точним для оцінки ефективності досліджень. Тому дослідження повинні прагнути покращувати як ACC, так і AUC одночасно. Для цього використали архітектури капсульних мереж, такі як CapsuleNet і ArCapsNet, для виконання класифікаційної задачі. ArCapsNet замінює динамічну маршрутизацію та активацію squash у CapsuleNet на маршрутизацію уваги та активацію капсули. Використовуючи дві різні капсульні архітектури, з метою скористатися сильними сторонами обох. Для підвищення продуктивності нейронної мережі були використані попередньо натреновані мережі, такі як VGG19, EfficientNet-B0 та EfficientNet-B4 як екстрактори ознак. Також продуктивність було підвищено шляхом об'єднання моделей. Наш запропонований метод демонструє кращі результати порівняно з сучасними техніками на наборах даних Celeb-DF і DFDC-P, що підтверджено розширеними експериментами. Таким чином, ACC та AUC було покращено одночасно.

— Проблема високих обчислювальних витрат: Системи виявлення діпфейків розроблені для автоматичного визначення фейкового контенту в соціальних мережах, але існує проблема високих обчислювальних витрат і тривалого часу обробки поточних систем виявлення. Очевидно, що є потреба в дослідженнях, які зменшать вартість і час виявлення. Вибір конкретних кадрів є важливим, оскільки використання всіх кадрів у відео збільшить обчислювальні витрати. Незважаючи на використання технік рандомізації або вибору до певного кадру, більш доцільно розпізнавати та вибирати найкорисніші кадри. Для цього пропонуємо новий метод вибору кадрів на основі "золотого перетину" для обличчя. Завдяки цьому методу намагаємося вибрати найбільш підходящий кадр, який може містити докази підробки, щоб покращити продуктивність верифікації та скоротити час виявлення. Експериментальні дослідження показують, що запропонована модель приносить менші обчислювальні витрати порівняно з іншими дослідженнями.

1.2 Огляд існуючих підходів

Після того, як діпфейковий контент викликав значне суспільне занепокоєння, дослідження в цій галузі стали більш популярними. Було проведено багато досліджень у різних підгрупах. Класифікація методів виявлення діпфейків поділяється на 5 груп, як показано на Рисунку 1.9: на основі відбитків камери, на основі часової послідовності, на основі візуальних артефактів, на основі біологічних сигналів та на основі загальних мереж. У цьому розділі розглядаєм техніки виявлення діпфейків та деякі дослідження щодо кожної з цих технік.

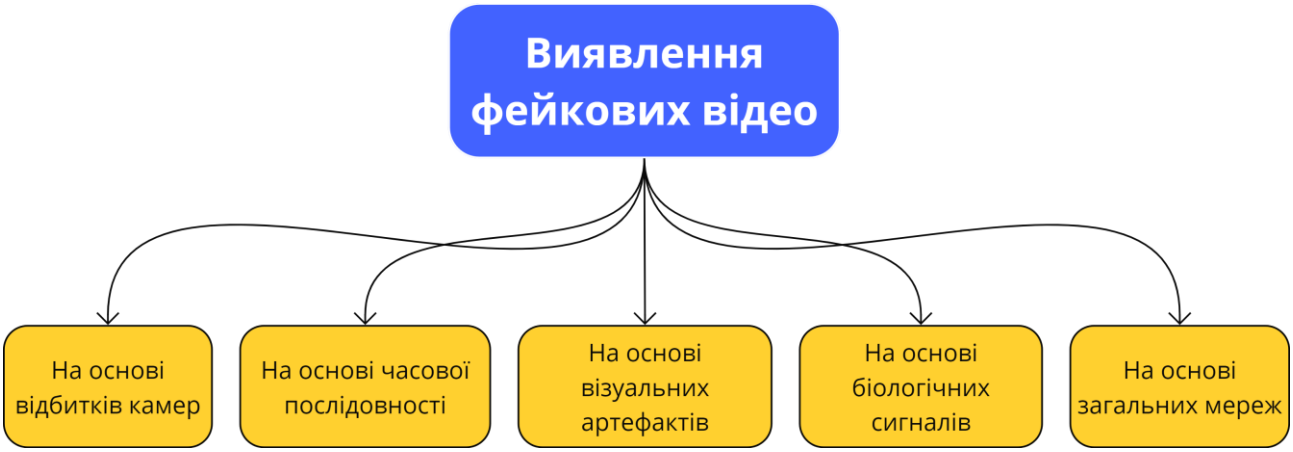


Рисунок 1.9 - Класифікація методів виявлення підробленого відео

1.2.1 Методи на основі відбитків камери

Деякі ранні дослідження аналізували відбитки камери, звані негомогенність реакції пікселів (PRNU), для виявлення діпфейків. PRNU виникає через неоднорідність кремнієвої пластини та різну чутливість пікселів до світла, що виникає в результаті дефектів виробничого процесу сенсора. Завдяки своїй унікальності та стабільності в кожному пристрої, модель PRNU вважається "відбитком пристрою", який може використовуватися для виконання багатьох криміналістичних завдань. Koopman та ін. [5] запропонували метод виявлення діпфейкових відео за допомогою PRNU. Вони розділили область обличчя кожного кадру відео на 8 груп і обчислили середнє значення PRNU для кожної групи. Експериментальні результати показали значну різницю в середньому нормалізованому коефіцієнті кореляції між реальним та діпфейковим відео. Lugstein та ін. [6] також використали PRNU і досягли гарних результатів у виявленні діпфейкових відео. Frank та ін. [7] вважають, що PRNU є ефективним для виявлення діпфейкових відео на невеликих наборах даних. Хоча на початку були отримані позитивні результати, методи на основі відбитків камери втратили свою ефективність через зразки, де можна було видалити відбиток.

1.2.2 Методи на основі часової послідовності

Відео мають спеціальну характеристику, яка називається часовою послідовністю. На відміну від зображень, відео — це серія візуальних кадрів із значною узгодженістю та послідовністю між ними. Коли кадри відео змінюються, послідовність між сусідніми кадрами втрачається, особливо коли виникають такі недоліки, як зміщення обличчя або стрибки відео. Враховуючи це, було розроблено багато стратегій виявлення.

З огляду на часову послідовність між кадрами зображення, Guera та ін. [8] рекомендують використовувати рекурентні нейронні мережі (RNN). Зображення облич у фейкових відео відтворюються покадрово. Автоенкодери (AE) не враховують попередні зображення під час відтворення, що може викликати аномалії між попередніми і наступними кадрами. Запропоновано кінцеву треновану систему виявлення діпфейкових відео для перевірки послідовності між сусідніми кадрами зображень. Ця система складається з конволюційної структури LSTM, яка обробляє послідовності кадрів зображень. Основні компоненти включають CNN як екстрактор ознак кадрів і LSTM для аналізу часових послідовностей. Попередньо треновану мережу InceptionV3 адаптували для обробки кожного кадру відео. Експерименти на створеній базі даних показали успішні результати.

Натхненні дослідженням Guera, Sabir та ін. [9] запропонували поєднувати попередню обробку обличчя та підходи часової мережі для виявлення фейкових відео. Хоча дослідження показало хороші результати для високоякісних відео, воно зазнало труднощів із стислими та низькоякісними відео. Montserrat та ін. [10] запропонували мережу CNN-RNN із системою автоматичного зважування, щоб підвищити стійкість алгоритму. Дослідження показали, що поєднання CNN та RNN дає високі результати на базі даних DFDC.

Інші дослідники, такі як Zhao та ін. [11], використовували оптичний потік для фіксації змін у виразах обличчя між сусідніми кадрами, але не змогли

досягти високої стійкості чи узагальнення. Wu та ін. [12] запропонували новий метод виявлення маніпуляцій, названий SSTNet, який використовує як низькорівневі ознаки, так і часові аномалії. Щоб підвищити узагальненість методу виявлення, Masi та ін. [13] запропонували двогалузеву мережу, де одна з гілок екстрагує початкові ознаки зображення, а інша пригнічує зміст обличчя.

Zhang та ін. [14] представили Video Transformer (VidTr), модель на основі трансформерів для розпізнавання дій у відео, яка безпосередньо застосовує увагу до сирих пікселів відео без використання згорток. Модель VidTr використовує розділену увагу для просторової та часової обробки, що знижує споживання пам'яті та забезпечує хорошу продуктивність з меншою кількістю обчислень у порівнянні з CNN. Zhao та ін. [15] запропонували метод ISTVT, який використовує розділену просторово-часову самоувагу в архітектурі трансформера, а також механізм само-віднімання для фокусування часової уваги на невідповідностях між кадрами.

1.2.3 Методи на основі візуальних артефактів

Дослідники намагаються виявити невідповідності між переднім планом і фоном обличчя, які можуть виникати під час створення фейкового обличчя. Деякі приклади таких невідповідностей наведено на Рисунку 1.10. Li та ін. [16] запропонували новий метод на основі глибинного навчання, заснований на спостереженнях невідповідностей між обличчями та фоном. У процесі навчання було використано чотири різні моделі CNN, зокрема VGG16, ResNet50, ResNet101 та ResNet152. Для баз даних першого покоління, UADFV та DeepfakeTIMIT, вони змогли досягти успішних результатів у порівнянні з попередніми дослідженнями.

Рисунок 1.10 - Невідповідності між новоствореною поверхнею інтер'єру та фоном

Li та ін. [17] пішли далі, запропонувавши нове дослідження. Вони представили нове представлення зображення, назване face X-ray, яке використовується для спостереження за тим, чи можна розбити вхідне зображення на переднє обличчя та фон. Було виявлено, що цей метод є більш ефективним для узагальнення. Проте, оскільки він надмірно зосереджений на межі між переднім планом і фоном, він не є успішним рішенням для ситуацій, які можуть заплутати цю межу.

Yang та ін. [18] представили підхід, що виявляє діпфейкові відео шляхом оцінки тривимірного положення голови на основі двовимірних точок обличчя. Відмінності між положеннями голови використовуються як вектор ознак для тренування SVM-класифікатора, який відрізняє реальні та фейкові відео. Однак продуктивність методу погіршується на відео з розмитою або низькою якістю зображення, оскільки складно точно визначити точки обличчя.

1.2.4 Методи на основі біологічних сигналів

Незважаючи на здатність GAN генерувати реалістичні зображення, вони поки що не здатні відтворювати біологічні сигнали. Біологічні сигнали, такі як моргання очей і частота серцевих скорочень, у фейкових зображеннях можуть відрізнятися від реальних, і завдяки цій різниці фейкові зображення можуть бути виявлені. Однак ця стратегія може стати неефективною, якщо GAN у майбутньому зможуть відтворювати біологічні сигнали.

Метод виявлення аномалій моргання очей був запропонований Jung та ін. [19], щоб визначати різницю між реальними та фейковими відео. Для виявлення моргань вони поєднали методи Fast-HyperFace [20] і EAR [21]. Спостерігаючи зміни в частоті моргань залежно від часу, статі, віку та інших змінних, використовується метод верифікації цілісності. Хоча в певних ситуаціях можна досягти успішних результатів, після вирішення проблеми частоти моргань цей метод вже не застосовується до сучасних задач виявлення діпфейків.

Ciftci та ін. [22] розробили підхід до виявлення фейкових облич, обчислюючи біологічні сигнали обличчя (такі як частота серцевих скорочень) за допомогою техніки rPPG. Витягнуті біологічні ознаки використовували для навчання моделей SVM і CNN. Результати показали високу точність виявлення, навіть для відео з низькою роздільною здатністю або низькою якістю.

Wu та ін. [23] запропонували двоетапну мережу для виявлення діпфейків за допомогою сигналів rPPG. Вони представили сигнали rPPG із відео обличчя за допомогою багатомасштабного просторово-часового подання, яке захоплює інформацію з кількох областей обличчя. Модуль маскованої локальної уваги виділяє змінені області на картах rPPG, а часовий трансформер взаємодіє з ознаками сусідніх карт rPPG для захоплення далеких залежностей. Експерименти на наборах даних FaceForensics++ і Celeb-DF показали, що запропонований метод перевершує інші методи на основі rPPG у виявленні фейкових відео.

DeepRhythm — це модель, представлена Qi та ін. [24], яка використовує ритм пульсу для ідентифікації діпфейкового відео. Вони розробили техніку, відому як MMSTR, для спостереження за сигналом серцевого ритму. Результати цього методу використовували для виявлення за допомогою мережі подвійної просторово-часової уваги. Аналогічно, DeepFakesOn-Phys [25] намагається виявляти фейкові відео, оцінюючи частоту серцевих скорочень через зміни кольору шкіри. Їхній метод досяг дуже успішних результатів.

1.2.5 Методи на основі загальних нейронних мереж

З огляду на успіх штучних нейронних мереж у задачах класифікації, ці методи також використовуються для виявлення діпфейкових відео, де необхідна класифікація на фейкове/реальне. У цьому методі зображення облич, отримані з відео, використовуються для навчання мережі виявлення. Навчена мережа застосовується до кадрів зображення відео, яке необхідно перевірити. Остаточне рішення щодо відео приймається шляхом усереднення результатів прогнозу, отриманих для кожного кадру. Методи на основі мереж є одними з перших методів, які використовувалися для виявлення діпфейкових відео. Після відкриття першого діпфейкового відео були запропоновані деякі методи, розроблені на основі існуючих мереж, які досягли успішних результатів у задачах класифікації зображень.

У своїй двопотоковій мережі Zhou та ін. [26] пропонують використовувати GoogleNet для дослідження структури облич і тернарну мережу для аналізу ознак стеганографії. Одна гілка використовує конволюційні нейронні мережі, тоді як інша екстрагує та класифікує фейкові ознаки. Рішення визначається шляхом поєднання оцінок класифікації обох гілок.

Wang та ін. [27] запропонували метод, заснований на моніторингу поведінки нейронів для виявлення фейкових облич. Rossler та ін. [28] використовували мережу Xception для виявлення на базі даних FF++. Схожі методи ідентифікації використовували Dolhansky та ін. [29] у дослідженні, що стало основою для оголошення бази даних DFDC. Було протестовано невелику нейронну мережу (DNN) з шести шарів згорток і повністю зв'язного шару, а також існуючу XceptionNet для створення базових показників. Початкові результати показали, що найкращий метод (XceptionNet) забезпечив точність 93%.

Nguyen та ін. [30] представили капсульну мережу для покращення продуктивності мереж виявлення. Спочатку зображення обличчя вводяться в частину попередньо натренованої мережі VGG-19. Витягнуті ознаки потім передаються в капсульну мережу, що включає кілька первинних капсул і дві вихідні капсули. Відповідність між ознаками, витягнутими первинними капсулами, динамічно обчислюється за допомогою алгоритму маршрутизації, і результати нарешті передаються до відповідної вихідної капсули.

Afchar та ін. [31] запропонували мережі Meso-4 і MesoInception-4. Виявлення діпфейків виконується за допомогою мезоскопічних ознак зображення. Метод був навчений і протестований на наборі даних Deepfake, і було отримано точність 98,4%.

Wodajo та ін. [32] запропонували метод, що поєднує CNN та Vision Transformer, і досягли конкурентоспроможних результатів на базі даних DFDC. Roy та ін. [33] використовували поєднання 3D ResNet та механізмів уваги. Cao та ін. [34] запропонували метод, заснований на навчанні реконструкції-класифікації, для покращення узагальнення та стійкості.

Wang та ін. [35] представили метод під назвою Representative Forgery Mining, який має два основні компоненти: карти уваги на підробки (Forgery Attention Maps) і підозріле стирання підробок (Suspicious Forgery Erasing). Експерименти показали, що RFM покращує продуктивність виявлення проти різних технік маніпуляції.

Zhang та ін. [36] запропонували 3D конволюційну нейронну мережу з часовим відсівом кадрів для виявлення діпфейкових відео. Вона використовує можливі невідповідності між кадрами відео, подаючи на вхід фіксовану кількість кадрів у 3DCNN і вводячи операцію випадкового відсіву кадрів.

Sun та ін. [37] запропонували метод під назвою Dual Contrastive Learning (DCL) для загального виявлення підроблених облич. DCL виконує контрастне навчання на різних рівнях для навчання дискримінаційних ознак. Експерименти

на кількох наборах даних показали, що метод дає хороші результати в узагальненні на нові домени.

Liu та ін. [38] запропонували новий метод під назвою FedForgery для узагальненого виявлення підроблених облич із використанням залишкового федеративного навчання. Він використовує варіаційний автоенкодер для навчання стійких залишкових карт ознак з реальних та фейкових облич для захоплення підроблених підказок.

Asha та ін. [39] представили візуальну мережу уваги, яка використовує VGG16 і оптичний потік для вилучення просторових і часових ознак з відео. Було створено власний мультимодальний набір даних, що поєднує реальні/фейкові зразки з існуючих наборів даних і нові діпфейки, створені за допомогою різних технік. Цей підхід досяг точності 97,6% на власному наборі даних.

Heidari та ін. [40] представили структуру під назвою Blockchain-Enabled Federated Deepfake Detection (BFLDL) для покращення виявлення фейкових відео з одночасним захистом конфіденційності даних. Щоб стандартизувати відео з різних джерел, застосовується процедура нормалізації даних. Потім використовуються моделі глибинного навчання для виявлення патернів, які вказують на діпфейки. Підхід федеративного навчання тренує глобальну модель із локальних оновлень моделей без обміну приватними даними.

Nguyen та ін. [41] представили метод під назвою LAA-Net для високоякісного виявлення діпфейків. LAA-Net пропонує явний механізм уваги в межах структури багатозадачного навчання для фокусування на вразливих точках, схильних до артефактів. Було введено також вдосконалену мережу піраміди ознак (E-FPN) для агрегування багатомасштабних ознак без надмірності, що дозволяє моделювати локалізовані ознаки. Експерименти на кількох наборах даних показали, що LAA-Net досягає 95,40% AUC на наборі даних Celeb-DF.

Kaddar та ін. [42] запропонували метод, що поєднує CNN із Vision Transformer і досягли конкурентоспроможних результатів на різних наборах даних.

1.3 Вибір перспективного шляху і постановка задачі дослідження

Актуальність теми дослідження діпфейків обумовлена швидким розвитком технологій штучного інтелекту, зокрема глибинного навчання, що створює нові виклики у сфері кібербезпеки та інформаційного простору. Діпфейки, які є результатом використання генеративних змагальних мереж (GAN) та автоенкодерів, дозволяють створювати відео- та зображувальні матеріали, що виглядають надзвичайно реалістично, але є фальсифікованими. Це створює небезпеку для суспільства, оскільки маніпулювання зображеннями публічних осіб може впливати на громадську думку, викликати соціальну дестабілізацію та поширювати дезінформацію.

Особливої актуальності набуває проблема виявлення діпфейкових матеріалів у політичних та соціальних процесах. Успішні атаки з використанням діпфейків можуть використовуватися для дискредитації лідерів, створення паніки або поширення фальшивих новин, що ставить під загрозу національну безпеку та стабільність суспільств. Наприклад, зростає кількість випадків використання діпфейків у передвиборчих кампаніях, що може кардинально змінити результати виборів та підірвати довіру до демократичних інститутів. Тому питання швидкого та точного виявлення діпфейків є критично важливим для забезпечення інформаційної безпеки.

Окрім політичного аспекту, діпфейки мають значний вплив на приватність та репутацію людей. Випадки створення фейкових відео за участю публічних осіб або звичайних громадян для компрометації та шантажу стали вже реальністю. Поява мобільних додатків, таких як FaceApp та ZAO, значно спростила процес створення фальшивих зображень та відео, роблячи їх

доступними для масового використання. Це зумовлює необхідність розробки ефективних систем виявлення діпфейків, які зможуть захистити громадян від незаконного використання їхньої ідентичності.

У науковій та промисловій сферах також відчувається необхідність у точних методах виявлення діпфейків. Відео- та фотоманіпуляції можуть використовуватися для створення фальшивих доказів, що впливають на судові процеси, або для підриву довіри до наукових даних. Крім того, зловмисне використання діпфейкових технологій може зашкодити репутації компаній, спровокувати корпоративні конфлікти або вплинути на фінансові ринки.

Отже, актуальність дослідження полягає в тому, що розробка нових ефективних методів для виявлення діпфейків є не лише науково-технічним викликом, а й має вагоме соціальне, політичне та економічне значення. Створення моделей на основі глибинного навчання, таких як CapsuleNet і ArCapsNet, здатних забезпечити високу точність і швидкість виявлення фальшивих відео, допоможе у розв’язанні цих актуальних проблем.

Мета дослідження полягає в розробці багатоступеневої моделі для виявлення діпфейкових відео на основі методів глибинного навчання, що забезпечить високу точність і ефективність виявлення маніпульованих медіа.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Провести огляд проблеми класифікації емоцій та визначити основні виклики.
2. Проаналізувати існуючі підходи до класифікації емоцій у різних контекстах.
3. Розробити модель IG-CNN для розпізнавання емоцій із розділенням інтерпретованих та автономних репрезентацій.
4. Впровадити обмеження інтерпретованості для підвищення ефективності навчання моделі.

5. Використати групову CNN для забезпечення некорельованості репрезентацій у процесі класифікації.

6. Дослідити методи розпізнавання емоцій на основі мовлення та інтегрувати їх у модель.

7. Провести експериментальне тестування та оцінити ефективність моделі в різних ситуаціях і умовах.

Висновки до розділу 1

1. Проведений аналіз сучасних технологій створення діпфейків та методів їх виявлення показав, що технології глибинного навчання, зокрема Генеративні Змагальні Мережі (GAN) та автоенкодери, стали основними інструментами для створення високоякісних та реалістичних фейкових медіа. Вони знаходять застосування в різних галузях, таких як кіноіндустрія, історичні реконструкції та онлайн-комунікація. Проте, з розвитком цих технологій виникає проблема їхнього зловживання у політичних та соціальних процесах, що вимагає нових підходів до виявлення фальсифікацій.

2. Аналіз існуючих методів виявлення діпфейків показав, що основні підходи поділяються на методи на основі відбитків камери, часової послідовності, візуальних артефактів, біологічних сигналів та загальних нейронних мереж. Кожен із цих методів має свої переваги та обмеження, але з огляду на швидкий розвиток технологій діпфейків, їх ефективність поступово знижується. Особливо це стосується методів, які базуються на обмежених характеристиках, таких як моргання або PRNU-відбитки.

3. У підсумку, для забезпечення точного та швидкого виявлення діпфейків необхідно використовувати комплексні підходи, що поєднують різні методи. Особливу увагу слід звертати на сучасні архітектури нейронних мереж, такі як CapsuleNet та ArCapsNet, які демонструють високу ефективність у класифікації фальшивих медіа.

2 БАГАТОСТУПЕНЕВА МОДЕЛЬ ДЛЯ ВИЯВЛЕННЯ ДІПФЕЙКОВИХ ВІДЕО

2.1 Багатоступенева модель для виявлення діпфейкових відео

Запропонована модель у цьому дослідженні включає шести ступінчастий процес, спрямований на виявлення діпфейкових відео. Блокова схема запропонованої моделі показана на рисунку 2.1, де виділено шість основних етапів.

1. Попередня обробка: На першому етапі здійснюється попередня обробка, де кадри вилучаються за допомогою виявлення та вирівнювання облич.

2. Вибір кадрів за зіницями: Далі, на етапі вибору кадрів за зіницями, обирається певна кількість кадрів з найпласкішим обличчям за допомогою співвідношення відстані від зіниць до межі обличчя на кадрах.

3. Вибір золотого кадру: На цьому етапі використовується інформація про золотий перетин обличчя для вибору найкращого кадру з кожного відео. Обирається кадр з найменшим значенням золотого перетину.

4. Витяг ознак: На наступному етапі витяг ознак виконується за допомогою попередньо натренованих моделей, таких як VGG19, EfficientNet-B0 та EfficientNet-B4, використовуючи метод трансферного навчання.

5. Класифікаційні мережі: Потім на етапі класифікації використовуються мережі CapsNet та ArCapsNet, і моделі навчаються з різними параметрами для отримання найкращих результатів.

6. Етап об'єднання моделей: Нарешті, на етапі об'єднання моделей найкращі моделі поєднуються, що призводить до значного покращення точності класифікації.

У подальших підрозділах кожен етап буде детально розглянуто.

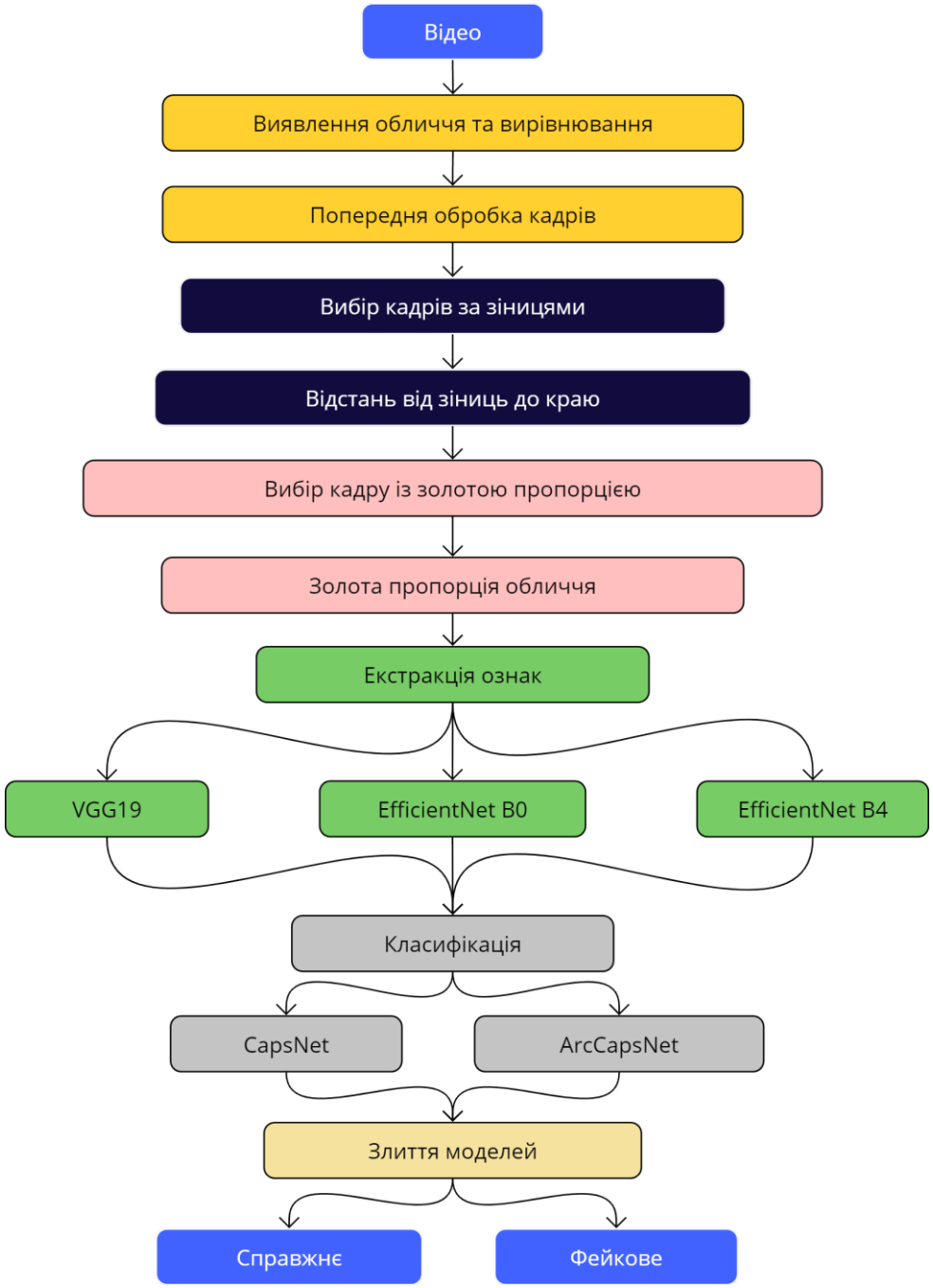


Рисунок 2.1 - Структура запропонованої моделі

2.2 Попередня обробка

Першим етапом запропонованого методу є попередня обробка, на якій виявляються та вирівнюються обличчя для вилучення кадрів. Faceswap [43] — це здебільшого програмне забезпечення для створення діпфейкових відео, яке має багато автоматизованих функцій, що полегшують процес створення таких

відео. Воно безкоштовне та відкрите для використання. Faceswap є одним із найбільш використовуваних програмних засобів для створення діпфейків і має інструменти для вилучення кадрів з відео шляхом виявлення та вирівнювання облич.

Faceswap пропонує інструменти MTCNN [44] та S3fd [45] для виявлення облич. Single Shot Scale-invariant Face Detector (S3Fd) — це сучасна глибока нейронна мережа, навчена для точного виявлення облич у різних умовах. S3fd сканує кожен кадр відео та визначає місце розташування та розмір кожного обличчя в кадрі. Після виявлення обличчя детектор створює рамку навколо нього, яка використовується для вилучення обличчя з кадру.

MTCNN (Multi-Task Cascaded Convolutional Neural Networks) — це високоефективний інструмент для виявлення облич, написаний на Python, що використовується для визначення положення зіниць і країв обличчя. Алгоритм MTCNN базується на каскадній архітектурі CNN, яка виконує три завдання послідовно: виявлення обличчя, локалізація контрольних точок обличчя та вирівнювання обличчя. На першому етапі MTCNN використовує мережу пропозицій для генерації набору рамок-кандидатів, які можуть містити обличчя. Потім на другому етапі рамки-кандидати уточнюються мережею уточнення, яка усуває помилкові позитивні результати та коригує координати рамки, щоб краще підігнати обличчя. І нарешті, на третьому етапі MTCNN використовує мережу контрольних точок для локалізації контрольних точок обличчя (таких як очі, ніс і рот) і вирівнювання обличчя в межах рамки.

Попри те, що S3fd здатний виявляти більше облич, перевага надається використанню MTCNN. Як показано на рисунку 2.2, S3fd виявляє інші обличчя на фоні, що призводить до фокусування на зразках, відмінних від основного обличчя. Тому було обрано MTCNN.

Рисунок 2.2 - Обличчя, розпізнані S3fd, позначені зеленою рамкою (ліворуч), обличчя, розпізнані Mtcnn, позначені синьою рамкою (праворуч)

Після виявлення обличчя наступним кроком є його правильне вирівнювання. Для цього використовується Face Alignment Network (Fan) [46], яка є детектором контрольних точок обличчя, що працює шляхом ідентифікації ключових точок на обличчі, таких як очі, ніс і рот. Fan використовує ці контрольні точки для правильного вирівнювання обличчя, забезпечуючи його центрування та правильну орієнтацію.

Кадри, вилучені за допомогою Faceswap, мають розмір 256 × 256 пікселів, що є поширеним розміром, який використовується в моделях глибинного навчання. Цей розмір забезпечує достатню деталізацію обличчя та швидкість обробки. Кроки етапу попередньої обробки показані на рисунку 2.3.

Рисунок 2.3 - Кроки етапу попередньої обробки. Спочатку обличчя визначається за допомогою Mtcnn і обрізається до розміру 256 × 256 пікселів. Після цього обличчя вирівнюється за допомогою вирівнювача FAN

2.3 Вибір кадрів за зіницями

Етап вибору кадрів за зіницями є другим етапом запропонованого підходу для виявлення діпфейкових відео. На цьому етапі здійснюється вибір кадрів, отриманих на попередньому етапі попередньої обробки, щоб отримати більш точні результати на наступних етапах. Метою цього етапу є вибір кадрів з найпласкішими обличчями для запобігання негативному впливу на результат наступного кроку.

Іноді в деяких вилучених кадрах може трапитися ситуація, коли обличчя повернене набік або не дивиться прямо в камеру, як показано на рисунку 2.4. Це може призвести до того, що частина обличчя не буде видимою, і це може негативно вплинути на результати наступних етапів. Тому вибір кадрів з найпласкішим обличчям є важливим для отримання точних результатів.

Рисунок 2.4 - Обличчя, що дивиться вбік (ліворуч), обличчя, що дивиться прямо (праворуч)

Для вибору кадрів з найпласкішим обличчям використовується співвідношення відстані між зіницями і краєм обличчя. Для визначення положення зіниць і країв обличчя використовується MTCNN. Приклад того, як MTCNN виявляє обличчя та області, можна побачити на рисунку 2.5.

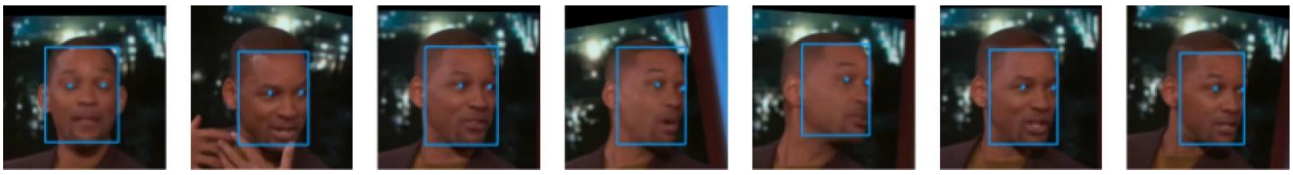


Рисунок 2.5 - Краї обличчя та зіниці за визначенням MTCNN

Основна мета цього етапу полягає у виборі кадрів з найпласкішими обличчями для отримання більш точних результатів на наступних етапах, тому використовується співвідношення відстані між зіницями і краєм обличчя для вибору кадрів з найпласкішим обличчям. Для цього здійснюються наступні кроки:

— Положення зіниць і країв обличчя визначаються за допомогою алгоритму MTCNN.

— Обчислюється піксельна відстань від лівої зіниці до лівого краю обличчя та від правої зіниці до правого краю обличчя.

— Визначається співвідношення меншого значення до більшого.

— Вибираються k кадрів із найвищим співвідношенням.

У математичному вигляді відстань від лівої зіниці до лівого краю обличчя та відстань від правої зіниці до правого краю обличчя можна позначити як dL_i і dR_i відповідно, де і представляє номер кадру. Співвідношення меншого значення до більшого можна обчислити за допомогою рівняння (2.1):

$$R_i = min\left(dL_i, dR_i\right) / max\left(dL_i, dR_i\right), \quad (2.1)$$

де *min()*- функція, яка повертає найменше значення;

max() - функція, яка повертає найбільше значення.

Кількість *k* кадрів з найбільшим співвідношенням можна вибрати як у рівнянні (2.2):

$$S = \{i_1, i_2, \ldots, i_k\} \, \text{that} R_{i,1} >= R_{i,2} >= \cdots >= R_{i,k} \quad , \qquad (2.2)$$

де S - набір вибраних кадрів;

i_1, i_2, …, i_k — це номери вибраних кадрів.

Загалом, етап вибору кадрів за зіницями використовує алгоритм MTCNN та математичні обчислення для вибору кадрів із найпласкішими обличчями, що забезпечує більш точні результати на наступних етапах.

2.4 Вибір золотого кадру

Етап вибору золотого кадру є одним із кроків запропонованого підходу для виявлення діпфейкових відео. На цьому етапі метою є вибір найбільш підходящого кадру з кожного відео за допомогою інформації про золотий перетин обличчя. Золотий перетин — це математичне співвідношення або пропорція, яка часто спостерігається в природі та мистецтві. Золотий перетин формується різними рисами обличчя, такими як висота та ширина обличчя. Фізичні розміри людського тіла, зокрема обличчя, часто демонструють золотий перетин.

Вибираючи кадр із найнижчим значенням золотого перетину, запропонований підхід намагається вибрати кадр, який відображає найбільш збалансовані та естетично приємні пропорції обличчя. Це має підвищити точність методу виявлення діпфейків. Значення золотого перетину обчислюється за допомогою програмного забезпечення з відкритим кодом GoldenFace [47]. Обчислення золотого перетину базується на принципах, показаних на рисунок 2.6.

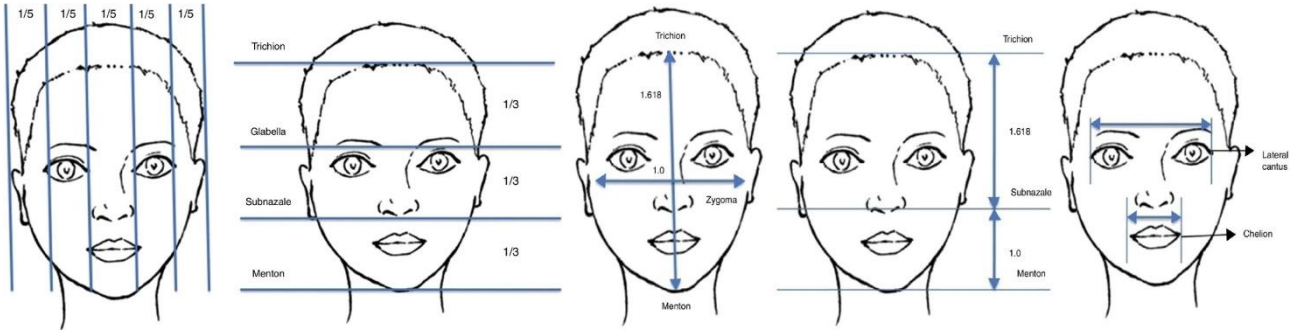


Рисунок 2.6 - Принципи золотого курсу на обличчі людини [48]

Для кожного відео значення золотого перетину обчислюється для k кадрів з попереднього етапу, де вже були вибрані найпласкіші обличчя. Кадр із найнижчим значенням золотого перетину обирається як "золоте обличчя". Деякі з вибраних кадрів для фейкових відео з бази даних Celeb-DF можна побачити на рисунку 2.7.

Важливість вибору найбільш підходящого кадру полягає в покращенні продуктивності моделей глибинного навчання, які використовуються на наступних етапах. Вибір прикладів, які використовуються для навчання глибинних мереж, є вирішальним для продуктивності моделі. Дослідження в літературі зазвичай вибирають кадри з відео випадково або з перших k кадрів. Однак використовуючи інформацію про золотий перетин обличчя, запропонована модель намагається вибрати найбільш підходящий кадр, що підвищить продуктивність моделі та скоротить загальний час навчання моделі.

Рисунок 2.7 - Вибрані кадри, які мають найнижчу золоту ціну обличчя у фейкових відео з бази даних Celeb-DF

2.5 Витягнення ознак

Після завершення попередньої обробки та вибору зображень на перших трьох етапах, зображення передаються на етап витягу ознак, який відіграє ключову роль у вилученні значущих представлень із необроблених даних зображень. Етап витягу ознак відповідає частині CNN перед основними капсулами в оригінальній архітектурі капсульної мережі [49].

На цьому етапі використовуються три потужні моделі згорткових нейронних мереж (CNN) як екстрактори ознак: VGG19, EfficientNet B0 і EfficientNet B4.

VGG19 — це впливова архітектура CNN, запропонована групою Visual Geometry Group (VGG) з Оксфордського університету. Назва "VGG19" походить від загальної кількості шарів у мережі, яка становить 19, включаючи згорткові, пулінгові та повністю зв'язані шари. Однією з ключових характеристик VGG19 є її простота та однорідність архітектури, що робить її легко зрозумілою та реалізованою. Мережа складається з повторюваних блоків 3 × 3 згорткових шарів, після яких ідуть 2 × 2 шари макс-пулінгу для прогресивного зменшення просторових розмірів ознак. Модель VGG19 використовується від першого шару до третього шару макс-пулінгу в процесі витягу ознак, як показано на рисунку 2.8 (a). Витягнуті ознаки з цієї мережі потім передаються на наступні етапи капсульної мережі.

EfficientNet — це сімейство архітектур CNN, представлених дослідниками з Google AI. Назва "Efficient" вказує на його унікальну характеристику — досягнення чудової точності при значно меншій кількості параметрів і обчислень у порівнянні з попередніми архітектурами. Ця ефективність досягається за допомогою техніки, яка називається складене масштабування (compound scaling), що балансує глибину моделі, ширину та роздільну здатність для досягнення оптимальної продуктивності.

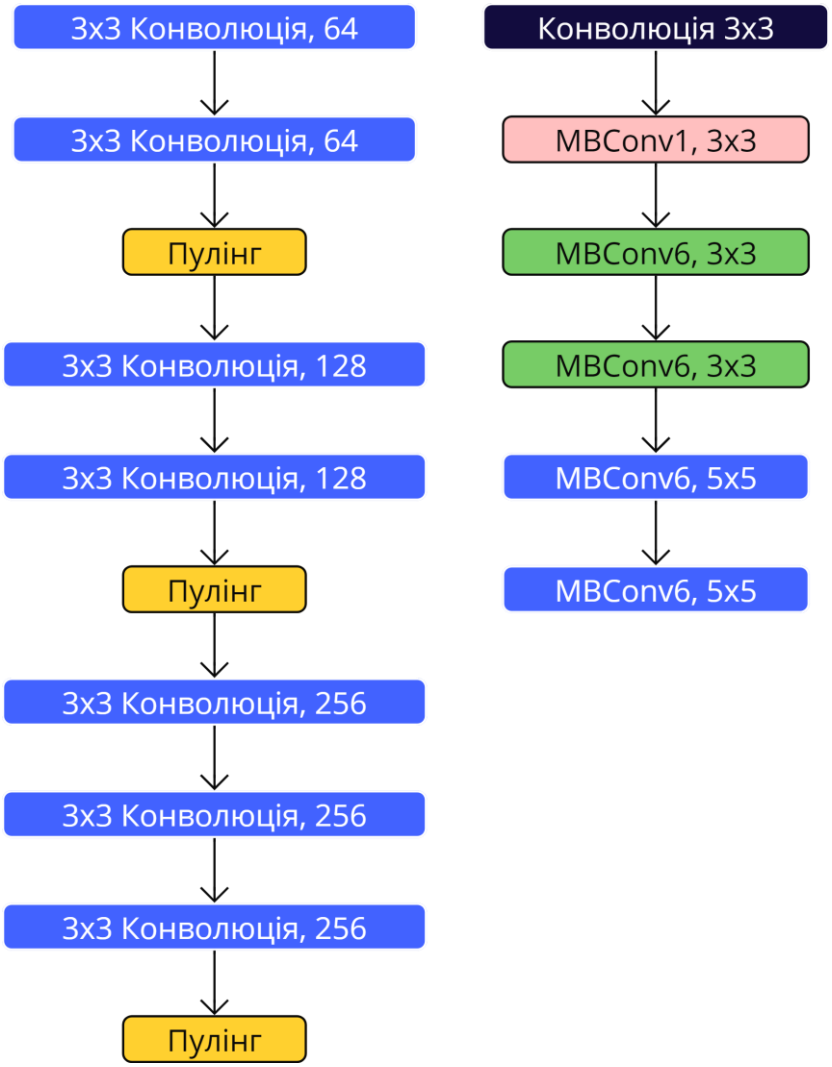


Рисунок 2.8 - (а) Використані шари VGG19 (б) Використані шари EfficientNet

EfficientNet B0 є базовою моделлю в сімействі EfficientNet. Він використовує метод складеного масштабування, що дозволяє рівномірно масштабувати глибину, ширину та роздільну здатність мережі в упорядкований спосіб. Це дозволяє знаходити оптимальний баланс між складністю моделі та продуктивністю, забезпечуючи ефективні та потужні можливості витягу ознак. EfficientNet B4, у свою чергу, належить до того ж сімейства, але представляє більшу та потужнішу модель порівняно з EfficientNet B0. Вона слідує тим самим принципам складеного масштабування, але з більшою глибиною, шириною та

роздільною здатністю, що дозволяє їй захоплювати більш складні закономірності та деталі у вхідних даних.

У процесі витягу ознак мережі EfficientNet B0 і B4 використовуються від першого блоку до четвертого блоку, як показано на рисунку 2.8 (b). Ці мережі демонструють чудову ефективність і надійність у вилученні ознак із зображень різної складності, що робить їх ідеальним вибором для наступних етапів капсульної мережі.

Завдяки використанню потужностей VGG19, EfficientNet B0 і EfficientNet B4 як екстракторів ознак, капсульна мережа може скористатися їх різноманітними можливостями представлення та узагальнення, що в кінцевому підсумку покращує продуктивність у різних завданнях комп'ютерного зору, таких як розпізнавання зображень, виявлення об’єктів і сегментація зображень. Усі мережі попередньо натреновані та використовуються як метод трансферного навчання. Використання попередньо навчених мереж як екстракторів ознак допомагає покращити точність класифікації та зменшити переобучення.

2.6 Класифікаційна мережа

На цьому етапі використовуються дві різні архітектури капсульних мереж. Ці мережі — Capsule Forensics [30] і ArCapsNet [50] (Attention Routing CapsuleNet), модифіковані від оригінальної архітектури капсульної мережі [49].

У 2011 році Hinton та ін. [51] представили першу капсульну мережу. Капсульні мережі спрямовані на подолання деяких обмежень CNN, особливо в області просторових відносин між об’єктами на зображенні. Капсульні мережі вводять новий тип одиниці, який називається капсулою. Капсулу можна уявити як групу нейронів, що кодують різні властивості, такі як наявність певного об’єкта та його позу (наприклад, положення, орієнтація, масштаб) на зображенні. Кожна капсула видає вектор, який називається активаційним вектором, що представляє різні властивості об’єкта, який вона представляє. Спочатку вони

стикалися з тими ж проблемами, що й CNN, включаючи обмеження продуктивності обладнання та нестачу ефективних методів. Ці проблеми було вирішено завдяки додаванню алгоритму динамічної маршрутизації [33]. За допомогою динамічної маршрутизації капсули в одному шарі можуть "спілкуватися" та взаємодіяти з капсулами на наступному рівні. Цей механізм допомагає капсулам вирішити, чи присутній об'єкт і які його характеристики. Мережа може збирати інформацію вищого рівня про об’єкти та їхні взаємодії, беручи до уваги узгодженість між капсулами.

Capsule Forensics — це адаптована версія оригінальної капсульної мережі. Вона містить кілька первинних капсул і дві вихідні капсули (реальна та фейкова). Кожна первинна капсула складається з 2D згорткової частини, шару статистичного пулінгу та 1D згорткової частини. Вихід первинних капсул динамічно маршрутизується до вихідних капсул. Остаточний результат обчислюється на основі активації вихідних капсул.

ArCapsNet — це інша версія капсульної мережі, модифікована від оригінальної. Вона замінює динамічну маршрутизацію та функцію активації squash з оригінальної капсульної мережі на маршрутизацію з увагою (attention routing) та активацію капсул. Вона має шар первинних капсул, шар згорткових капсул та повністю згортковий шар капсул. Крім того, мережа має активацію капсул, яка включає згортку 1×1 і активацію tanh, що застосовуються до виходу кожного шару.

У дослідженні використано архітектури капсульних мереж Capsule Forensics і ArCapsNet у тому ж вигляді, що і в оригінальних дослідженнях.

2.7 Об'єднання оцінок класифікації

На цьому важливому етапі результати ймовірності класифікації від різних моделей піддаються потужному процесу об'єднання. Об'єднання оцінок класифікації полягає в злитті кількох індивідуальних оцінок класифікації,

отриманих з різних моделей, в єдиний узгоджений результат. Цей підхід широко використовується в різних сферах, таких як машинне навчання, розпізнавання образів та аналіз даних, оскільки він дозволяє інтегрувати кілька джерел інформації, що підвищує точність і надійність прогнозів та ухвалення рішень.

Основна мета об'єднання оцінок класифікації полягає в тому, щоб використати сильні сторони та різноманітність, які надають різні класифікатори або джерела, синтезуючи їхні індивідуальні оцінки. Це стає особливо цінним при вирішенні складних завдань класифікації, де виграють від кількох точок зору та експертного досвіду в галузі. Злиття оцінок прагне досягти більш всебічного розуміння даних, що в кінцевому підсумку підвищує загальну продуктивність системи.

У цьому дослідженні стратегія об'єднання — це проста, але ефективна середня оцінка результатів моделей. Якщо у нас є N індивідуальних оцінок класифікації ($S_1, S_2, \ldots, S_n$) від N різних моделей, злиту оцінку F можна обчислити за допомогою середнього методу об’єднання:

$$F = \frac{1}{N} \sum_{i=1}^{N} S_i \,, \tag{2.3}$$

У цьому рівнянні додаємо всі індивідуальні оцінки та ділимо суму на кількість оцінок (N) для отримання середньої злитої оцінки F.

Важливо зазначити, що існують різні інші стратегії об'єднання, такі як зважена сума, зважене середнє, максимальне об'єднання, мінімальне об'єднання тощо, кожна зі своїми рівняннями для злиття оцінок. Вибір стратегії об'єднання залежить від конкретного застосування, характеристик моделей і бажаних властивостей злитої оцінки.

Цей метод сприяє об'єднанню прогнозів від кожної моделі в єдиний агрегований результат, що забезпечує надійний та збалансований результат.

Детальна інформація про результати об'єднання надається в розділі експериментальних результатів.

Загалом, об'єднання оцінок класифікації є ключовим кроком на шляху до підвищення точності та надійності прогнозів, що сприяє розвитку різних галузей, які значною мірою покладаються на ухвалення рішень, заснованих на даних.

2.8 Критерії оцінювання

В експерименті площа під кривою ROC (AUC) використовується як основний показник оцінки для всіх детекторів, оскільки це найчастіше використовувана міра в задачах бінарної класифікації. Точність (Accuracy) також є поширеним показником для оцінки продуктивності класифікаційної моделі, тому її використовували як додатковий критерій оцінки. Вона визначається як відношення правильно передбачених випадків до загальної кількості випадків у наборі даних і обчислюється за допомогою рівняння (2.4):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2.4}$$

— True Negative (TN): правильно визначене відсутність стану або події;

— False Negative (FN): неправильне невиявлення присутності стану або події;

— True Positive (TP): правильне виявлення присутності стану або події;

— False Positive (FP): неправильне визначення присутності стану або події, коли вона насправді відсутня.

Крива ROC (Receiver Operating Characteristic) не представляється за допомогою одного рівняння, але створюється шляхом побудови графіка залежності True Positive Rate від False Positive Rate на різних порогах класифікації. Однак загальну ідею можна виразити математично, як у рівнянні (2.5):

$$AUC = \int_0^1 TPR(FPR)\, dFPR\,, \tag{2.5}$$

де *AUC* - міра площі під кривою, утвореною побудовою графіка TPR проти FPR на різних порогах класифікації. Інтеграл обчислює площу під цією кривою, що надає єдине скалярне значення для оцінки продуктивності моделі бінарної класифікації.

TPR (True Positive Rate) і FPR (False Positive Rate) є показниками продуктивності, які часто використовуються в задачах бінарної класифікації, особливо в контексті машинного навчання та статистики. Вони обчислюються за допомогою рівнянь (2.6) і (2.7) відповідно:

$$TPR = Recall = \frac{TP}{TP+FN}\,, \tag{2.6}$$

$$FPR = \frac{FP}{FP+TN}\,, \tag{2.7}$$

Використовано такі показники продуктивності, як Precision, Recall та F1-score. Precision (точність) — це позитивна прогностична цінність, яка вимірює точність позитивних передбачень моделі, і її можна обчислити за допомогою рівняння (2.8). Recall (повнота), також відома як чутливість або True Positive Rate, вимірює здатність моделі виявляти всі позитивні випадки та обчислюється за допомогою рівняння (2.6). F1-score є гармонійним середнім між точністю і повнотою. Він забезпечує баланс між точністю та повнотою і обчислюється за допомогою рівняння:

$$Precision = \frac{TP}{TP+FP}, \tag{2.8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}\,, \tag{2.9}$$

У підсумку, основними критеріями оцінки продуктивності бінарних класифікаторів в експерименті виступають AUC (площа під кривою ROC) та Accuracy, які дають змогу об’єктивно порівнювати моделі на основі їхньої здатності правильно класифікувати випадки. Окрім цього, для глибшого розуміння роботи моделей розглядаються показники Precision, Recall (TPR) та F1-score, що відображають баланс між точністю позитивних прогнозів, здатністю моделі виявляти всі позитивні випадки та їхнім гармонійним поєднанням. Такий комплексний підхід до оцінювання продуктивності дозволяє отримати максимально повну картину ефективності застосованих методів класифікації.

Висновки до розділу 2

1. Запропонована багатоступенева модель для виявлення діпфейкових відео демонструє інноваційний підхід, що включає шість основних етапів, кожен з яких відіграє важливу роль у забезпеченні точності й ефективності класифікації. Ключовими етапами є попередня обробка кадрів, вибір кадрів за зіницями, використання золотого перетину для вибору найкращих кадрів, витяг ознак із зображень за допомогою попередньо натренованих моделей, а також класифікація та об'єднання результатів від кількох моделей. Це забезпечує цілісний процес, який покращує продуктивність системи на різних етапах обробки та класифікації відео.

2. Однією з головних переваг моделі є інтеграція кількох методів на кожному етапі, таких як використання MTCNN для виявлення облич, застосування золотого перетину для вибору найбільш підходящих кадрів, а також трансферне навчання з попередньо натренованими моделями VGG19, EfficientNet-B0 та EfficientNet-B4 для витягу ознак. Це дозволяє моделі досягати високої точності класифікації за допомогою поєднання різних підходів, що використовуються для аналізу зображень.

3. Модель передбачає об'єднання результатів від різних класифікаційних моделей (Capsule Forensics та ArCapsNet), що підвищує надійність і стабільність системи при виявленні діпфейкових відео. Об'єднання оцінок класифікації за допомогою методу середнього підходу дозволяє досягти високих результатів і уникнути впливу можливих похибок окремих моделей. Тестування на наборах даних, таких як Celeb-DF і DFDC-P, показало, що запропонована модель значно покращує показники точності й AUC у порівнянні з іншими сучасними методами.

3 ПРОГРАМНО-ТЕХНОЛОГІЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Аналіз еталонних наборів даних та експериментальні конфігурації для виявлення діпфейкових відео

Почнемо з представлення двох відомих еталонних наборів даних, надаючи детальну інформацію про наші експериментальні конфігурації. Потім проводимо експерименти з абляції, щоб ретельно оцінити ефективність нашого модуля попередньої обробки. Надалі ретельно перевіряємо наш інноваційний підхід на наборах даних DFDC-Preview та Celeb-DF. Крім того, представляємо порівняльні експерименти з найсучаснішими (SOTA) методами, щоб продемонструвати перевагу нашого методу. Далі оцінюємо злиття найкращих моделей.

Використання еталонних наборів даних відіграє ключову роль у оцінюванні продуктивності та ефективності різних алгоритмів і моделей. У цьому дослідженні використано два важливі та складні набори даних у цій галузі: Celeb-DF та DFDC-P.

Набір даних Celeb-DF (Celebrity Deepfake Dataset) — це комплексна колекція діпфейкових відео, що включає відомих знаменитостей, представлена Li та ін. [52] у 2020 році. Він містить різноманітні загальнодоступні відео з YouTube і є цінним ресурсом для дослідників і практиків. Цей набір даних не лише багатий на контент, але й охоплює широкий спектр складних сценаріїв, таких як різні умови освітлення, пози та вирази обличчя. Використання знаменитостей у цьому наборі додає складності через варіації в зовнішності та ідентичності. Celeb-DF, таким чином, пропонує реалістичне уявлення про діпфейкові відео, що робить його хорошим вибором для оцінки стійкості методів виявлення діпфейків. Celeb-DF було створено на основі реальних відео з YouTube від 59 різних знаменитостей і складається з 890 реальних відео та 5639 високоякісних діпфейкових відео.

Набір даних DFDC-P (DeepFake Detection Challenge Preview) [29] — це ще один важливий набір даних, який було включено в це дослідження. Він був спеціально створений для конкурсу DeepFake Detection Challenge, організованого Facebook у співпраці з кількома партнерами. DFDC-P містить високоякісні діпфейкові відео, створені професійними акторами, що забезпечує контрольований і стандартизований набір даних для суворої оцінки. Набір даних охоплює різноманітні рухи обличчя, емоції та етнічні групи, що представляє комплексний набір викликів для алгоритмів виявлення діпфейків. DFDC-P є цінним для оцінки продуктивності методів виявлення та слугує еталоном для оцінки прогресу в цій галузі. Він складається з 1131 реального відео та 4113 діпфейкових відео.

Розподіл даних для навчання, валідації та тестування для обох наборів даних Celeb-DF та DFDC-P детально описаний у Таблиці 1.

Для набору даних Celeb-DF використовували ту саму стратегію, що й у [53]: відео перших 13 знаменитостей із набору використовували для валідації, а решта 46 знаменитостей — для навчання. Загалом 610 реальних і 4299 фейкових відео було використано для навчання, 102 реальних і 1000 фейкових — для валідації та 178 реальних і 340 фейкових відео — для тестування, як показано у Таблиці 1. Набір для тестування в цьому наборі даних був попередньо визначений відповідним автором і використовувався у цьому дослідженні без змін. Відео з тестового набору не використовувалися на етапах навчання або валідації.

У наборі даних DFDC-P автор попередньо визначив набір для тестування, який також використовувався у цьому дослідженні. Після того, як відео з тестового набору були відокремлені, решту відео було випадково розподілено на навчальні та валідаційні набори в пропорції 80% та 20% відповідно. Загалом 685 реальних і 2899 фейкових відео було призначено для навчання, 170 реальних і 710 фейкових — для валідації та 276 реальних і 504 фейкових відео — для тестування, як показано у таблиці 3.1.

Таблиця 3.1 - Розподіл даних серед навчання, валідації та тестування

Набір даних	Набір	Навчання	Валідація	Тестування
Celeb-DF	Реальні	610	102	178
	Фейкові	4299	1000	340
DFDC-P	Реальні	685	170	276
	Фейкові	2899	710	504

3.2 Оцінка продуктивності

У цьому дослідженні використовували ті самі параметри (швидкість навчання, dropout, оптимізатор Adam), що й автори в архітектурах CapsNet [30] та ArCapsNet [50]. У моделях CapsNet швидкість навчання була встановлена на рівні 5×10^{-4}, щоб забезпечити баланс між швидкістю збіжності та точністю, що дозволяє моделям ефективно адаптуватися до патернів даних. Використано оптимізатор Adam з $\beta_1 = 0.9$, що сприяє ефективній оптимізації, поєднуючи переваги як імпульсу, так і адаптивних швидкостей навчання. Крім того, для запобігання переобученню було введено коефіцієнт dropout 5%, що сприяє стійкій генералізації шляхом випадкової деактивації невеликого відсотка одиниць нейронної мережі під час навчання. У моделях ArCapsNet швидкість навчання була встановлена на рівні 1×10^{-3}, а оптимізатор Adam використовувався з $\beta_1 = 0.9$. Для всіх моделей було обрано розмір пакету 16, а всі зображення мали розмір 256 × 256 пікселів. Всі моделі тренувалися протягом 100 епох.

Цей розділ включає кілька підрозділів:

— загальна оцінка — результати дослідження;

— оцінка об’єднання різних моделей — оцінка продуктивності об’єднання;

— порівняння з попередніми роботами — порівняння результатів з іншими дослідженнями в літературі;

— аналіз складності моделей — оцінка складності моделей;

— статистичний аналіз — застосування статистичних тестів.

3.2.1 Загальна оцінка

У цьому розділі представлено експериментальні результати, отримані в нашому комплексному дослідженні. Досліджено результати моделей CapsNet і ArCapsNet та порівнюємо їх ефективність за допомогою різних моделей для витягу ознак, зокрема VGG19 [54], EfficientNet B0 [55] та EfficientNet B4 [55], які детально представлені в таблицях 3.2 та 3.3.

Таблиця 3.2 - Продуктивність моделей на наборі даних Celeb-DF

Модель	Екстрактор ознак	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsuleNet	VGG19	94.49	90.93	99.25	74.16	84.89
	EfficientNet B0	97.93	89.58	96.97	71.91	82.58
	EfficientNet B4	97.58	92.66	98.61	79.78	88.20
CapsuleNet (Loss Weight)	VGG19	97.54	92.28	97.92	79.21	87.58
	EfficientNet B0	96.18	91.12	95.21	78.09	85.80
	EfficientNet B4	98.61	93.05	97.97	81.46	88.96
ArCapsNet (Layer1)	VGG19	89.88	87.26	96.67	65.17	77.85
	EfficientNet B0	92.89	84.94	93.86	60.11	73.29
	EfficientNet B4	91.40	88.80	96.88	69.66	81.05
ArCapsNet (Layer2)	VGG19	93.38	88.80	95.45	70.79	81.29
	EfficientNet B0	94.57	85.91	89.47	66.85	76.53
	EfficientNet B4	96.41	90.15	97.04	73.60	83.71
ArCapsNet (Layer3)	VGG19	94.02	89.38	94.89	73.03	82.54
	EfficientNet B0	88.77	83.40	97.92	52.81	68.61
	EfficientNet B4	90.86	85.71	96.43	60.67	74.48

Таблиці 3.2 та 3.3 надають повний аналіз результатів, отриманих шляхом використання цих моделей у контексті витягу ознак і застосування капсульних мереж. Зокрема, під час обчислення функції втрат у капсульній мережі особливу увагу приділяли вирішенню проблеми дисбалансу класів, і ці вагові коригування явно вказані поруч із назвами моделей.

Особливої уваги заслуговує архітектура ArCapsNet, яка має шарувату структуру, що складається з блоків капсул. Різні результати були досягнуті шляхом зміни кількості шарів (від 1 до 3). Це різноманіття архітектурних конфігурацій надає більш тонкий погляд на продуктивність моделі.

Таблиця 3.3 - Продуктивність моделей на наборі даних DFDC-P

Модель	Екстрактор ознак	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsuleNet	VGG19	79.68	73.68	60.47	75.36	67.10
	EfficientNet B0	81.57	72.52	58.68	77.17	66.67
	EfficientNet B4	81.89	73.81	60.52	76.09	67.42
CapsuleNet (Loss Weight)	VGG19	82.92	74.06	61.47	72.83	66.67
	EfficientNet B0	79.86	69.81	55.28	79.71	65.28
	EfficientNet B4	80.55	72.13	58.47	75.00	65.71
ArCapsNet (Layer1)	VGG19	82.63	76.39	66.43	68.12	67.26
	EfficientNet B0	82.57	75.48	64.05	71.01	67.35
	EfficientNet B4	82.94	76.90	65.40	74.64	69.71
ArCapsNet (Layer2)	VGG19	84.12	77.16	67.62	68.84	68.22
	EfficientNet B0	82.57	75.61	63.55	73.91	68.34
	EfficientNet B4	84.03	77.42	67.60	70.29	68.92
ArCapsNet (Layer3)	VGG19	85.28	77.55	68.75	67.75	68.25
	EfficientNet B0	81.64	75.74	64.86	69.57	67.13
	EfficientNet B4	87.56	80.52	71.93	74.28	73.08

У таблиці 3.2 наведено продуктивність різних моделей на наборі даних Celeb-DF, що демонструє їхню ефективність з точки зору AUC та Точності (ACC). Моделі, що розглядаються, включають CapsuleNet, CapsuleNet з коригуваннями ваг втрат, та ArCapsNet з різною кількістю шарів.

3.2.2 Аналіз продуктивності моделей на наборах даних Celeb-DF та DFDC-P

CapsuleNet, використовуючи VGG19 як екстрактор ознак, досягає AUC на рівні 94.49 та ACC 90.93. Використання EfficientNet B0 та B4 як екстракторів ознак демонструє покращену продуктивність, з AUC на рівні 97.93 (B0) і 97.58 (B4). Зокрема, EfficientNet B4 перевершує B0 та VGG19 за точністю (92.66).

Введення коригування ваг втрат у CapsuleNet значно покращує продуктивність. У поєднанні з VGG19 AUC та ACC досягають 97.54 і 92.28 відповідно. З EfficientNet B0 AUC становить 96.18, ACC — 91.12, а з EfficientNet B4 — 98.61 і 93.05 відповідно, що демонструє найвищу ефективність.

ArCapsNet, з його багатошаровою архітектурою, демонструє різну продуктивність залежно від кількості шарів. У Layer 1 значення AUC варіюються від 89.88 (VGG19) до 92.89 (EfficientNet B4), тоді як ACC коливається від 84.94 (EfficientNet B0) до 88.80 (EfficientNet B4). У Layer 2 продуктивність покращується, причому AUC досягає 96.41 (EfficientNet B4), а ACC — 90.15. У Layer 3 продуктивність залишається на конкурентному рівні, з AUC від 88.77 (EfficientNet B0) до 94.02 (VGG19), а ACC коливається від 83.40 (EfficientNet B0) до 89.38 (VGG19).

Аналіз показує, що модель CapsuleNet (Loss Weight) у поєднанні з EfficientNet B4 як екстрактором ознак є найбільш ефективною конфігурацією, досягаючи AUC 98.61 і ACC 93.05 на наборі даних Celeb-DF. Ця модель поєднує сильні сторони архітектури CapsuleNet з потужними можливостями витягу ознак EfficientNet B4, демонструючи її високу ефективність у виявленні діпфейків.

У таблиці 3.3 представлена продуктивність різних моделей на наборі даних DFDC-P, підкреслюючи їх ефективність з точки зору AUC і ACC. Подібно до аналізу, проведеного для таблиці 3.2, розглянемо кожну модель окремо.

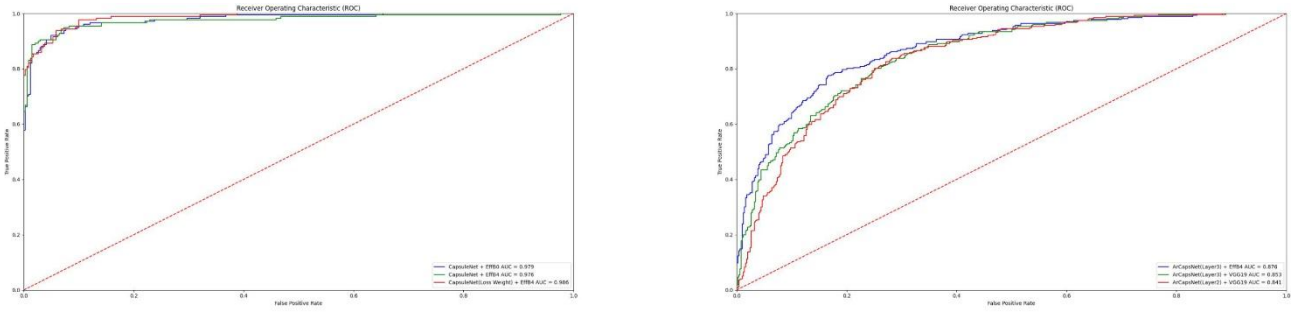
CapsuleNet, використовуючи VGG19 як екстрактор ознак, показує AUC 79.68 і ACC 73.68. Використання EfficientNet B0 та B4 підвищує AUC до 81.57 та 81.89, а ACC до 72.52 і 73.81 відповідно. Коригування ваг втрат у CapsuleNet дає покращені результати: VGG19 досягає AUC 82.92 і ACC 74.06, а EfficientNet B4 — AUC 80.55 і ACC 72.13.

ArCapsNet з багатошаровою архітектурою демонструє тонкі відмінності в продуктивності в залежності від шару. У Layer 1 AUC коливається від 82.57 (EfficientNet B0) до 82.94 (EfficientNet B4), а ACC — від 75.48 (EfficientNet B0) до 76.90 (EfficientNet B4). У Layer 2 продуктивність покращується: AUC становить 84.12 (VGG19), а ACC — 77.42 (EfficientNet B4). У Layer 3 досягається найвища продуктивність: AUC від 81.64 (EfficientNet B0) до 87.56 (EfficientNet B4), а ACC від 75.74 (EfficientNet B0) до 80.52 (EfficientNet B4).

Модель CapsuleNet (Loss Weight) з VGG19 як екстрактором ознак досягає AUC 82.92 і ACC 74.06 на наборі даних DFDC-P. Ця модель демонструє високу ефективність у розпізнаванні діпфейкового контенту в межах набору даних DFDC-P.

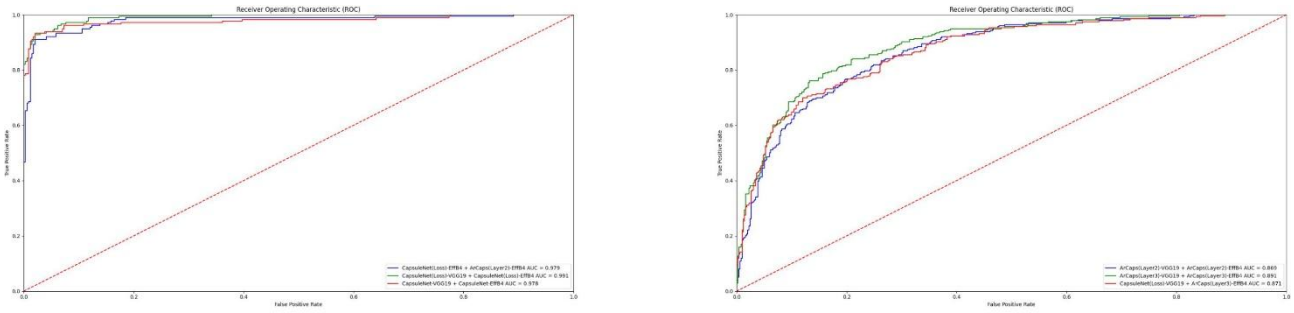
Порівняльний аналіз серед екстракторів ознак показує високу ефективність EfficientNet B4, яка послідовно забезпечує відмінні результати. Введення коригування ваг втрат помітно покращує продуктивність CapsuleNet. Багатошарова архітектура ArCapsNet пропонує гнучкість, демонструючи різну продуктивність залежно від кількості шарів.

Рисунки 3.1 та 3.2 представляють криві ROC для трьох найкращих результатів, отриманих із таблиць 3.2 та 3.3 відповідно.



(a) (b)

Рисунок. 3.1 - (а) РПЦ на Celeb-DF (б) РПЦ на DFDC-P



(a) (b)

Рисунок. 3.2 - (а) ROC моделей синтезу на Celeb-DF (b) ROC моделей синтезу на DFDC-P

3.2.3 Оцінка об'єднання різних моделей

Щоб додатково перевірити ефективність об'єднання, вводимо експериментальні бенчмарки для наборів даних Celeb-DF і DFDC-P. Як видно з таблиць 3.4 та 3.5, об'єднання прогнозованих оцінок добре працюючих моделей покращило результати.

Таблиця 3.4 - Результати об'єднання найкращих моделей для набору даних Celeb-DF (Результати класифікації Моделі 1 та Моделі 2 об'єднані)

Модель 1	Модель 2	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsuleNet (Loss) + VGG19	CapsuleNet (Loss) + EffB4	99.14	93.63	99.32	82.02	89.85
CapsuleNet + VGG19	CapsuleNet + EffB4	97.83	91.51	100.00	75.28	85.90
ArCaps (Layer1) + VGG19	ArCaps (Layer1) + EffB4	94.33	87.64	99.14	64.61	78.23
ArCaps (Layer2) + VGG19	ArCaps (Layer2) + EffB4	97.79	89.96	97.73	72.47	83.23
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB4	96.33	88.80	96.88	69.66	81.05
CapsuleNet (Loss) + EffB4	ArCaps (Layer1) + EffB4	95.55	89.38	97.67	70.79	82.08
CapsuleNet (Loss) + EffB4	ArCaps (Layer2) + EffB4	97.88	91.51	97.18	77.53	86.25
CapsuleNet (Loss) + EffB4	ArCaps (Layer3) + VGG19	97.44	89.38	94.89	73.03	82.54

У таблиці 3.4 представлені результати об'єднання найкращих моделей на наборі даних Celeb-DF, що демонструє вплив поєднання різних архітектур для підвищення продуктивності класифікації. Розглянуті моделі включають CapsNet (Loss), CapsNet та ArCapsNet з різними шарами, кожна з яких була об'єднана з VGG19 або EfficientNet B4 як екстракторами ознак. Операції об'єднання виконувались для кожної моделі з двома екстракторами ознак, які дали найкращі результати. Крім того, модель CapsNet, яка показала найкращі результати, була об'єднана з моделями ArCaps, які також показали найкращі результати.

Таблиця 3.5 - Результати об'єднання найкращих моделей для набору даних DFDC-P (Результати класифікації Моделі 1 та Моделі 2 об'єднані)

Модель 1	Модель 2	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsuleNet + VGG19	CapsuleNet + EffB4	83.50	73.29	59.89	75.72	66.88
CapsuleNet (Loss) + VGG19	CapsuleNet (Loss) + EffB4	83.40	75.35	62.54	76.81	68.94
ArCaps (Layer1) + VGG19	ArCaps (Layer1) + EffB4	86.58	79.87	69.87	76.45	73.01
ArCaps (Layer2) + VGG19	ArCaps (Layer2) + EffB4	86.86	79.10	70.21	71.74	70.97
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB4	89.08	82.84	76.00	75.72	75.86
CapsuleNet (Loss) + VGG19	ArCaps (Layer1) + EffB4	83.97	76.65	64.62	76.09	69.88
CapsuleNet (Loss) + VGG19	ArCaps (Layer2) + EffB4	84.46	77.29	66.45	73.19	69.66
CapsuleNet (Loss) + VGG19	ArCaps (Layer3) + EffB4	87.14	78.84	68.67	74.64	71.53

Модель CapsNet з коригуванням втрат у поєднанні з VGG19 та EfficientNet B4 як екстракторами ознак демонструє виняткову продуктивність при об'єднанні між собою. Об'єднання CapsNet (Loss) + VGG19 і CapsNet (Loss) + EffB4 забезпечує вражаючий AUC 99.14 і ACC 93.63. Об'єднання CapsNet + VGG19 і CapsNet + EffB4 також показує сильні результати, з AUC 97.83 та ACC 91.51.

ArCapsNet, зі своєю багатошаровою структурою, демонструє різні результати при об'єднанні. Об'єднання ArCaps (Layer1) + VGG19 і ArCaps (Layer1) + EffB4 дає AUC 94.33 і ACC 87.64, що демонструє конкурентоспроможну продуктивність. Подібні тенденції спостерігаються при об'єднанні ArCaps (Layer2) + VGG19 та ArCaps (Layer2) + EffB4, де AUC досягає 97.79, а ACC — 89.96. Об'єднання ArCaps (Layer3) + VGG19 та ArCaps (Layer3) + EffB4 дає AUC 96.33 і ACC 88.80.

Важливо відзначити, що об'єднання CapsNet (Loss) + VGG19 та CapsNet (Loss) + EffB4 досягає вражаючих показників AUC 99.14 та ACC 93.63, що підтверджує ефективність поєднання моделей для покращення продуктивності. ACC покращився на 0.58%, а AUC — на 0.53%.

У таблиці 3.5 наведені результати об'єднання найкращих моделей на наборі даних DFDC-P, що ілюструє результати поєднання різних архітектур виявлення діпфейків з використанням VGG19 або EfficientNet B4 як екстракторів ознак. Розглянуті моделі включають CapsNet, CapsNet (Loss) та ArCapsNet з різними шарами. Для кожної моделі операції об'єднання проводилися з використанням двох екстракторів ознак, які показали найкращі результати. Крім того, моделі ArCaps та CapsNet, які показали найкращі результати, також були об'єднані для створення найкращої моделі.

CapsNet, у поєднанні з VGG19 або EfficientNet B4, демонструє конкурентоспроможну продуктивність. Об'єднання CapsNet + VGG19 та CapsNet + EffB4 забезпечує AUC 83.50 та ACC 73.29, що підкреслює ефективність оригінальної архітектури CapsuleNet у комбінації з різними екстракторами ознак для виявлення діпфейків. Введення коригувань ваг втрат у CapsNet у моделях об'єднання CapsNet (Loss) + VGG19 та CapsNet (Loss) + EffB4 призводить до покращення продуктивності за ACC. Це об'єднання досягає AUC 83.40 та ACC 75.35.

ArCapsNet зі своєю багатошаровою структурою також демонструє багатообіцяючі результати при об'єднанні. Об'єднання ArCaps (Layer1) + VGG19 та ArCaps (Layer1) + EffB4 досягає AUC 86.58 та ACC 79.87, що свідчить про конкурентоспроможну продуктивність. Подібні тенденції спостерігаються при об'єднанні ArCaps (Layer2) + VGG19 та ArCaps (Layer2) + EffB4, де досягається AUC 86.86 та ACC 79.10. Об'єднання ArCaps (Layer3) + VGG19 та ArCaps (Layer3) + EffB4 демонструє найкращу продуктивність з AUC 89.08 та ACC 82.84, що підкреслює ефективність глибших шарів у ArCapsNet для виявлення діпфейків.

Серед об'єднаних моделей ArCaps (Layer3) + VGG19 та ArCaps (Layer3) + EffB4 виявляються найбільш ефективною комбінацією, з найвищим AUC 89.08 та ACC 82.84. Це демонструє ефекти об'єднання моделей з підвищенням ACC на 2.32% і AUC на 1.52% відповідно.

Окрім об'єднання двох моделей, також було перевірено ефективність об'єднання трьох моделей. Об'єднання проводилося між комбінаціями моделей з усіма трьома екстракторами ознак. Для набору даних Celeb-DF найкраща модель для кожного шару ArCaps була об'єднана з комбінацією CapsNet (Loss) + VGG19 та CapsNet (Loss) + EffB4, що дало найкращі результати в об'єднанні двох моделей. Схожу стратегію було використано для набору даних DFDC-P. Усі три типи екстракторів ознак кожної моделі були об'єднані. Крім того, модель з найкращим результатом для кожного CapsNet була об'єднана з комбінацією ArCaps (Layer3) + VGG19 та ArCaps (Layer3) + EffB4, що також дало найкращі результати в об'єднанні двох моделей.

Таблиця 3.6 показує продуктивність об'єднання трьох моделей для набору даних Celeb-DF. Як видно, найвищий AUC у об'єднанні трьох моделей становить 99.10, а ACC 92.86. Ці результати відстають від значень, отриманих в об'єднанні двох моделей. Об'єднання двох моделей досягає кращих результатів з AUC на 0.04 і ACC на 0.77 в порівнянні з об'єднанням трьох моделей.

Таблиця 3.6- Результати об'єднання 3 моделей для набору даних Celeb-DF

Модель 1	Модель 2	Модель 3	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsNet + VGG19	CapsNet + EffB0	CapsNet + EffB4	99.10	90.54	99.24	73.03	84.14
CapsNet (Loss) + VGG19	CapsNet (Loss) + EffB0	CapsNet (Loss) + EffB4	99.09	92.86	97.32	81.46	88.69
ArCaps (Layer1) + VGG19	ArCaps (Layer1) + EffB0	ArCaps (Layer1) + EffB4	95.35	87.07	97.44	64.04	77.29
ArCaps (Layer2) + VGG19	ArCaps (Layer2) + EffB0	ArCaps (Layer2) + EffB4	97.06	89.38	96.24	71.91	82.32
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB0	ArCaps (Layer3) + EffB4	96.60	87.26	99.12	63.48	77.40
CapsNet (Loss) + VGG19	CapsNet (Loss) + EffB4	ArCaps (Layer1) + EffB4	97.77	90.93	99.25	74.16	84.89
CapsNet (Loss) + VGG19	CapsNet (Loss) + EffB4	ArCaps (Layer2) + EffB4	98.34	92.28	97.92	79.21	87.58
CapsNet (Loss) + VGG19	CapsNet (Loss) + EffB4	ArCaps (Layer3) + VGG19	97.96	90.15	97.74	73.03	83.60

Таблиця 3.7 представляє об'єднання трьох моделей на наборі даних DFDC-P. Як видно, в найкращому випадку AUC становить 88.81, а ACC — 82.32. Ці результати також відстають від значень, отриманих в об'єднанні двох моделей. Об'єднання двох моделей досягає кращих результатів з AUC на 0.27 і ACC на 0.52 в порівнянні з об'єднанням трьох моделей. Отже, видно, що об'єднання двох моделей дає кращі результати, ніж об'єднання трьох моделей.

Об'єднання моделей доводить свою цінність як стратегія для підвищення продуктивності, з помітними покращеннями в AUC та ACC в різних конфігураціях. Комбінація моделей CapsuleNet та ArCapsNet, разом з стратегічним вибором екстракторів ознак, демонструє потенціал для покращення можливостей виявлення діпфейків. Підсумовуючи, експериментальні результати підкреслюють ефективність об'єднання моделей та важливість вибору екстракторів ознак для досягнення оптимальної продуктивності у завданнях виявлення діпфейків.

Таблиця 3.7 - Результати об'єднання 3 моделей для набору даних DFDC-P

Модель 1	Модель 2	Модель 3	AUC	ACC	Точність (Precision)	Повнота (Recall)	F1
CapsNet + VGG19	CapsNet + EffB0	CapsNet + EffB4	83.43	74.32	60.66	79.35	68.76
CapsNet (Loss) + VGG19	CapsNet (Loss) + EffB0	CapsNet (Loss) + EffB4	83.41	72.39	58.33	78.62	66.98
ArCaps (Layer1) + VGG19	ArCaps (Layer1) + EffB0	ArCaps (Layer1) + EffB4	86.20	78.58	67.86	75.72	71.58
ArCaps (Layer2) + VGG19	ArCaps (Layer2) + EffB0	ArCaps (Layer2) + EffB4	86.59	79.61	69.93	75.00	72.38
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB0	ArCaps (Layer3) + EffB4	88.18	81.29	73.14	75.00	74.06
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB4	CapsNet + EffB4	88.67	82.32	74.22	77.17	75.67
ArCaps (Layer3) + VGG19	ArCaps (Layer3) + EffB4	CapsNet (Loss) + VGG19	88.81	81.68	73.59	75.72	74.64

3.3 Порівняння з попередніми роботами

Рисунки 3.3 та 3.4 ілюструють порівняльну продуктивність різних методів на наборах даних Celeb-DF та DFDC-P відповідно. Важливо зазначити, що запропонований метод перевершує всі попередні дослідження як за точністю (ACC), так і за площею під кривою (AUC).

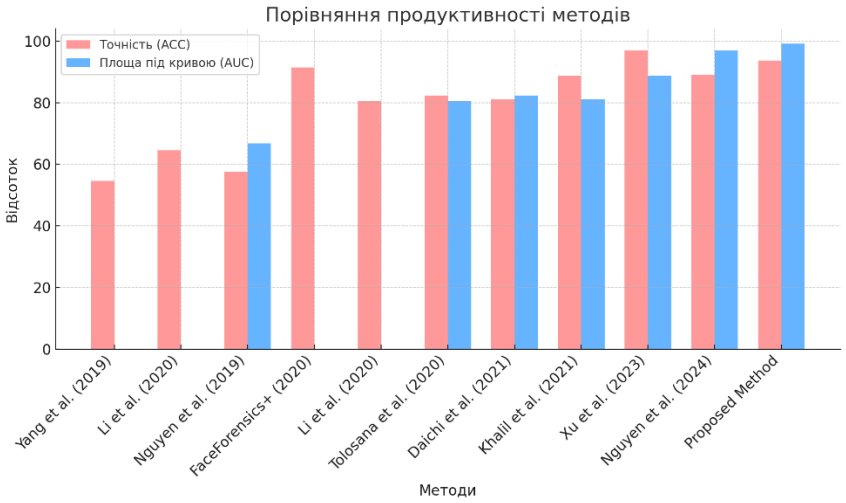


Рисунок. 3.3 - Порівняння результатів ACC та AUC (%) у наборі даних Celeb-DF

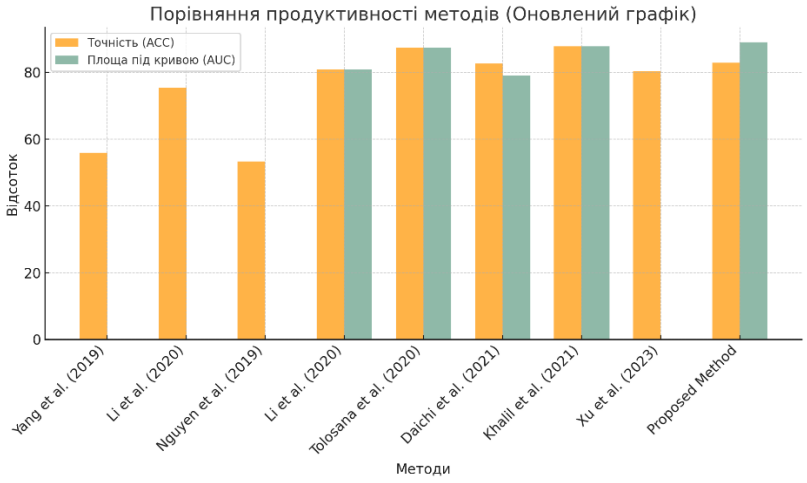


Рисунок. 3.4 - Порівняння результатів ACC та AUC (%) у наборі даних DFDC-P

У попередніх роботах 2019 року, як приклад Yang et al. [18] та Nguyen et al. [30], досягнуто відносно низьких показників AUC, що підкреслює еволюцію методів виявлення діпфейків з часом. З моменту свого зародження сфера виявлення діпфейків зробила суттєвий прогрес, і більш сучасні підходи дають помітно кращі результати. Запропонований метод демонструє еволюцію та досягнення у цій галузі, забезпечуючи значно вищу точність і показники AUC.

Порівнюючи роботи Nguyen [30] та Tolosana [56], обидві використовували схожу архітектуру капсульної мережі, але підхід Tolosana демонструє вищу продуктивність. Це розходження у результатах може бути пояснене різницями в обробці зображень або іншими нюансами реалізації, підкреслюючи важливість вибору архітектури та попередньої обробки даних у виявленні діпфейків.

Daichi et al. [36] (2021), Khalil et al. [53] (2021) та Xu et al. [57] (2023) представляють більш сучасні роботи, які досягли гідних результатів. Однак запропонований метод не тільки перевершує їхні результати за показниками ACC і AUC на обох наборах даних, але й демонструє суттєве покращення можливостей виявлення діпфейків.

Особливо запропонований метод виділяється як лідер у виявленні діпфейків. Досягнувши високої ACC 93.63 і вражаючого AUC 99.14 на наборі даних Celeb-DF, а також ACC 82.84 і AUC 89.08 на DFDC-P, він перевершує попередні найкращі методи з великим відривом.

Розглядаючи таблицю 3.8, можна побачити, що запропонована модель досягає кращих результатів, ніж інші дослідження, за метрикою F1 на наборі даних Celeb-DF.

Таблиця 3.8 - Порівняльна таблиця F1 на наборі даних Celeb-DF

Метод	ACC (%)	F1 (%)
CViT (2021) [32]	73.85	68.91
DSP-FWA (2020) [52]	64.60	61.93
Face X-ray (2020) [17]	54.87	50.79

Продовження таблиці 3.8

Метод	ACC (%)	F1 (%)
3D ResNet (2022) [33]	68.17	66.71
RECCE (2022) [34]	72.48	71.87
RFM (2021) [35]	70.72	69.94
Daichi et al. (2021) [36]	81.08	82.42
DCL (2022) [37]	75.58	72.70
FedForgery (2023) [38]	75.74	72.87
HCIT (2024) [42]	76.03	73.96
Запропонована модель	93.63	89.85

3.3.1 Аналіз складності моделей

Аналіз складності є важливим для розуміння компромісів між продуктивністю моделей та їхньою обчислювальною ефективністю. Таблиця 3.9 надає порівняння різних моделей глибокого навчання за кількістю параметрів та операцій з плаваючою точкою за секунду (FLOPs).

Базова модель CapsuleNet, позначена як "Capsule", з AUC 57.50% має відносно низьку продуктивність порівняно з іншими моделями та використовує 3.90 мільйонів параметрів і 11.15 гігафлопсів. VidTr, який досягнув помітного показника AUC 83.30%, має вищу складність із 93.00 мільйонами параметрів і 117.00 гігафлопсів. ISTVT при AUC 84.10% не надає інформації про кількість параметрів, але демонструє значне навантаження з 455.8 гігафлопсами.

CapsNet + VGG19 забезпечує баланс між продуктивністю та складністю, досягаючи вражаючого AUC 97.54% з 3.90 мільйонами параметрів і 8.14 гігафлопсів. Поєднання CapsNet із архітектурами EfficientNet демонструє ще більшу ефективність, де CapsNet + EffB4 досягає AUC 98.61% із лише 1.86 мільйонами параметрів та 1.08 гігафлопсів, що підкреслює ефективність використання економних архітектур без зниження продуктивності.

ArCapsNet також показує цікаві результати, досягаючи конкурентоспроможних значень AUC (від 82.89% до 98.61%) у різних конфігураціях шарів (Layer1, Layer2, Layer3) у поєднанні з VGG19 або EfficientNet. Незважаючи на високі показники кількості параметрів і FLOPs, продуктивність цих моделей залишається стабільною, що демонструє їхню здатність до масштабування при збереженні обчислювальної ефективності.

Таблиця 3.9 - Порівняння кількості параметрів та FLOPs

Модель	Celeb-DF AUC (%)	Параметри	FLOPs (G)	Кількість кадрів на відео
Capsule [30]	57.50	3.90M	11.15	Перші 100
VidTr [14]	83.30	93.00M	117.00	Усі кадри відео
ISTVT [15]	84.10	–	455.8	Усі кадри відео
CapsNet + VGG19	97.54	3.90M	8.14	1
CapsNet + EffB0	98.10	1.65M	0.89	1
CapsNet + EffB4	98.61	1.86M	1.08	1
ArCapsNet (Layer1) + VGG19	89.88	205.54M	7.35	1
ArCapsNet (Layer1) + EffB0	82.89	203.26M	0.09	1
ArCapsNet (Layer1) + EffB4	91.40	203.47M	0.28	1
ArCapsNet (Layer2) + VGG19	93.38	204.91M	7.35	1
ArCapsNet (Layer2) + EffB0	94.57	202.63M	0.09	1
ArCapsNet (Layer2) + EffB4	96.41	202.84M	0.28	1
ArCapsNet (Layer3) + VGG19	94.02	205.07M	7.35	1
ArCapsNet (Layer3) + EffB0	88.77	202.79M	0.09	1
ArCapsNet (Layer3) + EffB4	91.00	203.00M	0.28	1

Однією з примітних особливостей наших моделей є їхня ефективність у тренуванні. У той час як інші дослідження використовують стратегію випадкового вибору великої кількості кадрів із кожного відео під час тренування, наші моделі використовують лише один кадр на відео. Це не тільки спрощує процес навчання, але й сприяє покращенню обчислювальної продуктивності.

Значення складності для моделей були розраховані для обробки одного зображення, що забезпечує значну перевагу з точки зору часу обчислень.

3.3.2 Статистичний аналіз

Для подальшого уточнення результатів експериментів було використано методи статистичного тестування, такі як тест Вілкоксона та тест Фрідмана. Для наборів даних Celeb-DF та DFDC-P були відібрані випадки, які досягли найкращих результатів точності серед одиночних моделей з таблиць 3.2 та 3.3, моделей подвійного об'єднання з таблиць 3.4 та 3.5, а також моделей потрійного об'єднання з таблиць 3.6 та 3.7.

Після цього було застосовано тест Вілкоксона для визначення, чи є різниця між одиночною моделлю та іншими випадками об'єднання. Результати тесту Вілкоксона наведені в таблиці 3.10. Значення P-value менше 0.05 вказує на наявність значущої різниці між порівнюваними моделями, а більше 0.05 — на її відсутність. Як видно з результатів у Таблиця 3.10, є різниця між одиночною моделлю та подвійною об'єднаною моделлю, однак різниці між потрійною об'єднаною моделлю немає.

Таблиця 3.10 - Тест Вілкоксона

Випадок	Статистика тесту	P-value	Результат
Одиночна–Подвійне об'єднання	0	0.001953	Значуща різниця
Одиночна–Потрійне об'єднання	7	0.066316	Немає значущої різниці

Також було застосовано тест ранжування Фрідмана, який дозволяє порівнювати три моделі. Це дає можливість провести більш просунуте тестування наших моделей. Як видно з таблиці 3.11, одиночні, подвійні та потрійні моделі об'єднання мають значні відмінності у результатах. Ці тести показують, що модель подвійного об'єднання створює значну різницю в

результатах продуктивності, але потрійна модель об'єднання не приносить помітної різниці.

Таблиця 3.11 - Тест Фрідмана

Тест	Статистика тесту	P-value	Результат
Тест Фрідмана	13.282	0.001305	Значущі відмінності

Статистичний аналіз, проведений за допомогою тестів Вілкоксона та Фрідмана, підтвердив ефективність моделей подвійного об'єднання в порівнянні з одиночними моделями, демонструючи значущу різницю в продуктивності (P-value = 0.001953). Водночас, результати показали, що додавання третьої моделі в об'єднання не дає статистично значущого покращення (P-value = 0.066316), що свідчить про те, що потрійне об'єднання не приносить додаткових переваг. Тест Фрідмана також підтвердив наявність значних відмінностей між моделями, підкреслюючи, що подвійне об'єднання є найбільш ефективним підходом для покращення точності виявлення діпфейкових відео.

Висновки до 3 розділу.

1. Проведено аналіз і тестування запропонованих моделей для виявлення діпфейкових відео на двох еталонних наборах даних: Celeb-DF та DFDC-P. Ці набори даних надали різноманітні умови для перевірки ефективності моделей, що включали різні типи відео та їх складність. Результати показали високу продуктивність запропонованої моделі у поєднанні з трансферним навчанням та капсульними мережами.
2. Порівняння з іншими сучасними методами (SOTA) виявило значну перевагу нашого підходу. Запропоновані моделі показали покращені результати за такими показниками, як AUC та точність, що вказує на їхню високу здатність до виявлення діпфейкових відео. Використання комбінованих моделей також

суттєво підвищило загальну точність системи, що підкреслює важливість об'єднання різних архітектур для досягнення найкращих результатів.

3. Проведени аналіз складності моделей, показав оптимізацію між продуктивністю та обчислювальними витратами. Об'єднання різних моделей дозволило не тільки підвищити точність, але й зменшити час навчання. Статистичний аналіз результатів підтвердив значущу різницю між окремими моделями та моделями з об'єднанням, що вказує на ефективність цього підходу в завданні виявлення діпфейків.

ВИСНОВКИ

1. У ході цієї роботи була розроблена багатоступенева модель для виявлення діпфейкових відео, що поєднує різні техніки машинного навчання, трансферного навчання та капсульних мереж. Основною метою дослідження було покращення точності і надійності алгоритмів для виявлення діпфейків, що досягається шляхом поєднання різних моделей і глибоких архітектур. Запропонована модель пройшла тестування на двох еталонних наборах даних — Celeb-DF та DFDC-P, що є загальновизнаними у цій галузі.

2. Першим кроком було тестування моделі на наборі даних Celeb-DF, де модель показала високу точність з результатами AUC = 99.14% і точністю (ACC) = 93.63%. Важливо зазначити, що ці результати значно перевищують попередні методи, представлені в літературі, зокрема такі як CViT та Face X-ray, що досягли AUC на рівні 73.85% і 54.87% відповідно. Це свідчить про значне покращення завдяки застосуванню трансферного навчання і капсульних мереж, а також завдяки використанню інноваційного етапу попередньої обробки.

3. На наборі даних DFDC-P запропонована модель також продемонструвала високі показники. Результати показали AUC = 89.08% та точність (ACC) = 82.84%, що також значно перевищує результати сучасних методів, таких як 3D ResNet та RFM. У попередніх дослідженнях максимальний AUC на цьому наборі даних складав 84.12%, що підкреслює важливість нових підходів, представлених у цій роботі.

4. Запропонована модель використовувала декілька популярних екстракторів ознак, таких як VGG19, EfficientNet B0 та EfficientNet B4. Найкращі результати були досягнуті у поєднанні з EfficientNet B4, що забезпечило AUC = 98.61% на Celeb-DF і AUC = 87.56% на DFDC-P. Це свідчить про те, що обраний екстрактор ознак є ефективним і дозволяє отримати якісні результати без значного збільшення обчислювальних витрат.

5. Також було показано, що CapsuleNet з використанням коригування ваг втрат демонструє найкращі результати серед усіх архітектур. На наборі даних Celeb-DF вона досягла AUC = 98.61% та точності (ACC) = 93.05%. На наборі даних DFDC-P ця модель продемонструвала AUC = 80.55% і точність (ACC) = 72.13%, що також перевищує результати інших методів.

6. Об'єднання моделей CapsuleNet та ArCapsNet показало особливо високі результати. Комбінація цих моделей призвела до покращення точності і AUC. Наприклад, на наборі даних Celeb-DF об'єднання CapsuleNet (Loss Weight) та ArCapsNet (Layer1) з EfficientNet B4 досягло AUC = 99.14% і точності (ACC) = 93.63%, що підкреслює перевагу комбінованих підходів у задачах класифікації.

7. Важливо зазначити, що модель була протестована на складних умовах із різними освітленням, рухами облич та емоціями у відео, що дозволило перевірити її здатність працювати з різними сценаріями. Використання лише одного кадру з відео під час навчання значно зменшило обчислювальні витрати, що покращило продуктивність системи.

8. Статистичний аналіз результатів показав значну різницю між одиночними моделями та моделями з об'єднанням, що підтверджується тестами Вілкоксона і Фрідмана. Об'єднані моделі показали приріст у показниках точності та AUC на 0.53% та 0.77% відповідно порівняно з одиночними моделями. Це свідчить про важливість застосування стратегії об'єднання моделей для досягнення найкращих результатів.

Таким чином, запропонована багатоступенева модель для виявлення діпфейкових відео демонструє високу ефективність та надійність, що підтверджується її успішним застосуванням на двох великих еталонних наборах даних. Отримані кількісні результати свідчать про те, що модель може бути використана у реальних умовах для вирішення завдань класифікації діпфейкових відео.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. https://doi.org/10.1145/3422622
2. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*. https://arxiv.org/abs/1312.6114
3. McTaggart, I. (2021, March 5). Tom Cruise deepfake creator has spoken out and warns of the risks the technology poses. *The Telegraph*. https://www.telegraph.co.uk/news/2021/03/05/tom-cruise-deepfake-creator-has-spoken-warns-risks-technology/
4. Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., & Nyameko, R. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access, 11*, 143296-143323. https://doi.org/10.1109/ACCESS.2023.3342107
5. Koopman, M., Rodriguez, A. M., & Geradts, Z. (2018). Detection of deepfake video manipulation. In *Proceedings of the 20th Irish Machine Vision and Image Processing Conference* (pp. 133-136).
6. Lugstein, F., Baier, S., Bachinger, G., & Uhl, A. (2021). PRNU-based deepfake detection. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security* (pp. 7-12). https://doi.org/10.1145/3417027.3417037
7. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deepfake image recognition. In *International Conference on Machine Learning* (pp. 3247-3258).
8. Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). https://doi.org/10.1109/AVSS.2018.8639163

9. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces, 3*(1), 80-87.
10. Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., & Horváth, J. (2020). Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 668-669).
11. Zhao, Y., Ge, W., Li, W., Wang, R., Zhao, L., & Ming, J. (2020). Capturing the persistence of facial expression features for deepfake video detection. *Information and Communications Security: 21st International Conference, ICICS 2019, Revised Selected Papers*, 21, 630-645. https://doi.org/10.1007/978-3-030-52304-7_39
12. Wu, X., Xie, Z., Gao, Y., & Xiao, Y. (2020). SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2952-2956). https://doi.org/10.1109/ICASSP40776.2020.9054037
13. Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020). Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of the European Conference on Computer Vision (ECCV), Part VII* (pp. 667-684). https://doi.org/10.1007/978-3-030-58598-4_39
14. Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., & Brattoli, B. (2021). VidTr: Video transformer without convolutions. *arXiv preprint arXiv:2104.11746*. https://arxiv.org/abs/2104.11746
15. Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., & Tang, J. (2023). ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1335-1348. https://doi.org/10.1109/TIFS.2023.3239223
16. Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*. https://arxiv.org/abs/1811.00656

17. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., & Wen, F. (2020). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5001-5010). https://doi.org/10.1109/CVPR42600.2020.00503
18. Yang, X., Li, Y., & Lyu, S. (2019). Exposing deepfakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261-8265). https://doi.org/10.1109/ICASSP.2019.8683164
19. Jung, T., Kim, S., & Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access, 8*, 83144-83154. https://doi.org/10.1109/ACCESS.2020.2990749
20. Ranjan, R., Patel, V. M., & Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(1), 121-135. https://doi.org/10.1109/TPAMI.2017.2781233
21. Soukupova, T., & Cech, J. (2016). Eye blink detection using facial landmarks. In *Proceedings of the 21st Computer Vision Winter Workshop* (pp. 1-8).
22. Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.3038438
23. Wu, J., Zhu, Y., Jiang, X., Liu, Y., & Lin, J. (2024). Local attention and long-distance interaction of rPPG for deepfake detection. *The Visual Computer, 40*(2), 1083-1094. https://doi.org/10.1007/s00371-023-02011-1
24. Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., & Feng, W. (2020). Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 4318-4327). https://doi.org/10.1145/3394171.3413590

25. Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). DeepFakesOn-Phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*. https://arxiv.org/abs/2010.00400
26. Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1831-1839). https://doi.org/10.1109/CVPRW.2017.231
27. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Guo, H. (2019). Fakespotter: A simple yet robust baseline for spotting AI-synthesized fake faces. *arXiv preprint arXiv:1909.06122*. https://arxiv.org/abs/1909.06122
28. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1-11). https://doi.org/10.1109/ICCV.2019.00009
29. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*. https://arxiv.org/abs/1910.08854
30. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*. https://arxiv.org/abs/1910.12467
31. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-7). https://doi.org/10.1109/WIFS.2018.8630761
32. Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*. https://arxiv.org/abs/2102.11126
33. Roy, R., Joshi, I., Das, A., & Dantcheva, A. (2022). 3D CNN architectures and attention mechanisms for deepfake detection. In *Handbook of Digital Face*

Manipulation and Detection: From Deepfakes to Morphing Attacks (pp. 213-234). Springer International Publishing. https://doi.org/10.1007/978-3-030-87664-7_10

34. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4103-4112). https://doi.org/10.1109/CVPR52688.2022.00408
35. Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14923-14932). https://doi.org/10.1109/CVPR46437.2021.01467
36. Zhang, D., Li, C., Lin, F., Zeng, D., & Ge, S. (2021). Detecting deepfake videos with temporal dropout 3DCNN. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1288-1294). https://doi.org/10.24963/ijcai.2021/178
37. Sun, K., Yao, T., Chen, S., Ding, S., Li, J., & Ji, R. (2022). Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 2316-2324. https://doi.org/10.1609/aaai.v36i2.20060
38. Liu, D., Dang, Z., Peng, C., Zheng, Y., Li, S., Wang, N., & Huang, Q. (2023). FedForgery: Generalized face forgery detection with residual federated learning. *IEEE Transactions on Information Forensics and Security*. https://doi.org/10.1109/TIFS.2023.3298464
39. Asha, S., Vinod, P., & Menon, V. (2024). A defensive attention mechanism to detect deepfake content across multiple modalities. *Multimedia Systems, 30*, 87-102. https://doi.org/10.1007/s00530-023-01248-x
40. Heidari, A., Navimipour, N. J., Dag, H., Talebi, S., & Unal, M. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. *Cognitive Computation*. https://doi.org/10.1007/s12559-024-10145-2

41. Nguyen, D., Mejri, N., Singh, I. P., Kuleshova, P., Astrid, M., & Kacem, A. (2024). LAA-Net: Localized artifact attention network for high-quality deepfakes detection. *arXiv preprint arXiv:2401.13856*. https://arxiv.org/abs/2401.13856
42. Kaddar, B., Fezza, S. A., Akhtar, Z., Hamidouche, W., Hadid, A., & Serra-Sagristà, J. (2024). Deepfake detection using spatiotemporal transformer. *ACM Transactions on Multimedia Computing, Communications, and Applications*. https://doi.org/10.1145/3599892
43. Faceswap. (2021). Deepfakes software for all. Retrieved from https://faceswap.dev/
44. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503. https://doi.org/10.1109/LSP.2016.2603342
45. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). S3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 192-201). https://doi.org/10.1109/ICCV.2017.120
46. Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1021-1030). https://doi.org/10.1109/ICCV.2017.118
47. Aksoylu, A. (2021). Aksoylu/GoldenFace: An image processing library about calculating face golden ratio, facial cosine similarity and more. Retrieved from https://github.com/Aksoylu/GoldenFace
48. Kaya, K. S., Türk, B., Cankaya, M., Seyhun, N., & Coşkun, B. U. (2019). Assessment of facial analysis measurements by golden proportion. *Brazilian Journal of Otorhinolaryngology, 85*(4), 494-501. https://doi.org/10.1016/j.bjorl.2018.04.006
49. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information*

Processing Systems (pp. 3856-3866). https://dl.acm.org/doi/10.5555/3295222.3295296

50. Choi, J., Seo, H., Im, S., & Kang, M. (2019). Attention routing between capsules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).
51. Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. In *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN 2011)* (pp. 44-51). Springer. https://doi.org/10.1007/978-3-642-21735-7_6
52. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3207-3216). https://doi.org/10.1109/CVPR42600.2020.00326
53. Khalil, S. S., Youssef, S. M., & Saleh, S. N. (2021). ICaps-DFake: An integrated capsule-based model for deepfake image and video detection. *Future Internet*, *13*(4), 93. https://doi.org/10.3390/fi13040093
54. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://arxiv.org/abs/1409.1556
55. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105-6114). PMLR. https://arxiv.org/abs/1905.11946
56. Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R. (2021). Deepfakes evolution: Analysis of facial regions and fake detection performance. In *Proceedings of the International Conference on Pattern Recognition* (pp. 442-456). Springer. https://doi.org/10.1007/978-3-030-87664-7_40
57. Xu, K., Yang, G., Fang, X., & Zhang, J. (2023). Facial depth forgery detection based on image gradient. *Multimedia Tools and Applications*, 1-25. https://doi.org/10.1007/s11042-023-14771-7

58. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.
59. Комар М.П., Саченко А.О., Васильків Н.М., Загородня Д.І. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. – Тернопіль: ЗУНУ, 2024. – 32 с.
60. Колівушко Е., Лучка С., Матвійчук М. Інтелектуальні методи виявлення маніпулятивного контенту. Тренди та перспективи розвитку мультидисциплінарних досліджень: матеріали VI Міжнародної студентської наукової конференції, м. Кременчук, 11 жовтня, 2024 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2024, 112 -113 с.
61. Колівушко Е., Лучка С., Кондратюк Г., Кравчук Б., Тарасюк С., Прикладні аспекти використання машинного навчання для обробки текстових даних. Стратегічні напрямки розвитку науки: фактори впливу та взаємодії: збірник наукових праць з матеріалами V Міжнародної наукової конференції, м. Луцьк, 11 жовтня, 2024 р. / Міжнародний центр наукових досліджень. — Вінниця: ТОВ «УКРЛОГОС Груп, 2024, 154-156 с.

Додаток А

Апробація отриманих результатів

ЗБІРНИК НАУКОВИХ ПРАЦЬ

З МАТЕРІАЛАМИ V МІЖНАРОДНОЇ НАУКОВОЇ КОНФЕРЕНЦІЇ

11 ЖОВТНЯ 2024 РІК

М. ЛУЦЬК, УКРАЇНА

**«СТРАТЕГІЧНІ НАПРЯМКИ РОЗВИТКУ НАУКИ:
ФАКТОРИ ВПЛИВУ ТА ВЗАЄМОДІЇ»**

УДК 082:001
С 83

Організація, від імені якої випущено видання:
ГО «Міжнародний центр наукових досліджень»
Номер запису організації в Єдиному реєстрі громадських об'єднань: 1499141

Голова оргкомітету: Сотник С.Г.

Верстка: Білоус Т.В.
Дизайн: Бондаренко І.В.

Рекомендовано до видання Вченою Радою Інституту науково-технічної інтеграції та співпраці. Протокол № 57 від 10.10.2024 року.

*Конференцію зареєстровано Державною науковою установою у сфері управління Міністерства освіти і науки «Український інститут науково-технічної експертизи та інформації» в базі даних науково-технічних заходів України на поточний рік та бюлетені «План проведення наукових, науково-технічних заходів в Україні» (**Посвідчення № 351 від 12.06.2024**).*

*Збірник наукових праць з матеріалами конференції видано офіційно суб'єктом видавничої справи зі **Свідоцтвом ДК № 7860 від 22.06.2023***

С 83 **Стратегічні напрямки розвитку науки: фактори впливу та взаємодії**: збірник наукових праць з матеріалами V Міжнародної наукової конференції, м. Луцьк, 11 жовтня, 2024 р. / Міжнародний центр наукових досліджень. — Вінниця: ТОВ «УКРЛОГОС Груп, 2024. — 254 с.

ISBN 978-617-8440-16-9
DOI 10.62731/mcnd-11.10.2024

Викладено матеріали учасників V Міжнародної наукової конференції «Стратегічні напрямки розвитку науки: фактори впливу та взаємодії», яка відбулася 11 жовтня 2024 року у місті Луцьк.

УДК 082:001

ISBN 978-617-8440-16-9

СЕКЦІЯ XI.
ХІМІЯ, ХІМІЧНА ТА БІОІНЖЕНЕРІЯ

ЗАСТОСУВАННЯ НЕОРГАНІЧНИХ СПОЛУК ДЛЯ НАДАННЯ ВОЛОКНИСТИМ МАТЕРІАЛАМ БАКТЕРИЦИДНИХ ВЛАСТИВОСТЕЙ У МЕДИЧНИХ І ПРОФІЛАКТИЧНИХ ЦІЛЯХ
Качківський В.В., Попович Т.А. **130**

СЕКЦІЯ XII.
ЕЛЕКТРОНІКА ТА ТЕЛЕКОМУНІКАЦІЇ

ЗАСТОСУВАННЯ ТЕХНОЛОГІЇ БЕЗДРОТОВОЇ ЗАРЯДКИ ЗА ДОПОМОГОЮ РЕКТЕН ДЛЯ ЖИВЛЕННЯ СИСТЕМ РЕЗ
Сокіркаєв Д.В. **135**

СЕКЦІЯ XIII.
ЕКОЛОГІЯ ТА ТЕХНОЛОГІЇ ЗАХИСТУ НАВКОЛИШНЬОГО СЕРЕДОВИЩА

ІМПЛЕМЕНТАЦІЯ ЄВРОПЕЙСЬКОГО ДОСВІДУ ЕКОЛОГІЧНОГО РЕГУЛЮВАННЯ ВИКОРИСТАННЯ ТЕРИТОРІЙ ПРИРОДНО-ЗАПОВІДНОГО ФОНДУ
Голян В.М., Іванців В.В. **138**

СЕКЦІЯ XIV.
КОМП'ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

HARDWARE SUPPORT OF INFORMATION SYSTEMS
Pavlyk H.V. **142**

СЕКЦІЯ XV.
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА СИСТЕМИ

ПОБУДОВА ГНУЧКИХ СТРАТЕГІЙ УПРАВЛІННЯ ЗМІНАМИ В ML-ПРОЄКТАХ
Науково-дослідна група:
Висоцький А.В., Левандівський Н.Б., Вівюрка Н.М., Сулима Б.Я. **151**

ПРИКЛАДНІ АСПЕКТИ ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ОБРОБКИ ТЕКСТОВИХ ДАНИХ
Науково-дослідна група:
Колівушко Е., Лучка С., Кондратюк Г., Кравчук Б., Тарасюк С. **154**

ПРИКЛАДНІ АСПЕКТИ ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ОБРОБКИ ТЕКСТОВИХ ДАНИХ

НАУКОВО-ДОСЛІДНА ГРУПА:

Колівушко Едуард
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Лучка Святослав
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Кондратюк Георгій
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Кравчук Богдан
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Тарасюк Софія
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Ліп'яніна-Гончаренко Христина
канд. техн. наук, доцент,
доцент кафедри інформаційно-обчислювальних систем та управління
Західноукраїнський національний університет, Україна

Машинне навчання (ML) кардинально змінило підходи до обробки текстових даних, дозволяючи автоматизувати складні завдання, такі як класифікація тексту, аналіз емоцій та виявлення фейкових новин. У зв'язку зі зростанням обсягів текстової інформації в Інтернеті, особливо на соціальних платформах, машинне навчання стало ключовим інструментом для швидкої і точної обробки цих даних. За статистикою, кожного дня публікується понад 2,5 мільярда текстових записів на платформах соціальних мереж [5], що робить застосування машинного навчання надзвичайно актуальним.

Існує декілька основних методів машинного навчання, які активно використовуються для обробки тексту. Серед них найпопулярніші — наївний Баєс, метод опорних векторів (SVM), логістична регресія та дерева рішень. Проте з розвитком глибинного навчання в останні роки стали особливо популярними рекурентні нейронні мережі (RNN) та трансформери (наприклад, BERT), що забезпечують високий рівень точності в завданнях класифікації тексту і прогнозуванні наступних слів у реченнях [3].

Попередня обробка тексту є критично важливою для підготовки даних до аналізу. Найбільш ефективними методами є стемінг, лематизація та видалення стоп-слів. Стемінг перетворює слова до їх основних форм, що допомагає зменшити варіативність даних, тоді як лематизація враховує контекст і надає точніші результати. Видалення стоп-слів допомагає зменшити розмір тексту, видаляючи непотрібні слова, які не несуть суттєвої інформації для моделі [4].

Якість попередньої обробки тексту безпосередньо впливає на результати моделей машинного навчання. Якщо дані погано підготовлені, моделі можуть працювати неефективно через надмірну кількість шуму або неправильні структури даних. Наприклад, неправильне видалення стоп-слів або відсутність лематизації може призвести до того, що модель не зможе точно розпізнати ключові патерни тексту, що знижує точність класифікації [2]. Таким чином, ретельна обробка тексту є важливим етапом у процесі машинного навчання.

Векторизація тексту — це процес перетворення текстових даних у числові представлення, які можуть бути використані моделями машинного навчання. Найпоширеніші підходи включають Bag of Words (BoW), TF-IDF і word embeddings (наприклад, Word2Vec, GloVe). BoW і TF-IDF є простими методами, що використовуються для побудови векторів на основі частоти слів. Однак word embeddings дозволяють моделі навчитися семантичним зв'язкам між словами, що значно покращує якість аналізу тексту [3].

Вибір техніки векторизації залежить від характеру задачі. Якщо основний фокус на простій класифікації тексту, такі методи, як TF-IDF або Bag of Words, можуть бути достатніми. Однак для задач, що вимагають врахування контексту та семантики (наприклад, аналіз емоцій або прогнозування наступних слів), більш ефективними будуть word

embeddings або трансформери на основі BERT. Для задач, де точність є критично важливою, варто віддавати перевагу більш складним підходам [1].

Аналіз емоцій є одним з основних застосувань машинного навчання для обробки тексту, особливо у сфері маркетингу, соціальних мереж та підтримки клієнтів. Алгоритми машинного навчання, такі як нейронні мережі та методи класифікації, можуть використовуватися для визначення тональності тексту, що дозволяє автоматично аналізувати настрої користувачів. Успішні кейси використання цих методів вже застосовуються у великих компаніях, таких як Amazon і Google, для аналізу зворотного зв'язку клієнтів [2].

У результаті дослідження виявлено, що машинне навчання є ефективним інструментом для обробки текстових даних, особливо при правильній попередній обробці та виборі підходящої техніки векторизації. Для більш складних завдань, таких як аналіз емоцій або семантична класифікація, рекомендується використовувати сучасні моделі глибинного навчання, що базуються на embeddings або трансформерах. Подальші дослідження можуть зосередитися на оптимізації процесу векторизації та покращенні якості моделей на багатомовних даних.

Список використаних джерел:

1. Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 2017.
2. Pang, B., & Lee, L. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval*, 2008.
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*, 2013.
4. Manning, C., Raghavan, P., & Schütze, H. "Introduction to Information Retrieval." *Cambridge University Press*, 2008.
5. Statista. "Number of social media users worldwide from 2010 to 2021." [Online]. Available: www.statista.com.


МАТЕРІАЛИ VI МІЖНАРОДНОЇ
СТУДЕНТСЬКОЇ НАУКОВОЇ
КОНФЕРЕНЦІЇ
ТРЕНДИ ТА ПЕРСПЕКТИВИ
РОЗВИТКУ МУЛЬТИДИСЦИП-
ЛІНАРНИХ ДОСЛІДЖЕНЬ
ISO
М. КРЕМЕНЧУК, УКРАЇНА
11 ЖОВТНЯ
2024 РІК

УДК 082:001
Т 66

Голова оргкомітету: Коренюк І.О.
Верстка: Зрада С.І.
Дизайн: Бондаренко І.В.

Рекомендовано до видання Вченою Радою Інституту науково-технічної інтеграції та співпраці. Протокол № 57 від 10.10.2024 року.

Конференцію зареєстровано Державною науковою установою «УкрІНТЕІ» в базі даних науково-технічних заходів України та бюлетені «План проведення наукових, науково-технічних заходів в Україні» (Посвідчення №320 від 12.06.2024).

Т 66 **Тренди та перспективи розвитку мультидисциплінарних досліджень**: матеріали VI Міжнародної студентської наукової конференції, м. Кременчук, 11 жовтня, 2024 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2024. — 198 с.

ISBN 978-617-8440-34-3
DOI 10.62732/liga-inter-11.10.2024

Викладено матеріали учасників VI Міжнародної мультидисциплінарної студентської наукової конференції «Тренди та перспективи розвитку мультидисциплінарних досліджень», яка відбулася 11 жовтня 2024 року у місті Кременчук, Україна.

УДК 082:001

ISBN 978-617-8440-34-3

СЕКЦІЯ 11.
АГРАРНІ НАУКИ ТА ПРОДОВОЛЬСТВО

ЕКОСИСТЕМНІ ПОСЛУГИ ЛІСОВОГО ГОСПОДАРСТВА
Луців В.В., ***Науковий керівник: Бондар О.Б.*** **95**

СЕКЦІЯ 12.
ЕЛЕКТРОНІКА ТА ТЕЛЕКОМУНІКАЦІЇ

ПРИВАТНІСТЬ ТА ЕТИЧНІ ПИТАННЯ ВІДЕОСПОСТЕРЕЖЕННЯ У ГРОМАДСЬКИХ МІСЦЯХ
Зеленський Д.М. **97**

СЕКЦІЯ 13.
КОМП'ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

LOGIC-FREE ВЕКТОРНИЙ КОМП'ЮТИНГ
Лук'янов Я.І., ***Науковий керівник: Хаханова Г.В.*** **99**

THE EVOLUTION OF SOFTWARE ENGINEERING: FROM CODE MONKEYS TO ARCHITECTS
Almakadma M.I., Pershyna A.A. **103**

СЕКЦІЯ 14.
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА СИСТЕМИ

ВИКОРИСТАННЯ ІНСТРУМЕНТІВ AUTOML ДЛЯ ОПТИМІЗАЦІЇ ОБРОБКИ ДАНИХ
Власенко Є.С., ***Науковий керівник: Слюсар В.І.*** **105**

ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ ВИКОРИСТАННЯ СУЧАСНИХ ТЕХНОЛОГІЙ ДЛЯ ДВОФАКТОРНОЇ АУТЕНТИФІКАЦІЇ НА ОСНОВІ БІОМЕТРИЧНИХ ПАРАМЕТРІВ
Безпалий О.Р., ***Науковий керівник: Висоцька О.О.*** **107**

ІНТЕГРАЦІЯ МЕСЕНДЖЕРА ТА GPT ДЛЯ РЕАЛІЗАЦІЇ ВІРТУАЛЬНОГО АСИСТЕНТА
Перевозкін А.С., ***Науковий керівник: Слюсарь І.І.*** **110**

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ВИЯВЛЕННЯ МАНІПУЛЯТИВНОГО КОНТЕНТУ
Колівушко Е., Лучка С., Матвійчук М., ***Науковий керівник: Ліп'яніна-Гончаренко Х.*** . **112**

МЕТОД ПРОГНОЗУВАННЯ ЗАТРИМОК ТА ПЕРЕВИТРАТ В ІНЖЕНЕРНИХ ПРОЕКТАХ ЗА ДОПОМОГОЮ МОДЕЛЕЙ ГЛИБИННОГО НАВЧАННЯ
Сулима Б.Я. **114**

ШТУЧНИЙ ІНТЕЛЕКТ - РУШІЙНА СИЛА ДЛЯ РОЗВИТКУ РІЗНОМАНІТНИХ ТЕХНОЛОГІЧНИХ СФЕР ДІЯЛЬНОСТІ ЛЮДИНИ
Черепаха М.О., ***Науковий керівник: Сердюк Н.М.*** **116**

Колівушко Едуард, здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Лучка Святослав, здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Матвійчук Микола, здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Ліп'яніна-Гончаренко Христина, канд. техн. наук, доцент, доцент кафедри інформаційно-обчислювальних систем та управління
Західноукраїнський національний університет, Україна

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ВИЯВЛЕННЯ МАНІПУЛЯТИВНОГО КОНТЕНТУ

З розвитком цифрових технологій та глобальної мережі Інтернет дезінформація і маніпулятивний контент стали невід'ємною частиною сучасного інформаційного простору. Швидке поширення фейкових новин і діпфейків негативно впливає на суспільну думку, політичні процеси та економічні відносини. За даними досліджень, 86% користувачів Інтернету щонайменше один раз стикалися з фейковими новинами [5]. Це підвищує необхідність розробки ефективних технологій для автоматичного виявлення такого контенту.

Основною складністю виявлення дезінформації є її постійна еволюція та складність у визначенні об'єктивної правдивості інформації. Додатковим викликом є те, що фейкові новини можуть бути дуже схожими на правдиві за своїм стилем і структурою, а діпфейки часто використовують високотехнологічні методи, що робить їх майже непомітними для людського ока [2]. Це вимагає впровадження інтелектуальних систем на базі машинного навчання для більш ефективного розпізнавання маніпуляцій.

Для автоматичного виявлення фейкових новин широко використовуються алгоритми машинного навчання, які здатні аналізувати текстові дані, виявляти закономірності та класифікувати контент на правдивий або маніпулятивний. Серед найбільш популярних методів можна виділити логістичну регресію, метод опорних векторів (SVM) та нейронні мережі. Наприклад, нейронні мережі показали високий рівень точності при класифікації текстів, аналізуючи синтаксичні та семантичні ознаки [3].

Текстові методи виявлення маніпулятивного контенту мають низку переваг, таких як швидка обробка та висока точність при виявленні патернів у тексті. Однак їхня ефективність знижується при роботі з неоднозначними чи контекстно залежними даними. Відеоаналітичні методи, такі як виявлення діпфейків, можуть виявити візуальні аномалії, проте вимагають значних обчислювальних ресурсів і складних моделей для точного розпізнавання підробленого контенту [1], [2].

Моделі машинного навчання показують різну продуктивність залежно від типу

контенту. Для текстових даних методи, засновані на класифікації, такі як логістична регресія чи LSTM, досягають точності понад 90% при аналізі новинних статей [3]. Для відеоконтенту глибинні нейронні мережі, такі як CNN або GAN, успішно виявляють діпфейки, проте їхня точність може коливатися залежно від якості відео і складності підробки, досягаючи 80-90% точності у складних випадках [1].

Для підвищення точності виявлення маніпулятивного контенту варто комбінувати різні підходи, що дозволить аналізувати і текстову, і відеоінформацію. Наприклад, інтеграція текстових моделей для попередньої фільтрації новин з подальшим аналізом відео чи зображень може значно підвищити ефективність. Крім того, використання гібридних моделей, що поєднують машинне та глибинне навчання, може забезпечити більш гнучке розпізнавання різних типів контенту [4].

У результаті дослідження було встановлено, що машинне навчання є потужним інструментом для виявлення маніпулятивного контенту, проте його ефективність залежить від типу даних та складності підробки. Комбінування текстових і відеоаналітичних методів може суттєво підвищити точність і зменшити ймовірність помилкових спрацьовувань.

У майбутньому можна очікувати подальшого розвитку інтелектуальних систем виявлення дезінформації, зокрема завдяки вдосконаленню глибинних нейронних мереж та їхньої інтеграції з іншими алгоритмами. Особливий інтерес представляє використання мультимодальних підходів, що дозволяють аналізувати текст, аудіо та відео одночасно, що зробить виявлення маніпуляцій ще більш точним та оперативним.

Список використаних джерел:

1. H. Nguyen, J. Yamagishi, and I. Echizen. "Capsule-forensics: Using capsule networks to detect forged images and videos." ICASSP 2019.
2. T. Korshunov and S. Marcel. "Deepfakes: A New Threat to Face Recognition? Assessment and Detection." IEEE International Conference on Biometrics, 2020.
3. D. Shu, H. Mahmud, and P. Resnick. "Fake News Detection on Social Media: A Data Mining Perspective." ACM SIGKDD, 2019.
4. C. Conroy, V. Rubin, and Y. Chen. "Automatic deception detection: Methods for finding fake news." Proceedings of the Association for Information Science and Technology, 2015.
5. Statista. "Share of internet users worldwide who encountered fake news as of January 2021." [Online]. Available: www.statista.com.