Міністерство освіти і науки України

«Київський національний університет імені Тараса Шевченка»

Факультет інформаційних технологій

Кафедра кібербезпеки та захисту інформації

ДОПУСТИТИ ДО ЗАХИСТУ:
завідувач кафедри кібербезпеки
та захисту інформації
_____ Н.В. Лукова-Чуйко
« » червня 2021р.

ПОЯСНЮВАЛЬНА ЗАПИСКА

дипломної роботи
бакалавра
(назва освітнього рівня)

галузь знань 12 Інформаційні технології
(шифр і назва галузі знань)

спеціальність 125 Кібербезпека
(код і назва спеціальності)

освітня програма Кібербезпека
(назва освітньої програми)

на тему: «Технологія Deepfake як загроза інформаційній безпеці»

Виконавець: студентка IV курсу, групи КБ-41

_____ **Купріна Лада Олександрівна**
(підпис) (прізвище ім'я по-батькові)

	Прізвище, ініціали	**Підпис**
Керівник	Браіловський М.М.	

Нормоконтроль	Даков С.Ю.	

Київ 2021

Міністерство освіти і науки України

«Київський національний університет імені Тараса Шевченка»

Факультет інформаційних технологій

Кафедра кібербезпеки та захисту інформації

ЗАТВЕРДЖЕНО:
завідувач кафедри кібербезпеки
та захисту інформації
_____ Н.В. Лукова-Чуйко
«10» жовтня 2020 р.

ЗАВДАННЯ
на виконання дипломної роботи

спеціальності 125 Кібербезпека
(код і назва спеціальності)

освітньої програми Кібербезпека
(назва освітньої програми)

Студентці **КБ-41** (група) **Купріній Ладі Олександрівні** (прізвище ім'я по-батькові)

Тема дипломної роботи Технологія Deepfake як загроза інформаційній безпеці

1. ПІДСТАВИ ДЛЯ ПРОВЕДЕННЯ РОБОТИ

Тема дипломної роботи затверджена на засіданні кафедри кібербезпеки та захисту інформації протокол №2 від 08.10.2020 р.

2. ВИХІДНІ ДАНІ ДЛЯ ПРОВЕДЕННЯ РОБІТ

Техніки виявлення підробок у мультимедійному контенті засобами машинного навчання.

3. ЗМІСТ РОЗРАХУНКОВО-ПОЯСНЮВАЛЬНОЇ ЗАПИСКИ

Проаналізувати методи створення та виявлення Deepfake із зазначенням проблемних моментів у їхньому функціонуванні, проаналізувати процес виявлення Deepfake із використанням різних технік, створених засобами машинного навчання, зібрати статистичні дані для порівняння роботи алгоритмів, виробити рекомендації.

4. ВИМОГИ ДО РЕЗУЛЬТАТІВ ВИКОНАННЯ РОБОТИ

Практична цінність полягає у підвищенні якості результатів виявлення відеопідробок Deepfake засобами машинного навчання.

5. ДАТА ВИДАЧІ ЗАВДАННЯ

Дата видачі завдання:12 жовтня 2020 року

Завдання видав		М.М. Браіловський
	(підпис)	(ініціали, прізвище)
Завдання прийняла до виконання		Л.О. Купріна
	(підпис)	(ініціали, прізвище)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів робіт	Строки виконання робіт (початок-кінець)	Відмітка про виконання
1	Уточнення постановки задачі	25.01.2021 – 29.01.2021	*виконано*
2	Аналіз літератури	30.01.2021 – 11.02.2021	*виконано*
3	Написання загального плану роботи	12.02.2021 – 15.02.2021	*виконано*
4	Дослідження методів створення Deepfake	16.02.2021 – 04.03.2021	*виконано*
5	Дослідження методів виявлення Deepfake	05.03.2021 – 21.03.2021	*виконано*
6	Порівняльний аналіз методів виявлення Deepfake засобами машинного навчання	22.03.2021 – 08.04.2021	*виконано*
7	Вироблення рекомендацій щодо захисту від Deepfake у корпоративному середовищі	09.04.2021 – 18.05.2021	*виконано*
8	Оформлення пояснювальної записки	19.05.2021 – 08.06.2021	*виконано*
9	Підготовка до захисту дипломної роботи	09.06.2021 – 16.06.2021	*виконано*

Завдання видав		М.М. Браіловський
	(підпис)	(ініціали, прізвище)
Завдання прийняла до виконання		Л.О. Купріна
	(підпис)	(ініціали, прізвище)

Термін подання дипломної роботи до ЕК 08 червня 2021 року

РЕФЕРАТ

Пояснювальна записка до дипломної роботи «Технологія Deepfake як загроза інформаційній безпеці» складається зі списку скорочень, вступу, основної частини, що містить 3 розділи, висновків, списку літератури та джерел. Загальний обсяг роботи – 70 сторінок. Робота містить 10 рисунків. Список використаних джерел включає 73 джерела.

Об'єкт дослідження – процес виявлення підробок у медійному контенті засобами машинного навчання.

Мета роботи – здійснення порівняльного аналізу сучасних методів виявлення Deepfake на основі засобів машинного навчання, в результаті якого будуть окреслені можливі майбутні тенденції як створення, так і виявлення підробок у мультимедійному контенті.

Предмет дослідження – методи та математичні алгоритми виявлення Deepfake засобами машинного навчання.

Метод дослідження – аналізу сучасних методів виявлення Deepfake.

Практична цінність отриманих результатів забезпечується підвищенням якості результатів виявлення відеопідробок Deepfake засобами машинного навчання, порівняно із аналогічними дослідженнями з даної тематики.

Ключові слова: Deepfake, штучний інтелект, глибинне навчання, заміна обличчя, генеративні мережі, дезінформація.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

CNN – Convolutional Neural Network;

RNN – Recurrent Neural Network;

IoT – Internet of Things (Інтернет речей);

GAN – Generative Adversarial Network;

ШІ – Штучний інтелект;

ІБ – Інформаційна безпека;

FP – False Positive;

TP – True Positive;

FN – False Negative;

TN – True Negative;

TTS – Text-to-Speech;

CM – Confusion Matrix;

ПЗ – Програмне забезпечення;

IT – Information Technology;

API – Application Programming Interface.

ВСТУП

Поширення у світі цифрових смарт-пристроїв, таких як мобільні телефони, планшети, ноутбуки та цифрові камери, призвело до експоненціального зростання мультимедійного вмісту (наприклад, зображень та відео) у кіберпросторі. Крім того, еволюція соціальних медіа за останнє десятиліття дозволила людям швидко обмінюватися мультимедійним контентом, що призвело до значного збільшення його обсягу та полегшення доступу до нього. У той же час ми стали свідками величезного прогресу в галузі машинного навчання завдяки впровадженню складних алгоритмів, які можуть легко маніпулювати мультимедійним вмістом для розповсюдження дезінформації в Інтернеті через платформи соціальних медіа. Враховуючи легкість створення та поширення неправдивої інформації, стає все важче виявити правду та довіряти інформації, і це може призвести до шкідливих наслідків.

В останні роки фальшиві новини стали проблемою, яка загрожує суспільному дискурсу, людському суспільству та демократії. Під фальшивими новинами розуміється вигаданий вміст у стилі новин, який сфабриковано для обману громадськості. Неправдива інформація швидко поширюється через соціальні мережі, де може вплинути на мільйони користувачів. На сьогоднішній день кожен п’ятий користувач Інтернету отримує свої новини через сервіс YouTube, що поступається лише Facebook. Це зростання популярності відео наголошує на необхідності інструментів для підтвердження справжності медіа та новин, оскільки нові технології дозволяють переконливо маніпулювати відеоматеріалами.

З огляду на легкість отримання та поширення дезінформації за допомогою платформ соціальних медіа, стає все важче знати, чому довіряти, що, серед іншого, призводить до шкідливих наслідків для прийняття обґрунтованих рішень. Дійсно, сьогодні ми живемо в епоху, яку деякі називають епохою "постправди", яка характеризується цифровою дезінформацією та інформаційною війною, яку ведуть

зловмисні актори, що проводять фальшиві інформаційні кампанії для маніпулювання громадською думкою.

Розвиток комп'ютерного зору та глибинного навчання призвів до появи технології, відомої як Deepfake. Deepfake (від англ. Deep learning – «глибоке навчання» і fake – «фальшивий») – це реалістична маніпуляція аудіо- і відеоматеріалами за допомогою штучного інтелекту. Знаменитості та політики – першочергові об'єкти Deepfake, оскільки вони мають величезну кількість відео та фотографій, доступних в Інтернеті. Таким чином, Deepfake-матеріали можуть використовуватися для розпалювання політичної чи релігійної напруженості між країнами, обману громадськості та впливу на результати виборів або створення нестабільності на фінансових ринках шляхом поширення неправдивої інформації.

Deepfake застосовує можливості штучного інтелекту для синтезу людського зображення: об'єднує кілька знімків, на яких людина зображена з різних ракурсів і з різним виразом обличчя, і робить з них відеопотік. Аналізуючи фотографії, спеціальний алгоритм навчається тому, як виглядає і може рухатися людина. При цьому працюють дві нейромережі. Перша з них генерує зображення, а друга відповідає за пошук відмінностей між ними і справжніми зразками. У разі якщо друга нейросеть виявляє підробку, зображення відправляється назад до першої для удосконалення.

Суттєвим фактором, який збільшує загрозу Deepfake, є обсяг, масштаб і витонченість задіяної технології, оскільки майже кожен, хто має комп'ютер, може виготовляти фальшиві відеоролики, які практично не відрізнити від автентичних. Deepfake працює за допомогою відкритих алгоритмів машинного навчання і бібліотек, що дозволяє досягти найвищої якості контенту. Нейромережа отримує зображення з бібліотеки і навчається за допомогою роликів на відеохостингу. Штучний інтелект тим часом зіставляє фрагменти вихідних портретів з тим, що є на відео, і в підсумку виходить правдоподібний матеріал. Підробки важко виявити, оскільки вони використовують реальні кадри, можуть мати аутентичне звучання та оптимізовані для швидкого розповсюдження в соціальних мережах. Таким чином, багато глядачів вважають, що відео, яке вони переглядають, справжнє.

Технологія Deepfake має величезний спектр застосувань, які можуть бути як позитивними, так і негативними, проте більшу частину часу вона використовується у зловмисних цілях. Неетичне використання технології Deepfake має шкідливі наслідки для нашого суспільства як у короткостроковій, так і в довгостроковій перспективі. Люди, які регулярно користуються соціальними мережами, мають величезний ризик наразитися на Deepfake. Однак належне використання цієї технології може принести багато позитивних результатів. Нижче докладно описано як негативне, так і позитивне застосування технології Deepfake.

Позитивне застосування: Хоча більшу частину часу ця технологія застосовується для зловмисних цілей, вона все ж має позитивне застосування також у кількох секторах. Створення Deepfake більше не залишається обмеженим лише експертами, тепер воно стає набагато простішим та доступнішим для будь-кого. У наш час конструктивне використання цієї технології широко зросло. Для створення нових витворів мистецтва, залучення аудиторії та надання їм унікального досвіду була використана ця технологія [1]. Нещодавно музей Далі в Санкт-Петербурзі, штат Флорида, США, надав шанс своїм відвідувачам познайомитись із Сальвадором Далі та більш інтерактивно взаємодіяти з його життям, щоб пізнати цю велику особистість за допомогою штучного інтелекту [2]. Технологія Deepfake використовується як в рекламних, так і в бізнес-цілях, а також для копіювання відомих творів мистецтва, таких як створення відео відомих робіт Мона Лізи за допомогою зображення [3]. Технологія Deepfake може заощадити величезні гроші та час кіноіндустрії, використовуючи можливості технології Deepfake для редагування відео тощо.

Негативне застосування: Deepfake та пов'язані з цим технології швидко розповсюджуються в реаліях сьогодення і часто використовуються для спричинення шкоди людині, особливо знаменитостям та політичним лідерам. Існують різні причини створення Deepfake-контенту, наприклад з метою пожартувати, але іноді це використовується для помсти, шантажу, викрадення особистості людини тощо. Найбільш зловмисним використанням Deepfake є експлуатація фото- та відеоматеріалів світових лідерів та політиків, створюючи їх фейкові відео, що може

бути великим ризиком для світового миру. Майже всі світові лідери, включаючи колишніх президентів США Барака Обаму та Дональда Трампа, політика США Ненсі Пелосі, канцлера Німеччини Ангелу Меркель та багато інших, зазнали шкоди через підроблені відео, і навіть засновник Facebook Марк Цукерберг стикався з подібним явищем [4].

Також Deepfake можуть становити загрозу не лише для громадських діячів, а й для простих людей. Так само останнім часом китайський додаток Zao став вірусним, оскільки менш кваліфіковані користувачі можуть накласти свої обличчя на тіла кінозірок і вставити себе у відомі фільми та телевізійні кліпи [5]. Ці форми фальсифікації створюють величезну загрозу порушення конфіденційності особистості та впливають на багато аспектів людського життя.

За останні кілька років технологія клонування голосу також стала дуже витонченою. На відміну від відео-Deepfake, виявленню аудіо-Deepfake приділяється менше уваги. Однак клонування голосу є не тільки загрозою як для автоматизованих систем розпізнавання особистості за голосом, так і для систем з голосовим управлінням, які широко використовуються у сфері Інтернету речей (IoT) [6].

Клонування голосу має величезний потенціал для знищення довіри громадськості та надання можливості злочинцям маніпулювати діловими відносинами або приватними телефонними дзвінками. Наприклад, нещодавно було зареєстровано три випадки, коли грабіжники банків використовували голосове клонування виступу керівника компанії, щоб обдурити своїх підлеглих та перевести сотні тисяч доларів на секретний рахунок. Deepfake із маніпуляцією голосом була використана для шахрайства, яке обійшлося генеральному директору компанії у $243000 [7]. Очікується, що інтеграція голосового клонування в deepfake зробить виявлення deepfake ще більш складним. Тому важливо, щоб, на відміну від сучасних підходів, що зосереджуються лише на виявленні маніпуляцій з відеоматеріалами, аудіо-підробки також були досліджені.

Із вище наведеного випливає, що розробка технологій, які можуть автоматично визначати та оцінювати автентичність цифрових візуальних носіїв інформації, є необхідною. У відповідь на все більш досконалий та реалістичний

маніпульований вміст дослідницьке співтовариство докладає великих зусиль для розробки вдосконалених методів виявлення маніпуляцій із відео- та аудіоматеріалами. Проблемою розробки алгоритмів виявлення Deepfake займалися у своїх дослідженнях такі закордонні науковці: Y. Mirsky, W. Lee, T. Nguyen, M. Snoeck, B. Baesens, C. Bravo, O. Caelen, T. Eliassi-Rad, S. K. Jha, K. K. Tharakunnel, J. C. Westland, R. J. Wang, Y. Huang, Y. Liu, N.M. Adams, M. Nawaz та інші. Серед вітчизняних науковців таких досліджень не проводилось.

Однак, жодна із наявних на сьогодні публікацій не охоплює створення та виявлення Deepfake як із маніпуляцією відеоматеріалами, так і із клонуванням голосу. Більшість існуючих праць зосереджуються лише на аналізі підробок у зображеннях та відео. Тому метою даної дипломної роботи детальний аналіз методів створення аудіо- та відео-Deepfake, розгляд загроз, які вони становлять для інформаційного суспільства, та наведення методів виявлення таких підробок.

Для досягнення даної мети необхідно вирішити наступні задачі:

1. провести аналіз та класифікацію методів створення Deepfake;
2. провести аналітичний огляд наявних методів виявлення відео- та аудіо-Deepfake;
3. скласти порівняльну характеристику наявних методів виявлення Deepfake із зазначенням проблемних моментів у їхньому функціонуванні;
4. провести дослідну експлуатацію та оцінку якості моделей виявлення Deepfake;
5. виробити рекомендації щодо захисту від Deepfake у корпоративному середовищі.

Об’єктом дослідження є процес виявлення підробок у відео- та аудіоматеріалах.

Предметом дослідження є методи виявлення відео- та аудіо-Deepfake засобами машинного навчання.

Для досягнення мети дипломної роботи були використані наступні методи дослідження:

1. у розділі 1 було проаналізовано еволюцію технології Deepfake, проведено аналіз методів виявленя Deepfake та проблемних моментів у їхньому функціонуванні;

2. у розділі 2 було здійснено аналіз основних методів виявлення Deepfake, наявних на сьогоднішній день, зазначено їхні переваги та недоліки. В результаті дослідження було вирішено дослідити ефективність методів виявлення, що базуються на згорткових нейронних мережах;

3. у розділі 3 було протестовано чотири згорткові нейронні мережі з метою перевірити ефективність виявлення підробок у відео-матеріалах; було отримано результати, на основі яких зроблено висновки щодо майбутніх напрямів розвитку методів виявлення Deepfake.

Наукова новизна дипломної роботи полягає у аналітичному огляді та систематизації методів виявлення Deepfake із зазначенням проблемних моментів у їхньому функціонуванні, а також дослідженні та обгрунтованому запропонуванні найкращого методу для виявлення підробок у відеоматеріалах.

Практичне значення отриманих результатів забезпечується підвищенням якості результатів виявлення відео- та аудіо-Deepfake засобами машинного навчання, порівняно із аналогічними дослідженнями з даної тематики.

Основні результати дипломної роботи доповідалися та обговорювалися на VII Міжнародній науково-практичній конференції «Information Technology and Interactions» (IT&I-2020) (Київ, 2020); IV Міжнародній науково-практичній конференції «Проблеми кібербезпеки інформаційно-телекомунікаційних систем» (Київ, 2021).

РОЗДІЛ 1
АНАЛІЗ МЕТОДІВ СТВОРЕННЯ DEEPFAKE

1.1 Еволюція Deepfake

Найбільш ранній приклад маніпуляції мультимедійним контентом стався ще у 1860 р., коли портрет південного політика Джона Калхуна був майстерно викривлений шляхом заміни його голови на голову президента США Авраама Лінкольна [8]. Зазвичай така маніпуляція здійснюється шляхом додавання, видалення та реплікації об'єктів усередині або між двома зображеннями. Потім застосовуються відповідні кроки пост-обробки, такі як масштабування, обертання та регулювання кольорів, щоб поліпшити візуальний вигляд, масштаб та узгодженість перспективи. Окрім цих традиційних методів маніпуляцій, вдосконалення комп'ютерної графіки та технік глибинного навчання призвело до появи різноманітних автоматизованих підходів до цифрових маніпуляцій із кращою семантичною послідовністю. Останній тренд передбачає синтез відео з нуля за допомогою автокодерів для різних застосувань і, більш конкретно, фотореалістичну генерацію людського обличчя на основі будь-якого атрибута [9].

Ще однією поширеною маніпуляцією, яка називається «дрібні підробки» або «дешеві підробки», є аудіо-візуальні маніпуляції, створені з використанням більш дешевого та доступного програмного забезпечення. Дрібні підробки передбачають основний монтаж відео з використанням уповільнення, прискорення, вирізання та вибіркового зрощування існуючих кадрів, що може змінити весь контекст наданої інформації. У травні 2019 року відео з інтерв’ю спікера США Ненсі Пелосі було відбірково відредаговано, щоб виставити її таким чином, наче вона нерозбірливо вимовляла слова і була напідпитку або розгублена. Відео було поширено у Facebook і за 48 годин набрало понад 2,2 мільйона переглядів. [10]

Відео-маніпуляції для індустрії розваг, зокрема у виробництві фільмів, виконуються вже десятки років. Раннім помітним академічним проектом стала

програма Video Rewrite Program [11], призначена для застосування у дублюванні фільмів, опублікована в 1997 р. Це було перше програмне забезпечення, що використовувалося для автоматичної реанімації рухів обличчя в існуючому відео на іншу звукову доріжку, і вона досягла напрочуд переконливих результатів. Перший справжній Deepfake з'явився в Інтернеті у вересні 2017 року, коли користувач Reddit на ім'я «deepfake» опублікував серію згенерованих відео відомих актрис із обличчями, накладеними на порнографічні відеоматеріали [12]. Після цього випадку Deepfake стали широко відомими у Інтернет-суспільстві. Рисунок 1.1 показує еволюцію Deepfake протягом багатьох років.

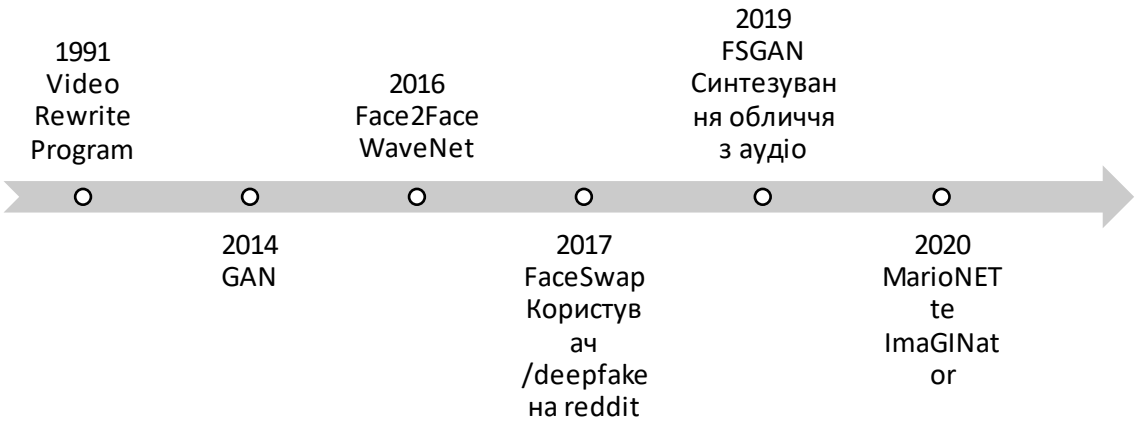

Рисунок 1.1 – Еволюція Deepfake

Сьогодні технології та додатки для ствоення Deepfake, такі як FakeApp, FaceSwap та ZAO, є легкодоступними для широкого загалу, і користувачі, які не мають досвіду у сфері комп'ютерної інженерії, можуть створити фейкове відео за лічені секунди. Більше того, проекти з відкритим кодом на GitHub, такі як DeepFaceLab та відповідні докладні інструкції до них, є легкодоступними на YouTube. Сучасними академічними проектами, що сприяють розвитку технології Deepfake, є Face2Face та Synthesizing Obama, опубліковані у 2016 та 2017 роках відповідно. Face2Face фіксує вираз обличчя людини в реальному часі, після чого накладає цей вираз на цільове обличчя. Synthesizing Obama – це програма, яка використовується для модифікації рухів рота людини на відео, щоб зобразити наче вона вимовляє слова, що містяться в довільній аудіодоріжці.

Окрім візуальних маніпуляцій, існують також аудіо-Deepfake – це нова форма кібератаки, яка може завдати серйозних збитків людям завдяки високотехнологічним методам синтезу мовлення, представленим насамперед у додатках WaveNet, Tacotron та deep voice1 [13]. Фінансові афери із використанням підробок голосу значно зросли у 2019 році завдяки прогресу в технології синтезу мовлення. У серпні 2019 року керівник європейської компанії, обдурений аудіо-Deepfake, здійснив переказ на рахунок шахраям у розмірі 243 000 доларів США [7]. Для клонування голосу жертви було використано програмне забезпечення, що імітує голос шляхом тренування алгоритмів машинного навчання за допомогою аудіозаписів, отриманих з мережі Інтернет. Якщо такі методи можна використовувати для імітації голосу вищого урядовця чи військового лідера та застосовувати їх у великому масштабі, це може мати серйозні наслідки для національної безпеки.

Список інших доступних додатків, програмного забезпечення та проектів із відкритим кодом для створення аудіо- та відео-Deepfake наведено в таблиці 1.1.

Таблиця 1.1

Огляд програмного забезпечення та Інтернет-платформ для генерації аудіовізуальних підробок

Інструмент	**Тип**	**Розробник / посилання**	**Техніка**
Маніпуляція атрибутами обличчя			
FaceApp	Мобільний додаток	FaceApp Inc	Згорткові нейронні мережі
Adobe	Комерційне ПЗ	Adobe	Згроткові нейронні мережі + фільтри
Rosebud	Комерційний веб-додаток	www.rosebud.ai/	Пропрієтарна з використанням штучного інтелекту (ШІ)
Заміна обличчя			
Reflect	Мобільний додаток	Neocortext, Inc	Пропрієтарна

Продовження таблиці 1.1

Impressions	Мобільний додаток	Synthesized Media, Inc	Пропрієтарна
FaceSwap	Open source реалізація	github.com/shaoanlu/faceswap-GAN	Пара кодер-декодер
Faceswapweb	Комерційний веб-додаток	www.faceswapweb.com	Генеративно-змагальна мережа
Клонування голосу			
Overdub	Комерційний веб-додаток	descript.com/overdub	Пропрієтарна з використанням ШІ
Respeecher	Комерційний веб-додаток	respeecher.com	Алгоритми обробки цифрових сигналів + техніки глибокого генеративного моделювання
ResembleAI	Мобільний додаток	Zoezi AB	Пропрієтарна

Зважаючи на доступність технології Deepfake для пересічних користувачів і на велику кількість відповідних програм та інтернет-платформ, даний вид мережевої активності зростає дуже стрімко. Як можна побачити на рисунку 1.2, лише за 2020 рік кількість Deepfake, виявлених у мережі Інтернет, зросла більш ніж удвічі [14].

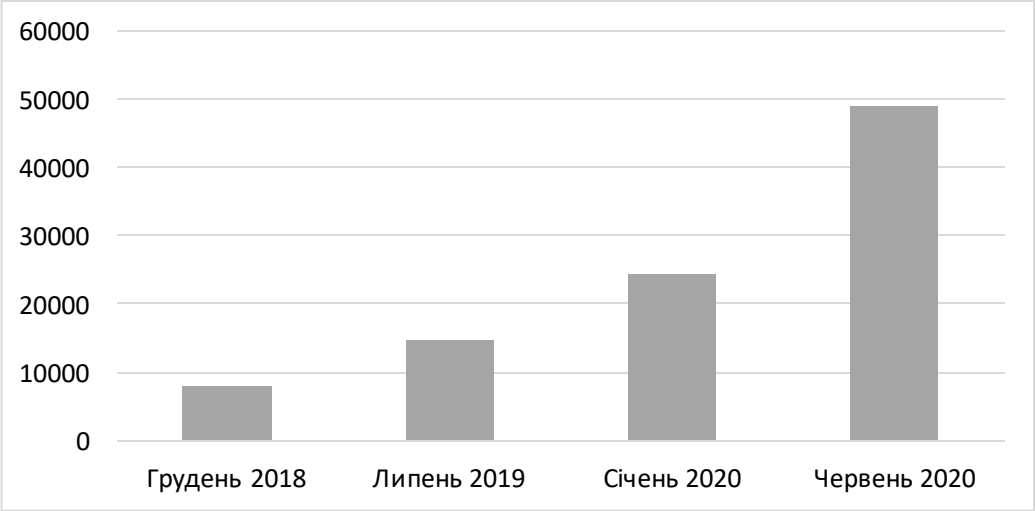

Рисунок 1.2 – Кількість Deepfake, виявлених у мережі Інтернет

1.2. Огляд методів створення Deepfake

1.2.1 Заміна обличчя

Традиційні техніки заміни обличчя зазвичай виконуються у три кроки.

1. Спочатку відбувається виявлення обличчя у вихідних зображеннях, потім з бібліотеки обличь вибирається таке зображення обличчя-кандидата, яке схоже на вихідне обличчя.

2. Далі, відбувається заміна очей, носа та рота обличчя, а також додаткове регулювання освітлення та кольору зображення обличчя кандидата відповідно до зовнішнього вигляду вихідних зображень, таким чином поєднуються обидва обличчя.

3. Нарешті, третій крок класифікує отримане зображення, обчислюючи відстань відповідності в області перекриття.

Цей підхід може давати хороші результати за певних умов, але має два основних обмеження. По-перше, він повністю замінює вихідне обличчя цільовим, і вирази вихідного обличчя втрачаються. По-друге, синтезований результат є дуже жорстким, а замінене обличчя виглядає неприродно, тобто для отримання хороших результатів потрібно ретельно підбирати відповідну позу.

Останнім часом підходи, засновані на глибокому навчанні, стали популярними для синтезування зображень завдяки їхнім реалістичним результатам. У той же час Deepfake показали, як ці підходи можна застосовувати за допомогою автоматизованих цифрових мультимедійних маніпуляцій. У 2017 році перший відео-Deepfake, який з’явився в Інтернеті, було створено за допомогою підходу заміни обличчя. Цей підхід використовував нейронну мережу, щоб накласти обличчя жертви на чужі риси, при цьому зберігаючи оригінальний вираз обличчя. Із часом програми для заміни обличчя, такі як FakeApp та FaceSwap, спростили та прискорили задачу створення Deepfake з більш переконливими результатами.

Ці підходи зазвичай використовують дві пари кодер-декодер. Кодер використовують для вилучення латентних рис обличчя із зображення, а потім

декодер використовують для реконструкції обличчя. Щоб поміняти обличчя на вихідному та цільовому зображенні потрібні дві пари кодера та декодера, де кожен кодер спочатку навчається на вихідному, а потім на цільовому зображенні. Після завершення навчання декодери міняються місцями, так що оригінальний кодер вихідного зображення та декодер цільового зображення використовуються для регенерації цільового зображення з особливостями вихідного зображення. Отримане зображення являє собою вихідне обличчя, накладене на цільове, при цьому зберігаючи вираз цільового обличчя.

Нещодавно запущені додатки ZAO, REFACE та FakeApp набирають популярність завдяки своїй ефективності у створенні реалістичних Deepfake на основі заміни обличь. FakeApp дозволяє вибірково модифікувати частини обличчя. ZAO та REFACE останнім часом стали вірусними, оскільки менш досвідчені користувачі можуть накласти своє обличчя на обличчя відомих людей і таким чином вбудуватися у відомі фільми та телевізійні кліпи. Існує багато загальнодоступних реалізацій технології заміни обличчя за допомогою глибоких нейронних мереж, таких як FaceSwap, DFaker, DeepFaceLab, DeepFake-tf та FaceSwapGAN.

Донедавна більшість досліджень зосереджувались на прогресі в технології заміни облич, використовуючи модель на основі GAN (Generative Adversarial Network, Генеративна Змагальна Мережа) [15]. Їх головним принципом є, по суті, конкуренція генератора та дискримінатора. У системі GAN одна з мереж (генеративна або G-мережа) генерує зображення, а інша (дискримінантна або D-мережа) намагається відрізнити правильні вибірки від неправильних. Використовуючи набір прихованих просторових змінних, генеративна мережа намагається створити новий шаблон шляхом змішування кількох вихідних зразків. Дискримінантна мережа вчиться розрізняти справжні та підроблені зразки, а результати розмежування подаються на вхід генеративної мережі, щоб вона могла вибрати найкращий набір прихованих параметрів, а дискримінантна мережа більше не могла відрізняти справжні зразки від підроблених.

Цей підхід перевершує декілька існуючіх методів, заснованих на парі кодера-декодера, оскільки вони працюють, не проходячи явної підготовки щодо конкретних

зображень. Більше того, ітераційний характер робить їх добре придатними для маніпуляцій з обличчями, таких як створення реалістичних підроблених зображень облич.

Однак нещодавно був запропонований підхід, заснований на конволюційній (або згортковій) нейронній мережі (CNN, Convolutional Neural Network), який передавав семантичний вміст, наприклад, позу, вираз обличчя та умови освітлення вихідного зображення, щоб відтворити цей стиль в іншому зображенні. Такий підхід використовує функцію втрати, яка є зваженою комбінацією втрати стилю, втрати вмісту, втрати світла та загальної регуляції варіацій. Цей метод генерує більш реалістичні підробки порівняно з традиційними підходами, однак вимагає великої кількості навчальних даних. Більше того, навчена модель може бути використана для одночасного перетворення лише одного зображення.

Оклюзії обличчя – це проблеми, які завжди складно вирішити при заміні обличчя. У багатьох випадках лицьова область у вихідному або цільовому зображенні може бути частково закрита волоссям, окулярами, рукою чи якимись іншими предметами. Це призводить до появи візуальних артефактів та невідповідностей отриманого зображення. На сьогодні існує фреймворк FaceShifter, що генерує замінене обличчя з високою точністю і зберігає цільові атрибути, такі як поза, вираз та оклюзія.

1.2.2 Синхронізація рухів губ

Підхід до синхронізації рухів губ (Lipsync) передбачає синтез цільового відео таким чином, щоб область рота в маніпульованому відео відповідала довільній аудіодоріжці. Ключовим аспектом візуального синтезу мовлення є рухи та зовнішній вигляд нижньої частини рота та прилеглої до нього області. Щоб передавати повідомлення більш ефективно і природно, важливо генерувати правильні рухи губ разом із виразом обличчя. З наукової точки зору, синхронізація губ має багато застосувань в індустрії розваг, наприклад, створення аудіокерованих фотореалістичних цифрових персонажів у фільмах чи іграх, голосових ботів та

дубляж фільмів іноземними мовами. Більше того, це також може допомогти людям з вадами слуху зрозуміти сценарій, читаючи губи з відео.

Існуючі роботи з синхронізації руху губ [16, 17] вимагають повторного виділення кадрів із відеозапису чи транскрипції разом із цільовими емоціями для синтезу руху губ. Ці підходи обмежуються певним емоційним станом і погано узагальнюють невидимі обличч. Однак моделі глибокого навчання здатні вивчати та прогнозувати рухи за допомогою аудіофункцій. У роботі [18] запропоновано підхід до створення реалістичного відео із синхронізацією рухів губ, використовуючи цільове відео та довільний аудіокліп. Модель, що базується на рекуррентній нейронній мережі (RNN, Recurrent Neural Network), була використана для вивчення картографування між звуковими функціями та формою рота для кожного кадра, а пізніше використовувала повторний вибір кадру для заповнення текстури навколо рота на основі орієнтирів.

Цей синтез проводився на нижніх ділянках обличчя, тобто роті, підборідді, носі та щоках, і застосував низку етапів пост-обробки, таких як згладжування розташування щелепи та зміна таймингів на відео, щоб вирівняти голосові паузи, або рухи голови при розмові, щоб створити відео, яке виглядає більш природним та реалістичним.

Модель Speech2Vid [19] взяла аудіокліп та статичне зображення цільового суб’єкта як вхідні дані та створила відео, яке синхронізується з аудіокліпом. Ця модель використовувала коефіцієнти Mel Frequency Cepstral Coefficients (MFCC), витягнуті з вихідного аудіо, і подавала їх у систему кодер-декодер на основі CNN. Для пост-обробки була використана окрема CNN для розмивання кадрів та посилення різкості, щоб зберегти якість візуального вмісту. Ця модель добре узагальнює невидимі обличчя і тому не потребує перепідготовки для нових ідентичностей. Однак ця робота не здатна синтезувати емоційну міміку.

У роботі [20] використовувалася часова GAN, що складається з RNN, для генерації фотореалістичного відео безпосередньо з нерухомого зображення та мовного сигналу. Отримане відео включало синхронізовані рухи губ, кліпання очима та природний вираз обличчя, не покладаючись на ручно виконані аудіо-

візуальні функції. Для контроля якості кадру, аудіовізуальної синхронізації та загальної якості відео використовувались кілька дискримінаторів. Ця модель може генерувати синхронізацію губ для будь-якої людини в режимі реального часу.

У роботі [21] для вивчення аудіо-візуального зображення було використано метод змагального навчання. Кодер мовлення був навчений проектувати як звукові, так і візуальні репрезентації в один і той же латентний простір. Перевага використання такого подання полягала в тому, що як аудіо, так і відео могли служити джерелом мовної інформації під час процесу генерації. В результаті вдалося створити реалістичні послідовності розмовляючих облич на довільній особистості за допомогою синхронізованого руху губ.

У роботі [22] представлена система Vdub, яка фіксує високоякісну тривимірну модель обличчя як джерела, так і цільового актора. Обчислювана модель обличчя була використана для фотореалістичної реконструкції 3D-моделі рота, яка застосовувалась до цільового актора. Для кращого вирівнювання синтезованого візуального вмісту зі звуком було проведено аналіз аудіоканалу. Цей підхід краще обробляє текстуру зубів, проте він не може синтезувати високоякісну внутрішню область рота.

У роботі [23] пропонується метод особистого перекладу LipGAN, який синтезує відео із обличчя будь-якої людини, використовуючи лише одне наявне зображення та аудіо сегмент. LipGAN складається з генераторної мережі для синтеза портретних відеокадрів із модифікованою областю рота та щелепи із заданих звукових та цільових кадрів, а також використовує дискримінантну мережу, щоб вирішити, чи є синтезоване обличчя синхронізоване з даним звуком. Цей підхід не може забезпечити часову послідовність синтезованого вмісту, оскільки в отриманому відео можна спостерігати розмитість і тремтіння.

Нещодавно у роботі [24] було запропоновано модель wav2lip, яка може точно синхронізувати рух губ у відеозаписі із заданим аудіокліпом. Цей підхід використовує попередньо навчений дискримінатор синхронізації рухів губ, який проходить подальшу підготовку на шумно сформованих відеозаписах за відсутності генератора. Ця модель використовує кілька послідовних кадрів замість одного кадру

в дискримінаторі, таким чином підвищуючи візуальну якість і враховуючи часову кореляцію.

Останні підходи можуть синтезувати фотореалістичні фальшиві відеоролики з мови (speech-to-video) або тексту (text-to-video) з переконливими результатами. Ці методи можуть відредагувати існуюче відео людини згідно до бажаної промови шляхом модифікації рухів рота та мовлення відповідно. Ці підходи більше зосереджені на синхронізації рухів губ шляхом синтезу лише області навколо рота.

1.2.3 Реконструкція обличчя

«Ляльковод», також відомий як реконструкція обличчя – це ще одна поширена варіація Deepfake, яка маніпулює мімікою людини, наприклад, передаючи вирази обличчя, рухи очей і голови на вихідне відео. Реконструкція обличчя має на меті деформувати рух рота людини, щоб створити сфабрикований вміст. Цей метод має різні застосування, наприклад зміна виразів обличчя та руху рота учасника багатомовної відеоконференції в Інтернеті при перекладі промови на іншу мову, редагування голови та виразів обличчя актора в системах постпродукції кіноіндустрії або створення фотореалістичної анімації для фільмів, ігор тощо.

Спочатку для реконструкції обличчя пропонували підходи, засновані на 3D-моделюванні обличчя, завдяки їх здатності точно фіксувати геометрію та рух, а також для поліпшення фотореалізму на реконструйованих обличчях. У роботах [25, 26] представлено перший метод передачі міміки в реальному часі від актора до цільової людини. Датчик RGB-D був використаний для відстеження та реконструкції тривимірної моделі актора джерела та цілі.

Face2Face – це вдосконалена форма техніки реконструкції обличчя, представлена в [27]. Цей метод працює в режимі реального часу і може змінювати рухи обличчя, наприклад, на відео YouTube, використовуючи стандартну веб-камеру. Підхід до реконструкції тривимірної моделі поєднувався з методами відтворення зображень для отримання результату. Це створює переконливий і миттєвий рендеринг цільового актора за відносно простої домашньої настройки. Ця

робота була додатково розширена для керування мімікою людини у цільовому відео на основі інтуїтивних жестів руками за допомогою інерціального вимірювального блоку.

GAN були успішно застосовані для реконструкції обличчя завдяки їх здатності генерувати фотореалістичні зображення. Pix2pixHD [28] створює зображення з високою роздільною здатністю з кращою точністю, поєднуючи багатомасштабну умовну архітектуру GAN. У роботі [29] запропоновано підхід, що дозволяє повною мірою реанімувати портретні відео, наприклад, змінюючи позу голови, погляд очей та моргання, а не просто модифікуючи вираз обличчя цільової особистості. Таким чином це дає фотореалістичні результати дубляжу. Спочатку застосовувався підхід до реконструкції обличчя для отримання параметричного зображення обличчя та інформації про освітленість з кожного відеокадру для отримання синтетичного відтворення цільової особистості. Потім це відтворення подавалося в мережу на основі cGAN (Conditional Generative Andversarial Network) для прогнозування синтетичного рендерингу у фотореалістичні відеокадри. Цей підхід вимагає тренування на відео із зображенням цільової особистості.

Нещодавно було запропоновано декілька підходів до реконструкції обличчя, для досягнення реконструкції з використанням кількох або навіть одного вихідного зображення. У роботі [30] було запропоновано модель X2face, що використовує різні способи, такі як керування кадрами, орієнтири на обличчі або звук для передачі пози та виразу обличчя на цільове обличчя. X2face використовував дві мережі кодера-декодера: мережу вбудовування та мережу керування. Мережа вбудовування вивчає представлення обличчя з вихідного кадру, а керуюча мережа вивчає інформацію про позу, рухи та вирази обличчя.

У роботі [31] представлено підхід до навчання, де мережу спочатку навчали на декількох особистостях, а потім допрацьовували на цільовій особистості. По-перше, кодування цільової особистості було отримано шляхом усереднення виразів цілі та пов’язаних орієнтирів з різних кадрів. Потім було використано GAN pix2pixHD [28] для генерування цільової ідентифікації з використанням вихідних орієнтирів як вхідних даних та кодування особистості через шари AdaIN. Цей підхід добре працює

під похилими кутами і безпосередньо передає вираз, не вимагаючи проміжного граничного латентного простору чи інтерполяційної карти. У роботі [32] запропоновано структуру на основі автокодера, щоб дізнатися приховане відображення вигляду обличчя цілі та форми обличчя джерела. Ці функції використовувались як вхідні дані до залишкових блоків SPADE для завдання реконструкції обличчя, які зберігали просторову інформацію та багатоконтурно об’єднували карту об’єктів із декодера реконструкції обличчя. Цей підхід може краще впоратися із великими змінами пози та перебільшеними діями обличчя. У FaRGAN [33] функції, які можна дізнатись із шарів згортки, використовувались як вхідні дані до модуля SPADE замість використання багатомасштабних масок орієнтиру, як у [32]. Зазвичай навчання з кількома знімками не дозволяє повністю зберегти ідентичність джерела у згенерованих результатах у випадках, коли між еталонним та цільовим зображенням існує велика різниця пози. MarioNETte [34] було запропоновано пом'якшити витік ідентичності шляхом використання блоку уваги та вирівнювання цільових функцій. Це допомогло моделі краще враховувати варіації між структурами обличчя. Нарешті, ідентичність було збережено за допомогою нового трансформаторного орієнтира, на який вплинула модель обличчя 3DMM [35].

1.2.4 Синтез обличчя та маніпуляція атрибутами

Маніпуляції з обличчям можна розділити на дві категорії: генерація обличчя та редагування атрибутів обличчя. Генерація обличчя передбачає синтез фотореалістичних образів людського обличчя, яких не існує в реальному житті. На відміну від цього, редагування атрибутів обличчя передбачає зміну зовнішнього вигляду обличчя існуючої вибірки шляхом модифікації специфічної для атрибута області, залишаючи інші регіони обличчя незмінними. Редагування атрибутів обличчя включає видалення / додавання окулярів, зміну точки зору, ретушування шкіри (наприклад, розгладження шкіри, видалення шрамів та мінімізацію зморшок) і навіть деякі модифікації вищого рівня, такі як вік, стать тощо. Люди все частіше

використовують комерційно доступні засоби редагування обличчя на основі ШІ та мобільні програми, такі як FaceApp, які автоматично змінюють зовнішній вигляд вихідного зображення.

Величезний розвиток глибоких генеративних моделей зробив їх широко прийнятими інструментами для синтезу та редагування зображень. Генеративні моделі глибокого навчання, такі як GAN [15] та VAE [36], успішно використовуються для створення фотореалістичних фальшивих зображень людського обличчя. Метою синтезу обличчя є створення неіснуючих, але реалістичних облич. Синтез обличчя представив широкий спектр корисних програм, таких як автоматичне створення персонажів для відеоігор та 3D-моделювання обличчя.

Синтез обличчя на основі ШІ також може бути використаний у зловмисних цілях, наприклад, для синтезування фотореалістичного фальшивого зображення для облікових записів соціальних мереж із фальшивою цифровою ідентичністю для поширення дезінформації. Запропоновано кілька підходів для створення реалістичних образів обличчя, які люди не в змозі розпізнати, чи є вони справжніми чи синтезованими. На рис. 1.3 показані синтетичні зображення обличчя та покращення їх якості між 2014 і 2019 роками, які майже неможливо відрізнити від реальних фото.

Рисунок 1.3 – Еволюція якості синтезованих облич

З моменту появи GAN у 2014 р. докладено значних зусиль для поліпшення якості синтезованих зображень. Зображення, створені з використанням першої моделі GAN, мали низьку роздільну здатність і не були дуже переконливі. DCGAN [37] був першим підходом, який ввів у генератор шар деконволюції для заміни повністю зв’язаного шару, що дозволило досягти кращих показників при генерації синтетичних зображень.

У роботі [38] запропоновано CoGAN на основі VAE для вивчення спільного розподілу дводоменних зображень. Ця модель навчила декілька GAN, а не одну, і кожна відповідала за синтез зображень в одному домені. Розмір сформованих зображень все ще залишався відносно невеликим, наприклад 64 × 64 або 128 × 128 пікселів.

Генерація зображень із високою роздільною здатністю була обмежена раніше через обмеження пам'яті. У роботі [39] представлена ProGAN, навчальна методологія для GAN, яка використовувала адаптивне поступове збільшення роздільної здатності, залежно від поточної вихідної роздільної здатності, додаючи рівні до мереж під час навчального процесу. StyleGAN [40] – це вдосконалена версія ProGAN. Замість зіставлення прихованого коду z з роздільною здатністю була використана мережа відображення, яка навчилася відображати вхідний прихований вектор (Z) на проміжний прихований вектор (W), який контролював різні візуальні особливості. Поліпшення полягає в тому, що проміжний прихований вектор вільний від будь-яких певних обмежень розподілу, і це зменшує кореляцію між ознаками (роз'єднання).

Управління рівнями генераторної мережі здійснюється за допомогою операції AdaIN, яка допомагає визначити особливості вихідного рівня. StyleGAN досягла найсучаснішої високої роздільної здатності у створених зображеннях, тобто 1024 × 1024, з дрібними деталями. StyleGAN2 [41] ще більше покращила сприйману якість зображення, видаливши небажані артефакти, такі як зміна напрямку погляду та вирівнювання зубів. Таким чином, створені зображення були фотореалістичними та дуже близькими до реальних зображень.

Нещодавно було запропоновано кілька підходів, заснованих на GAN, для редагування атрибутів обличчя, таких як колір шкіри, зачіска, вік та стать шляхом додавання / видалення окулярів та зміни виразу обличчя тощо. Під час цієї маніпуляції GAN бере вихідне зображення обличчя та генерує відредаговане зображення обличчя із заданим атрибутом.

1.2.5 Синтез голосу

Звукова маніпуляція, синтезована за допомогою штучного інтелекту – це тип Deepfake, який може клонувати голос людини і зображати, наче вона говорить щось обурливе, дезінформуюче, чого насправді ніколи не говорила. Нещодавні досягнення в алгоритмах синтезу мовлення та клонування голосу показали потенціал для створення реалістичних фальшивих голосів, які майже неможливо відрізнити від справжньої мови. Ці алгоритми можуть генерувати синтетичне мовлення, яке звучить реалістично, на основі тексту або висловлювання цілі, з надзвичайно переконливими результатами [42].

Синтетичний голос широко пристосований для розробки різних додатків, таких як автоматизований дубляж для телевізора та кіно, чат-боти, ШІ-помічники, читачі тексту та персоналізовані синтетичні голоси для людей з обмеженими можливостями. Окрім цього, синтетичні/підроблені голоси стали підвищеною загрозою для голосових біометричних систем і використовуються у зловмисних цілях, таких як політичні вигоди, фальшиві новини, шахрайства тощо.

Більш складний синтез звуку може поєднувати потужність ШІ та ручне редагування. Наприклад, моделі синтезу голосу на основі нейронних мереж, такі як Google Tacotron, Wavenet або AdobeVoco [43], можуть створити реалістичні та переконливі фальшиві голоси, що нагадують голос жертви. Пізніше програмне забезпечення для редагування звуку, наприклад Audacity, може використовуватися для поєднання різних фрагментів оригінального та синтезованого аудіо для створення більш переконливих аудіодоріжок.

Уособлення на основі ШІ не обмежується лише візуальним вмістом; нещодавні досягнення у синтезованих голосах допомагають створювати надзвичайно реалістичні відеоролики deepfake. Останні події в синтезі мовлення показали їхній потенціал для створення реалістичних та природних аудіофайлів, що представляють реальні загрози суспільству. Поєднання синтетичного аудіовмісту з візуальними маніпуляціями може зробити глибокі відеофайли більш переконливими та посилити їх шкідливий вплив. Однак до цих пір у цих синтезованих промовах відсутні деякі аспекти якості голосу, такі як виразність, шорсткість, задишка, стрес та емоції тощо, специфічні для цільової особистості.

TTS (Text-to-Speech) – це технологія, якій вже багато десятиліть, і вона здатна синтезувати природне мовлення людини із заданого вхідного тексту. Таким чином, це дозволяє використовувати голос для кращої взаємодії людини з комп'ютером. VC (Voice Conversion) – це ще одна методика, яка модифікує вихідну аудіодоріжку із промовою, щоб вона звучала так, як голос цільового мовця, зберігаючи при цьому мовний зміст незмінним [44]. Останні ініціативи щодо синтезу мовлення викликають більше занепокоєння [45]. Загалом, важливі зрушення у синтезі мовлення були зроблені за допомогою методів конкатенації мовлення або параметризації. Конкатенативні системи TTS засновані на розділенні високоякісного записаного мовлення на дрібні фрагменти з наступним об'єднанням у нове мовлення. За останні роки цей метод став застарілим та непопулярним, оскільки не є масштабованим та послідовним. На відміну від цього, параметричні моделі базуються на виділенні акустичних особливостей із заданого тексту та перетворенні їх у звуковий сигнал за допомогою вокодерів. Цікаві результати параметричної TTS завдяки покращенню ефективності параметризації мови, моделюванню голосових шляхів та впровадженню глибоких нейронних мереж, очевидно, вказують на майбутнє синтезу мовлення. Рисунок 1.4 демонструє принципову конструкцію сучасних методів TTS.

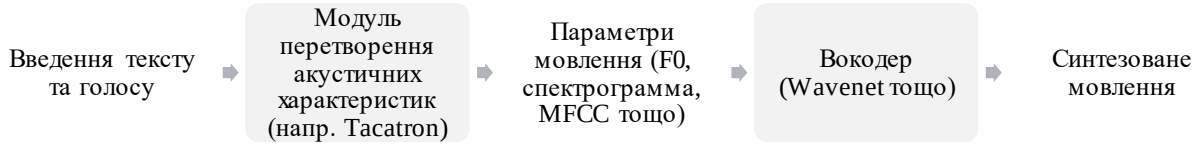

Рисунок 1.4 – Конструкція сучасних методів TTS

1.3. Аналіз проблем, що виникають при створенні Deepfake

Незважаючи на те, що були продемонстровані великі зусилля для поліпшення візуальної якості згенерованих Deepfake, все ще існує кілька проблем, які потрібно вирішити. Деякі з них розглядаються нижче.

Узагальнення: Генеративні моделі керуються даними, і тому вони відображають вивчені під час навчання особливості на виході. Для створення високоякісних Deepfake для навчання необхідна велика кількість даних. Більше того, сам навчальний процес вимагає години для створення переконливого глибокого аудіовізуального контенту. Зазвичай легше отримати датасет, але доступність достатньої кількості даних для конкретної жертви є складним завданням. Також перекваліфікація моделі для кожної конкретної цільової особистості є обчислювально складною. Через це потрібна узагальнена модель, яка дозволяє використовувати навчену модель для кількох цільових особистостей, невидимих під час навчання або з кількома доступними зразками навчання.

Навчання в парі: Навчена контрольована модель може генерувати високоякісний результат, але за рахунок об’єднання даних. Парування даних стосується отримання бажаного результату шляхом виявлення подібних вхідних прикладів з навчальних даних. Цей процес є трудомістким і не застосовним до тих сценаріїв, коли на етапі навчання використовуються різні види поведінки обличчя та різні особистості.

Варіації пози та відстань від камери: Існуючі методи дають хороші результати для фронтального виду обличчя. Однак якість маніпульованого вмісту значно погіршується для сценаріїв, коли людина не дивиться прямо у камеру. Це

призводить до небажаних візуальних артефактів навколо лицьової області. Крім того, ще однією великою проблемою для створення переконливої підробки є відстань до камери, оскільки збільшення відстані від пристроїв для зйомки призводить до неякісного синтезу обличчя.

Умови освітлення: Сучасні підходи до генерації Deepfake дають переконливі результати в контрольованому середовищі з постійними умовами освітлення. Однак різка зміна умов освітлення, таких як внутрішні та зовнішні сцени, призводить до невідповідності кольорів та появи дивних артефактів у отриманих відео.

Оклюзії: Однією з головних проблем у формуванні Deepfake є поява оклюзії, яка виникає, коли область обличчя джерела та жертви затуляється рукою, волоссям, окулярами або будь-якими іншими предметами. Більше того, оклюзія може бути результатом прихованого обличчя або частини ока, що врешті-решт спричиняє суперечливі риси обличчя у вмісті, яким маніпулюють.

Часова узгодженість: Ще одним недоліком сформованих підробок є наявність очевидних артефактів, таких як мерехтіння та тремтіння серед кадрів. Ці ефекти виникають тому, що фреймворки генерації Deepfake працюють на кожному кадрі без урахування часової узгодженості. Щоб подолати це обмеження, деякі роботи або надають цей контекст генератору або дискримінатору, розглядають втрати тимчасової когерентності, використовують RNN або використовують комбінацію всіх цих підходів.

Відсутність реалізму в синтезованому аудіо: Хоча якість, безумовно, стає набагато кращою, її все ще потрібно вдосконалювати. Основними проблемами аудіофайлів є відсутність природних емоцій, пауз, задишки та темпу, з яким ціль говорить.

Виходячи з вищезазначених обмежень, можна стверджувати, що існує потреба у розробці ефективних методів генерації Deepfake, стійких до змін в умовах освітлення, тимчасової когерентності, оклюзії, варіації пози, відстані камери, витоку ідентичності та парному тренуванню.

Висновки за розділом 1

У даному розділі було проведено ряд досліджень та отримано наступні результати:

1. Проаналізовано еволюцію технології Deepfake з метою розуміння того, як розвивалися методи маніпуляції медійним контентом, та в яких сферах вони застосовуються. Наведено найбільш поширені та легкодоступні додатки для створення Deepfake із зазначенням того, які техніки вони використовують. Наведено статистику для кількості Deepfake, виявлених у мережі Інтернет, із метою зазначення актуальності та стрімкого поширення даної проблеми.

2. Проведено огляд та аналіз сучасних методів створення Deepfake із розглядом принципів їхньої дії, а також сучасних досліджень, що прагнуть покращити ефективність наведених методів. Був проведений аналіз методів сворення не лише відео-, а й аудіопідробок, оскільки сучасні дослідження не приділяють достатньо уваги даному аспекту фальшивого мультимедійного контенту.

3. Здійснено аналіз проблем, які виникають при створенні Deepfake, з метою зорієнтувати дослідницьке суспільство на питання, які потрібно розглянути, щоб покращити сфери як генерації, так і виявлення Deepfake.

Таким чином, у першому розділі описано механізми створення Deepfake, а також їхні поточні обмеження та майбутні напрямки. Отримана інформація буде використовуватися для дослідження питання виявлення підробленого мультімедійного контенту.

РОЗДІЛ 2
ОСНОВНІ МЕТОДИ ВИЯВЛЕННЯ DEEPFAKE

2.1. Активне та пасивне виявлення

При відсутності обчислювальної потужності або її незначної кількості доведеться погодитися на ручне виявлення. За наявності обчислювальної потужності можна використовувати алгоритмічне виявлення, яке потім це можна розділити на активне виявлення, для чого потрібна попередня інформація, вбудована в зображення; та пасивне виявлення, яке можна використовувати, коли попередня інформація відсутня [46]. Прикладами активного виявлення є:

• водяний знак, який виявляє маніпуляції внаслідок слідів змін на непомітних водяних знаках;

• хешування зображень, що дозволяє аутентифікацію зображень за допомогою спільних секретних ключів.

Наразі на більшості пристроїв не реалізовано програмне забезпечення, яке підтримувало б активне виявлення зображень. Пасивне виявлення можна розділити на:

• методи глибокого навчання, які керуються даними, тобто вони отримують великий набір даних і знаходять власні причини для того, щоб вирішити, чи є зображення реальним чи ні;

• традиційні методи, в які люди або самі вкладають ці причини, і які можуть виділити сліди маніпуляцій, але залишають людям рішення щодо цілісності зображення.

Що стосується ручного виявлення неозброєним оком, два дослідження користувачів показують, що Deepfakes швидко стають занадто реалістичними для людей, щоб їх виявити:

• У роботі [47] проведено дослідження користувача за допомогою алгоритму Face2Face. Для версії, яка створює більш реалістичні відео, точність людини без

стиснення становила 60,57%; при жорсткому стисненні точність людини становила 48,93%. Дослідження показують, що навіть для необроблених та високоякісних зображень людська точність не перевищує 80%.

• У роботі [48] виконане дослідження, в якому користувачі повинні сказати, чи є зображення реальним чи ні. Для датасета FaceForensics ++, учасники випробування вважають 8,4% справжніми; Для датасета Celeb-df, учасники вважають 61,0% реальними; Для датасета DeeperForensics-1.0, учасники вважають до 64,1% реальними.

Що стосується різниці між активним та пасивним виявленням, активне виявлення, як правило, є більш надійним, ніж пасивне виявлення, але його застосовність обмежена, оскільки воно потребує спеціального обладнання; також для повністю сформованих зображень, водяного знака для вилучення може і не існувати, оскільки підробка не базується на первозданному зображенні. Джонстон та Еліан [49] також визнають, що більшість існуючих відео не мають водяних знаків.

Пасивні методи виявлення можуть мати низьку точність і можуть бути обчислювально дорогими [50]. Для повністю згенерованих зображень пасивні методи виявлення можуть не виявити слідів маніпуляцій, оскільки жодної модифікації зображень не відбулося [51].

Що стосується методів глибокого навчання, то вони полягають у тому, що вимагають багато часу та великих об’ємів даних. Алгоритм глибокого навчання, навчений певному типу маніпуляцій, не виявить іншого типу маніпуляції, навіть якщо ці два типи маніпуляцій семантично схожі. Методи глибокого навчання також зазнають того недоліку, що їхні правила прийняття рішень не може інтерпретувати людина. Це також призводить до невизначеності потенціалу глибокого навчання в області виявлення маніпуляцій із зображеннями. Більше того, методи глибокого навчання, швидше за все, будуть упередженими до навчального набору даних, тобто вони розпізнаватимуть зображення, що походить із певного набору даних [52]. Інше занепокоєння щодо методів глибокого навчання полягає в тому, що вони можуть бути включені в тренінг GAN, навчаючи дискримінатора GAN (див. Розділ 3.2.5) обходити такі методи [53].

З іншого боку, традиційні методи добре працюють, але залежать від припущень, зроблених при розробці алгоритму виявлення. Також не узгоджується, чи добре працюють традиційні методи виявлення на методах глибокого навчання.

Більшість алгоритмів пасивного виявлення розроблено для зосередження на виявленні певного типу слідів. Існує дві основні категорії, які слід розрізняти: міжкадрове та кадрове виявлення. Міжкадрове розпізнавання використовує зв’язки між послідовними кадрами для виявлення підробки. Це можна зробити за допомогою семантичних артефактів, таких як:

- занадто мало моргання очима;
- розбіжності в частоті серцевих скорочень записаного індивіда, наприклад шляхом вимірювання невеликих відмінностей у кольорі шкіри;
- аналіз орієнтирів обличчя за моделлю, побудованою на автентичних відеороликах відомих людей.

Також за допомогою статистичних невідповідностей, які можна виявити, наприклад:

- розбіжності в оптичному потоці послідовних кадрів;
- порівняння ключових точок зображення в послідовних кадрах;
- виявлення тимчасових невідповідностей шляхом застосування LSTM для зображення залишків кадрів;
- дескриптори мультимедійного потоку.
- застосування моделі глибокого навчання до необроблених RGB-даних, дозволяючи моделі знаходити власні кореляційні зв'язки між кадрами послідовностей для ознак маніпуляцій. Це переважно полягає у використанні довготермінової пам’яті для аналізу часових розбіжностей між відеокадрами.

Розпізнавання кадрів витягує окремі кадри з відео та розглядає їх як незалежні зображення. Алгоритми, які використовують кадрове виявлення, а потім усереднюють вихідну оцінку за декілька кадрів, вважаються кадровими методами виявлення. Кадрове виявлення можна поділити на семантичне та статистичне. Статистичне виявлення використовує функції низького рівня, прикладами яких є:

- локальні дескриптори об’єктів, вилучення, наприклад, ключових точок або країв зображення;
- викривляючі артефакти, що є суперечливістю у роздільній здатності, коли замінене обличчя перетворюється назад у вихідне зображення;
- матриці співвходження, витягнуті з вихідних RGB-даних;
- частота насичених пікселів, поданих у SVM11;
- гістограми інтенсивності, що показують кількість пікселів на рівень інтенсивності;
- різниці у прогнозованому та фактичному рівнях стиснення;
- артефакти моделі камери;
- артефакти на основі GAN.

В свою чергу, семантичне виявлення може базуватися на:

- відсутніх деталях у зубах та очах;
- невідповідному кольорі очей;
- розташування орієнтирів на обличчі, дозволяючи SVM виявляти неприродні відхилення;
- відмінності у вираженні конкретних емоцій;
- відсутні відбиття (у дзеркалі тощо);
- неправильно змодельована геометрія, що спричиняє артефакти навколо носа та краю обличчя.

2.2 Просторове виявлення

2.2.1 Виявлення на основі експертизи зображень

Традиційні методи експертизи перевіряють диспропорції на рівні пікселів, що досліджується останніми працями на тему виявлення DeepFake. Вони дають пояснювані підказки при виявленні та вводять відмінності між реальним та підробленим. Однак ці праці страждають від проблем надійності, коли зображеннями та відеозаписами маніпулюють за допомогою простих перетворень.

У роботі [54] спостерігається, що відмінності між синтезованими та реальними обличчями виявляються в компонентах кольоровості, особливо у залишковій області. Вони пропонують навчити однокласний класифікатор на реальних особах, використовуючи відмінності в компонентах. Однак їх ефективність проти атак порушення, таких як перетворення зображень, невідома.

Шаблон неоднорідності фотовідгука (PRNU, Photo Response Non-Uniformity) – це шаблон шуму в цифровому зображенні, викликаний датчиком освітленості в камері, який можна використовувати для відрізняння Deepfake від автентичних відео. Інші досліджують використання матриць співвходження для розрізнення реальних і підроблених облич. Сенс цих робіт очевидний, але їх ефективність у вирішенні складних проблем Deepfakes високої якості не ясна.

Аналогічним чином, при роботі з підробленими відео дослідники також запозичують ідеї з традиційної відео-криміналістики, використовуючи особливості локального руху, захоплені з реальних відео, для виявлення ненормальності змінених відео. Використання методів судової експертизи зображень та відео – це ідея для боротьби з DeepFake за рахунок зосередження уваги на низькорівневих функціях, але вони не практичні для розгортання в реальних умовах.

2.2.2 Виявлення на основі DNN

Ці методи повністю керуються даними шляхом використання існуючих або проектування нових моделей на основі DNN шляхом вилучення просторових особливостей для підвищення ефективності та узагальнювальної здатності виявлення. Однак усі ці методи виявлення, засновані на DNN, страждають від атак з додатковими шумами, і дослідження не змогли оцінити їх ефективність у боротьбі з подібними атаками. Наявні дослідження, що використовують DNN для виявлення Deepfakes, можна класифікувати на наступні дві категорії.

Вдосконалення узагальнюючих здібностей. Звичайні DNN широко застосовуються для виявлення підробок, але вони переобучаються на конкретні типи маніпуляцій та страждають від проблем перенесення. Таким чином,

мотивована процесами соціального сприйняття та соціального пізнання людського мозку, нова ієрархічна мережа пам'яті (HMN, Hierarchial Memory Network) використовується для виявлення фальшивих облич для вирішення проблем перенесення та підвищення ефективності взаємодії із невідомими GAN.

Дослідження признаків артефактів. Для того, щоб зосередити увагу на власних криміналістичних підказках, попередня обробка зображень за допомогою згладжуючої фільтрації або шуму застосовується для знищення нестабільних артефактів низького рівня на зображеннях, синтезованих GAN. Дослідження внутрішніх підказок може суттєво покращити здатність узагальнення моделі CNN при виявленні невідомих GAN.

2.2.3 Виявлення очевидних артефактів

Через обмеження існуючих методів ШІ, створені Deepfake демонструють деякі очевидні артефакти, які можна використати для виявлення за допомогою деяких простих моделей DNN. У роботі [55] досліджено, що локальні патчі мають зайві артефакти, які можна використовувати для розрізнення фальшивих облич. Повністю згортковий підхід застосовується щоб навчити класифікатори зосередити увагу на плямах зображень. Цей підхід можна добре узагальнити для різних мережевих архітектур, наборів даних зображень тощо.

Невідповідність між обличчями та їхнім контекстом є ще одним артефактом для виявлення підробок. Мережа ідентифікації облич навчається з використанням області обличчя для ідентифікації людини, тоді як мережа розпізнавання контексту навчається за допомогою контексту обличчя, наприклад волосся, вуха, для ідентифікації людини. Два вектори з вищезазначених двох мереж порівнюються для виявлення розбіжностей між особистостями. Цей підхід також має хороші здібності до узагальнення для GAN. Для кожної людини можна вирізнити певну манеру руху обличчя, коли вона говорить. Це можна використати для захисту знаменитостей, маючи великий обсяг даних для навчання. Ці підходи просто використовують артефакти як підказки для виявлення без введення нових моделей DNN, отже, вони

будуть недійсними, коли GAN оновиться або артефакти будуть виправлені в новій версії.

2.3 Виявлення, засноване на частоті пікселів

Замість того, щоб досліджувати візуальні артефакти, деякі дослідники працюють над дослідженням недосконалості існуючих GAN, що забезпечує очевидні сигнали для розрізнення реальних та фальшивих облич. Вони зазвичай працюють у частотній області.

У роботі [56] досліджується архітектура моделі генератора і спостерігається, що внутрішнє значення генератора нормалізується, що обмежує частоту насичених пікселів. Потім простий класифікатор на основі SVM навчається вимірювати частоту насичених та недостатньо експонованих пікселів на кожному зображенні обличчя для розрізнення фальшивих облич. Однак стійкість проти атак збурень не вивчається.

У роботі [57] вперше представлено відбитки пальців GAN для класифікації зображень як реальних або синтезованих із GAN. Відбитки GAN можуть бути надалі використані для прогнозування джерела зображень. Дослідження показало, що незначні відмінності в навчанні GAN можуть призвести до чітких відбитків GAN. Однак відбитки можуть бути легко знищені простими атаками збурень, такими як розмиття, стиснення JPEG тощо.

Інші дослідження також використовують відбитки GAN для розрізнення фальшивих облич, синтезованих GAN. Артефакти GAN є перспективним ключем для виявлення, однак артефакти можуть бути легко пошкоджені за допомогою деяких простих перетворень зображень, таких як неглибока реконструкція за допомогою PCA тощо.

2.4 Виявлення, засноване на біологічних сигналах

Справжні нерухомі зображення обличчя та відео створюються за допомогою камер, які є природними порівняно із синтезованими підробленими обличчями. Артефакти біологічних сигналів на синтезованих підроблених обличчях дають очевидні підказки для виявлення підробки. Ці біологічні сигнали можна класифікувати на наступні категорії.

2.4.1 Аудіовізуальні невідповідності

Поєднання візуального та звукового сигналів для виявлення невідповідності у фальшивих обличчях є новим способом для розрізнення Deepfake. Ці методи можуть добре пояснити, чому відео є фальшивим. Сіамська мережа використовується для моделювання візуального та звукового сигналів у відеозаписах із комбінацією двох функцій триплетних втрат для вимірювання подібності.

Зокрема, одна функція втрат призначена для обчислення подібності між візуальним та звуковим сигналами, інша функція втрат розроблена для обчислення сигналів ефекту, наприклад сприйняті емоції. Експерименти показують, що цей метод перевершує звичайні методи на основі DNN при виявленні фальшивих відео.

Синхронізація губ – це типовий Deepfake, що синтезує рухи рота людини, щоб відповідати мовленню. Основне розуміння полягає в тому, що динаміка форми рота іноді не узгоджується з розмовною фонемою у підроблених відео.

Зокрема, губи повинні бути закриті, коли вимовляються деякі слова, що починаються на М, Б, П. Однак це правило порушується у фальшивих відео. Дослідники використали це знання для виявлення Deepfake із синхронізацією губ.

2.4.2 Візуальні невідповідності

Візуальна невідповідність свідчить про те, що синтезовані обличчя не є природними, особливо форма, риси та орієнтири облич. У роботі [58] відзначають, що синтезовані підроблені обличчя завжди мають фіксовані розміри через обмеженість обчислювальних ресурсів та час виготовлення алгоритмів Deepfake. Фіксований розмір синтезованих облич залишає артефакти деформації, які можна використовувати для виявлення DeepFake. Потім модель CNN проходить навчання для виявлення артефактів. Відсутність моргання очима – це ще одна ознака для виявлення Deepfakes. CNN у поєднанні з рекурсивною нейронною мережею тренується для розрізнення стану очей. Невідповідні орієнтири на підроблених обличчях невидимі для людських очей, але їх легко виявити за допомогою глибинного навчання.

Візуальні артефакти, такі як очі, зуби, контури обличчя, є важливою ознакою для викриття Deepfake [59]. Невідповідне дзеркальне виділення рогівки між двома очима – ще одна підказка для виявлення синтезованих за допомогою GAN облич. Ця невідповідність головним чином пов’язана з відсутністю фізичних / фізіологічних обмежень у існуючих популярних GAN. Усі ці методи засновані на спостереженні, що на підроблених обличчях виявляються очевидні артефакти для людських очей, особливо на невідповідностях, що з'явилися на обличчі порівняно з реальними обличчями. Вони надають вагомі гарантії для пояснення рішення при розрізненні справжнього чи фальшивого, але вони будуть недійсними, коли будуть запропоновані розширені GAN. Крім того, їхня стійкість проти атак збурень незрозуміла.

2.4.3 Біологічні сигнали на відео

Біологічні сигнали у відео непросто відтворити. У FakeCatcher виділяється шість різних біологічних сигналів для використання просторової та часової

узгодженості для автентифікації реальних відео, зроблених камерою. Дослідження показали, що частоту серцевих скорочень можна використовувати для виявлення фальшивих відео, однак отримання частоти серцевих скорочень із відеозаписів є трудомістким завданням. У роботі [60] використовують нейронне звичайне диференціальне рівняння (Neural-ODE), навчене на оригінальних відео, щоб передбачити частоту серцевих скорочень на тестованих відео. DeepRhythm також виставляє відео DeepFake, відстежуючи ритми серцебиття. Зокрема, вони розробляють збільшене в русі просторово-часове подання (MMSTR, Motion-Magnified Spatial-Temporal Representation) у відео для виділення сигналів серцевого ритму.

2.5 Аналіз проблем, що виникають при виявленні Deepfake

Незважаючи на те, що були досягнуті значні успіхи у роботі глибинних детекторів, є численні занепокоєння щодо сучасних методів виявлення, які потребують уваги. Нижче розглядаються деякі проблеми підходів щодо виявлення Deepfake.

Якість Deepfake-датасетів. Доступність великих баз даних є важливим фактором у створенні методів розкриття підробок. Однак аналіз якості відео з цих наборів даних виявляє кілька неоднозначностей у порівнянні з фактичним маніпульованим вмістом, знайденим в Інтернеті. Різними візуальними артефактами, які можна візуалізувати в цих базах даних, є:

- тимчасове мерехтіння в деяких випадках під час виступу;
- розмитість навколо лицьових областей;
- надмірна гладкість текстури обличчя / відсутність деталей текстури обличчя;
- відсутність рухів головою;
- відсутність предметів, що закривають обличчя, таких як окуляри, ефект блискавки тощо;

- чутливість до змін у поставі або погляду, невідповідності кольору шкіри та витоку ідентичності;
- обмежена доступність комбінованого якісного аудіо-візуального набору даних.

Вищезазначені неоднозначності набору даних зумовлені недосконалими етапами в техніках маніпуляцій. Крім того, маніпульований вміст низької якості не може бути переконливим або створити враження що він справжній. Отже, навіть якщо підходи виявлення демонструють кращу ефективність порівняно з такими відео, не гарантується, що ці методи будуть добре працювати, якщо використовувати їх на реальних підроблених відео, що циркулюють в Інтернеті.

Оцінка ефективності. В даний час методи виявлення підробок сформульовані як проблема двійкової класифікації, де кожен зразок може бути як реальним, так і фальшивим. Таку класифікацію простіше побудувати в контрольованому середовищі, де ми генеруємо та перевіряємо методи розкриття підробок, використовуючи аудіо-візуальний вміст, який є оригінальним або виготовленим. Однак для сценаріїв реального світу відео можна змінювати іншими способами, тому вміст, не виявлений як маніпульований, не гарантує, що відео є оригінальним. Крім того, контент глибокої підробки може бути предметом багатьох типів змін, тобто аудіо / візуального, і тому окремий ярлик може бути не зовсім точним. Більше того, у візуальному контенті з обличчями кількох людей, як правило, одним або кількома з них маніпулюють за допомогою глибоких підробок над сегментом кадрів. Тому бінарну схему класифікації слід вдосконалити до багатокласової, щоб справлятися із викликами реальних сценаріїв.

Відсутність пояснень у методах виявлення. Існуючі підходи до виявлення підробок, як правило, призначені для проведення пакетного аналізу за великим набором даних. Однак, коли ці методи застосовуються на місцях журналістами або правоохоронними органами, для аналізу може бути наявний лише невеликий набір відео. Цифрова оцінка, паралельна ймовірності того, що аудіо чи відео є справжнім чи підробленим, не є настільки цінною для практиків, якщо вона не може бути підтверджена. У таких ситуаціях дуже часто вимагають пояснень числової оцінки,

щоб можна було повірити в аналіз перед публікацією чи використанням у суді. Однак більшості методів виявлення підробок не вистачає такого пояснення, особливо тим, які базуються на підходах DL через їх природу «чорного ящика».

Часова агрегація. Існуючі методи виявлення підробок засновані на двійковій класифікації на рівні кадру, тобто перевірці ймовірності кожного відеокадру як реального чи маніпульованого. Однак ці підходи не враховують часову узгодженість між кадрами і страждають від двох потенційних проблем:

- вміст deepfake відображає часові артефакти;
- реальні або підроблені кадри можуть з'являтися з послідовними інтервалами.

Крім того, ці методи вимагають додаткового кроку для обчислення оцінки цілісності на рівні відео, оскільки ці методи повинні поєднувати оцінку з кожного кадру для отримання кінцевого значення.

«Відмивання» соціальних мереж. Соціальні платформи, такі як Twitter, Facebook або Instagram, є основними мережами, що використовуються для поширення аудіо-візуального контенту серед населення. Щоб зберегти пропускну здатність мережі або забезпечити конфіденційність користувача, такий вміст перед завантаженням позбавляється метаданих, та істотно стискається. Ці маніпуляції, які зазвичай називають відмиванням соціальних мереж, усувають підказки щодо підробок і врешті-решт збільшують показники помилково позитивних виявлень.

Висновки за розділом 2

У цьому розділі представлено огляд методів виявлення підробленого мультимедійного контента, які можна розділити на активні та пасивні, а також просторові, частотні та біологічні.

Семантичне виявлення базується, наприклад, на фізичних або фізіологічних артефактах. Це означає, що ці стратегії концептуально легко зрозуміти людям: це стосується типу артефактів, які ми б самі помітили (якби вони були досить чіткими). Недоліком методів, заснованих на конкретних візуальних артефактах, є те, що вони

не можуть бути універсально застосовними; наприклад, виявити розмиті зуби можливо лише на зображеннях, на яких особа зображена із відкритим ротом. Більше того, візуальні артефакти навряд чи будуть корисними в майбутньому, оскільки якість Deepfake швидко покращується. Наприклад, стратегію виявлення відсутності моргання очима було обійдено шляхом додавання зображень із закритими очима на етапі навчання алгоритмів генерації [61].

Методи, зосереджені на фізичних артефактах, залежать від конкретних припущень моделювання. Вони більш надійні, ніж виявлення на рівні пікселів, але в реалістичних ситуаціях методи виявлення на рівні пікселів перевершують фізичні методи, засновані на артефактах. Зокрема, побудова моделі для виявлення розбіжностей на обличчі є ефективною; однак, це вимагає побудови індивідуальної моделі для кожної людини, до якої вона застосовується. Отже, неможливо застосувати цей метод до кожної людини у кожному відео.

Можливо, найбільшим недоліком виявлення із використанням згорткових моделей на суто RGB-даних є те, що висновки мереж не можна легко пояснити для людей. Більше того, коли застосовується складна пост-обробка, такий тип виявлення може не спрацювати. Відповідно, висловлюються сумніви щодо того, чи є (чи буде) достатньо візуальних підказок у Deepfakes для чисто згорткових (конволюційних) мереж, щоб їх помітити. Також зазначається, що згорткові мережі мають перекваліфікуватися кожного разу, коли публікується алгоритм нового покоління; це проблематично, враховуючи темп публікації алгоритмів нового покоління.

Однак, навіть якщо чисто згорткові мережі не вважаються недостатньо надійними для глибокого виявлення, їх все одно варто дослідити: вони також є основою багатьох методів статистичного виявлення, тому їх розуміння є надзвичайно важливим.

РОЗДІЛ 3
ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ DEEPFAKE

3.1 Згорткові мережі

Більшість методів виявлення маніпуляцій у медійному контенті використовують згорткові (конволюційні) нейронні мережі. Згорткова нейронна мережа (CNN, Convolutional Neural Network) – це пряма нейронна мережа, широко використовувана у програмах, пов’язаних із зором, що отримала свою назву завдяки застосуванню принаймні одного так званого згорткового шару [62]. Згортковий шар розділяє заданий вхід на менші частини, використовує ядро для виконання операції згортки, яка витягує шаблони об’єктів з кожної частини і, нарешті, виводить карту об’єктів.

Дві основні конфігурації, які ми можемо змінити, щоб змінити поведінку згорткових шарів, - це розмір ядра та заповнення. Розмір ядра може бути великим, наприклад 11x11, або мати менший розмір, такий як 3x3, або що-небудь між ними. В даний час 3x3 є найбільш широко використовуваним розміром ядра [62]. Заповнення відноситься до техніки, коли ми додаємо додатковий шар нулів як межу навколо введення. По суті, це означає, що ми будемо збирати більше інформації про ці середні частини, і, оскільки розмір вхідного обсягу зменшується, інформація з кутових частин може загубитися [63]. Щоб ми зберегли якомога більше інформації про початкові вихідні дані, ми можемо застосувати доповнення нулями навколо входу, тому інформація з кутових частин втрачається, втрачається лише заповнення.

Оскільки згенеровані функції із згорткового шару можуть бути досить великими, за цими типами шарів часто слідує шар об'єднання з метою зменшення складності. Цей процес згортання та об'єднання шарів можна побачити на рисунку 3.1.

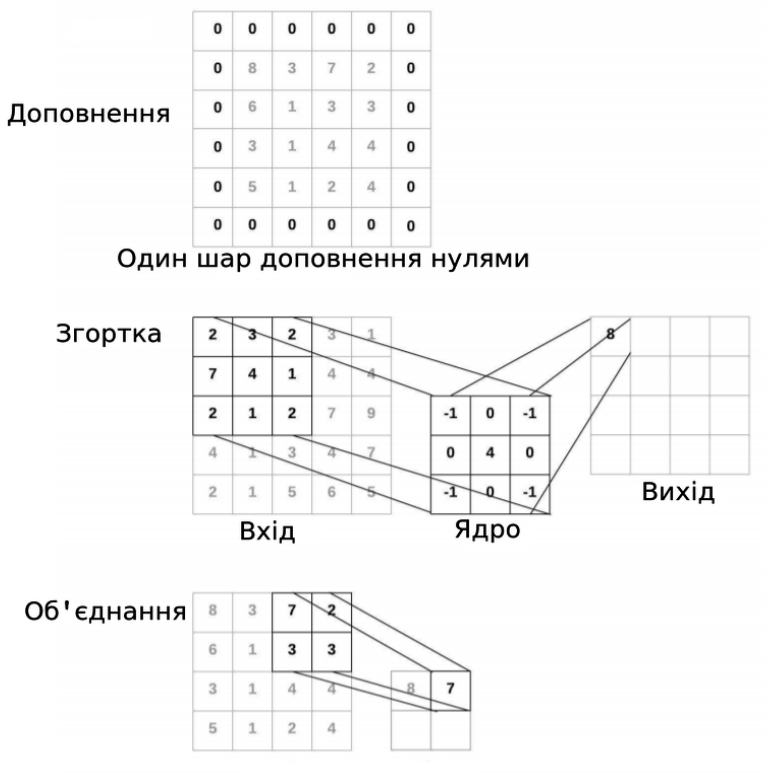

Рисунок 3.1 – Згорткова нейронна мережа

У CNN за цією базою згорткових та об'єднуючих шарів слідує класифікаційна частина, яка, як правило, складається з поєднання повністю з'єднаних шарів та шарів, що випадають. Випадаючий шар має певну функцію в нейронних мережах, пов’язану з запобіганням проблемі перенавчання. Шар «випадає» з випадкового набору нейронів, змушуючи нейронну мережу переконатися, що вона може надати правильну класифікацію, не надто покладаючись на конкретні активовані нейрони, не даючи їй повністю запам’ятовувати дані тренувань [64].

Популярність CNN в першу чергу пояснюється її здатністю до вилучення ознак, що допомагає виявляти особливості на різних рівнях. В основному, CNN, яка навчається на зображеннях людських облич, вчиться розпізнавати особливості нижчого рівня, такі як лінії, краї, кути, кола та інші основні форми, а також особливості вищого рівня, такі як ніс, очі, губи та інші частини обличчя. Далі модель CNN використовуватиме виявлені ознаки для прийняття рішення щодо класифікації.

Процес класифікації відео майже однаковий, оскільки відео по суті являє собою безліч кадрів зображень, складених у цілісне рухоме зображення. Як і у випадку з іншими методами машинного навчання, були докладені зусилля для покращення роботи CNN та пошуку рішень для їхніх проблем. Оскільки розташування компонентів CNN, як виявилося, відіграє центральну роль у досягненні підвищеної продуктивності, еволюція архітектур CNN почала наростати [65]. Слово архітектура стосується загальної структури мережі, наприклад, скільки шарів у неї є і як ці шари пов'язані між собою.

Для даної роботи були обрані ті моделі CNN, вихідний код яких знаходиться у публічному доступі, а саме XceptionNet, EfficientNet, ResNet та MesoNet

Архітектура Xception є частиною категорії: Багатоз'єднальні CNN на основі ширини. Ця архітектура використовує іншу форму процесу згортки у своїх згорткових шарах. Xception використовує ідею глибоко відокремлюваних згорткових шарів, що є варіантом традиційного згорткового шару, але ці типи шарів розділяють ядро на два окремих ядра для виконання двох звивин замість одного, з метою покращення обчислювальних характеристик. Метод XceptionNet [66] бере свій початок з наукової роботи, опублікованої на початку 2019 року, а потім оновленої в серпні того ж року. Модель XceptionNet – це традиційна модель CNN, побудована на основі архітектури CNN Xception. Цю архітектуру вибрали автори, оскільки вони бажали моделі виявлення, яка здатна досягти переконливих результатів на зображеннях зі слабким або відсутнім стисненням, зберігаючи при цьому розумну продуктивність на низькоякісних зображеннях. Теоретично це означало б, що модель більше підходить для обробки відео з різними станами, що часто буває в реальних ситуаціях.

Архітектура ResNet є частиною двох категорій: CNN на основі глибини та CNN на основі декількох шляхів [67]. Для першої категорії глибинні архітектури CNN базуються на припущенні, що збільшена глибина мережі відіграє важливу роль у рівні успіху класифікації. Архітектура ResNet покращила завдання розпізнавання зображень та локалізації, в той же час вимагаючи меншої складності обчислень, ніж запропоновані раніше мережі. Що стосується другої категорії, багатопроменеві

архітектури CNN здатні вирішувати загальні проблеми, з якими стикаються CNN під час навчання, такі як проблема зменшення градієнта, за допомогою концепції багатопроменевого або міжшарового зв'язку. Архітектура ResNet використовує цю ідею, систематично підключаючи один шар до іншого, забезпечуючи спеціалізований потік інформації між шарами. У поєднанні ResNet зробив революцію в архітектурній гонці CNN, представивши свою концепцію залишкового навчання та запропонувавши істотно глибокі варіації моделей з 50, 101 і 152 шарами

Google EfficientNet – це потужна архітектура згорткової нейронної мережі (CNN), яка була використана в найефективніших моделях у Deepfake Detection Challenge, яке було оголошено в червні минулого року на Facebook [68]. EfficientNet використовує мобільну перевернуту згортку вузьких місць (MBConv). Це забезпечує хороший компроміс між параметрами мережі та точністю класифікації. Крім того, мережа була розроблена з використанням алгоритмів пошуку нейронної архітектури (NAS), в результаті чого вийшла мережа, яка є одночасно компактною та точною. Насправді ця мережа перевершила попередні найсучасніші підходи в наборах даних, таких як ImageNet, маючи менше параметрів.

MesoNet починається з послідовності з чотирьох шарів послідовних звивин та об'єднання, а за нею - щільна мережа з одним прихованим шаром [69]. Для поліпшення узагальнення згорткові шари використовують функції активації ReLU, які вводять нелінійності та пакетну нормалізацію, щоб регулювати їх вихід і запобігати ефекту зникаючого градієнт. MesoNet має низьку кількість шарів для зосередження уваги на мезоскопічних властивостях зображення.

Таблиця 3.1

Інформація про досліджувані моделі

Модель	Розмір файла	Посилання	Дата публікації
XceptionNet	79,6 МБ	https://arxiv.org/pdf/1610.02357.pdf	2017
EfficientNet	244 МБ	https://arxiv.org/pdf/1905.11946.pdf	2019
ResNet	96 МБ	https://arxiv.org/pdf/1905.10596.pdf	2018
MesoNet	122 МБ	https://arxiv.org/pdf/1809.00888.pdf	2019

3.2 Оцінка моделей машинного навчання

3.2.1 Параметри

Точність моделі можна визначити як співвідношення між кількістю правильно класифікованих прогнозів та загальною кількістю прогнозів, як це видно з рівняння:

$$\text{точність} = \frac{\#\ \text{правильних передбачень}}{\#\ \text{передбачень в цілому}}$$

Незважаючи на те, що ми прагнемо розробити високоточні моделі, однієї лише точності може бути недостатньо для забезпечення хороших показників, оскільки ця оціночна метрика не робить різниці між різними класами, а отже, ми не будемо знати, чи виникають помилки, оскільки модель не виявляє Deepfake або неправильно класифікує реальні відео як Deepfake. У цьому випадку перегляд матриці помилок [70] результатів покаже більш детальну розбивку правильних та неправильних прогнозів для кожного класу. В контексті цього дослідження матриця помилок буде розділена на чотири частини, як видно з таблиці 3.2.

Таблиця 3.2

Матриця помилок

True Positive (TP) Зображення чи відео підроблене Передбачення: Deepfake	False Positive (FP) Зображення чи відео справжнє Передбачення: Deepfake

False Negative (FN) Зображення чи відео підроблене Передбачення: справжнє	True Negative (TN) Зображення чи відео справжнє Передбачення: справжнє

Чутливість та специфічність використовуються у двійковій класифікації для вимірювання продуктивності моделі, вказуючи наскільки вагомими є результати її тесту. Ці виміри використовують для своїх розрахунків різні частини матриці помилок. Чутливість, яка називається відстоком справжніх позитивних спрацювань (TPR, True Positive Rate) або ймовірністю виявлення, вимірює відсоток глибоких підробок, які правильно визначені як підроблені, і, як видно з рівняння, це обчислюється діленням кількості справжніх позитивних результатів (TP) на суму справжніх позитивних (TP) та помилкових негативних (FN).

$$\text{чутливість} = \frac{TP}{TP + FN}$$

Специфічність, яку також називають відсотком справжніх негативних спрацювань (TNR, True Negative Rate), оцінює частку фактичних негативів, які правильно визначені як такі, тобто відсоток реальних відео, правильно класифікованих як реальні. Як показує рівняння, специфічність обчислюється шляхом ділення справжніх негативних (TN) на суму справжніх негативних (TN) та помилкових позитивних спрацьовувань (FP):

$$\text{специфічність} = \frac{TN}{TN + FP}$$

Якщо результат тесту показує високу чутливість і низьку специфічність, модель має високий рівень виявлення Deepfake, але водночас вона також неправильно класифікувала багато справжніх відео як Deepfake. Навпаки, якщо тест

показує низьку чутливість та високу специфічність, модель не може виявити багато підробок, неправильно класифікуючи їх як реальні, але також рідко класифікує реальні відео як підроблені. В контексті цього дослідження були б бажаними як висока чутливість, так і висока специфічність, однак, загалом, ці два показники часто демонструють антикореляційні зв'язки на тестах [71]. Отже, можна стверджувати, що висока чутливість важливіша для моделі виявлення фальшивих фальшивок, ніж висока специфічність. Імовірно, було б краще виявити більше підробок, навіть якщо деякі реальні відео були помилково виявлені як такі, ніж якби не вдалося виявити підробки.

Функція втрат була згадана як частина того, як ANN обчислюють свої помилки та вимірюють якість своїх прогнозів. Чим менше значення втрат виробляє функція, тим вищий ступінь правильних прогнозів, що робить це значення корисною метрикою як для оцінки ефективності моделі, так і для порівняння різних моделей. Значення втрат враховує, наскільки певною є модель при правильному прогнозуванні. Якщо прогноз відхиляється від фактичної класифікації, модель карає себе за впевненість, все ще помиляючись, і збільшує свою втрату на основі того, наскільки оцінка її прогнозу відрізнялася від правильного класу. Аналогічним чином, значення втрат може показувати ознаки того, коли модель перенавчається, в такому випадку втрати тренувань зменшуються до тих пір, поки вона не буде нижчою, ніж втрата перевірки.

AUC означає область під кривою ROC (Reciever Operating Characteristic). Крива ROC ілюструє ймовірність виявлення, викладаючи відношення відсотку TP до відсотку FP [72]. Оскільки крива ROC надає деталі поведінки моделі, може бути складно порівняти кілька кривих ROC між собою, і, отже, AUC використовується як спосіб узагальнити показники в одне число, яке легко порівняти. Значення AUC говорить нам, наскільки модель здатна розрізняти різні класи. AUC завжди буде між 0,0 і 1,0, чим вищий AUC, тим краща модель прогнозує реальні відео як реальні, так і підробки як підробки. Отже, низька AUC небажана до того, що жодна практична модель не повинна мати AUC менше 0,5, що дало б значення гірше, ніж випадкове вгадування у двійковому випадку [72].

3.2.2 Датасет

Як правило, моделі глибокого навчання, такі як CNN, залежать від якісних даних для вивчення та вдосконалення їх алгоритму, що робить доступність великомасштабних та сучасних Deepfake-датасетів вирішальним фактором у розробці методів виявлення Deepfake. Крім того, вважається необхідним, щоб візуальна якість відеороликів deepfake відповідала реальним deepfake, що циркулюють в Інтернеті, щоб гарантувати що результати вийдуть максимально реалістичними. Підробки з низькою якістю візуального зображення навряд чи будуть переконливими в реальних сценаріях, і, відповідно, висока ефективність виявлення таких відео може не мати великого значення.

Зважаючи на вищесказане, для оцінки алгоритмів виявлення Deepfake було обрано масштабний складний набір відеокліпів DeepFake, CelebDF, який містить 5 639 високоякісних відео DeepFake знаменитостей, створених за допомогою вдосконаленого процесу синтезу [73].

Всього в наборі даних Celeb-DF є 5 639 відео DeepFake, що відповідають понад 2 мільйонам кадрів. Реальні відеоролики базуються на загальнодоступних відеокліпах YouTube із 59 знаменитостей різної статі, віку та етичних груп. Відео DeepFake створюються за допомогою вдосконаленого методу синтезу DeepFake. Як результат, загальна візуальна якість синтезованих відео DeepFake у Celeb-DF значно покращується порівняно з існуючими наборами даних, із значно меншою кількістю помітних візуальних артефактів, див. Рис.2. На основі набору даних Celeb-DF та інших існуючих наборів даних ми проводимо оцінку поточних методів виявлення DeepFake

Набір даних Celeb-DF складається з 590 реальних відео та 5639 відео DeepFake (що відповідає понад двох мільйонам відеокадрів). Середня тривалість усіх відеороликів становить приблизно 13 секунд при стандартній частоті кадрів 30 кадрів в секунду. Реальні відео вибираються із загальнодоступних відео YouTube, що відповідає інтерв'ю 59 знаменитостей з різним розподілом за статтю, віком та етнічними групами. 56,8% суб'єктів у реальних відео – чоловіки, а 43,2% – жінки.

8,5% – віком від 60 років, 30,5% – від 50 до 60 років, 26,6% – 40 років, 28,0% – 30 років, 6,4% – молодше 30 років. 5,1% – азіати, 6,8% – афроамериканці та 88,1% – білі . Крім того, реальні відео демонструють широкий спектр змін у таких аспектах, як розмір обличчя випробовуваних (у пікселях), орієнтація, умови освітлення та фони. Відео DeepFake створюються шляхом обміну гранями для кожної пари з 59 предметів. Остаточні відео представлені у форматі MPEG4.0.

Зважаючи на обмеженість обчислювальних ресурсів та об'єм пам'яті, було вибрано використовувати зменшену версію датасета, що складається із 518 відео, із яких 340 – Deepfake, 178 – справжні відео.

3.2.3 Умови тестування

Для оцінки роботи алгоритмів було обрано детектор deepfake, що забезпечує порівняльний аналіз (бенчмаркинг), навчання та виявлення одиничних відео deepfake за допомогою можливості завантажити обрані методи виявлення Deepfake.

Цей експеримент може виконуватися в будь-якому бажаному середовищі із версією Python вижче, ніж Python 3. У цьому дослідженні використовуваною версією Python була 3.9, а експеримент проводився на операційній системі Windows 10 з використанням ЦП для обчислень під час роботи і використанням Thonny у якості редактора коду.

Для досягнення мети буде використано метод «benchmarking», що дозволяє обрати бажаний метод виявлення Deepfake та датасет, на якому цей метод буде тестуватися. Це робиться за допомогою команди:

python deepfake_detector/dfdetector.py --benchmark True --data_path path\celebdf --detection_method <method>

Де у якості <method> будуть використовуватися параметри xception_celebdf, efficientnetb7_celebdf, resnet_lstm_celebdf та mesonet_celebdf відповідно.

3.3 Результати тестування

В результаті тестування моделей виявлення Deepfake було отримано результати, оформлені у вигляді матриць помилок для кожної моделі. На рис. 3.2-3.3 можна побачити візуальне представлення матриць помилок для XceptionNet, EfficientNet, ResNet та MesoNet:

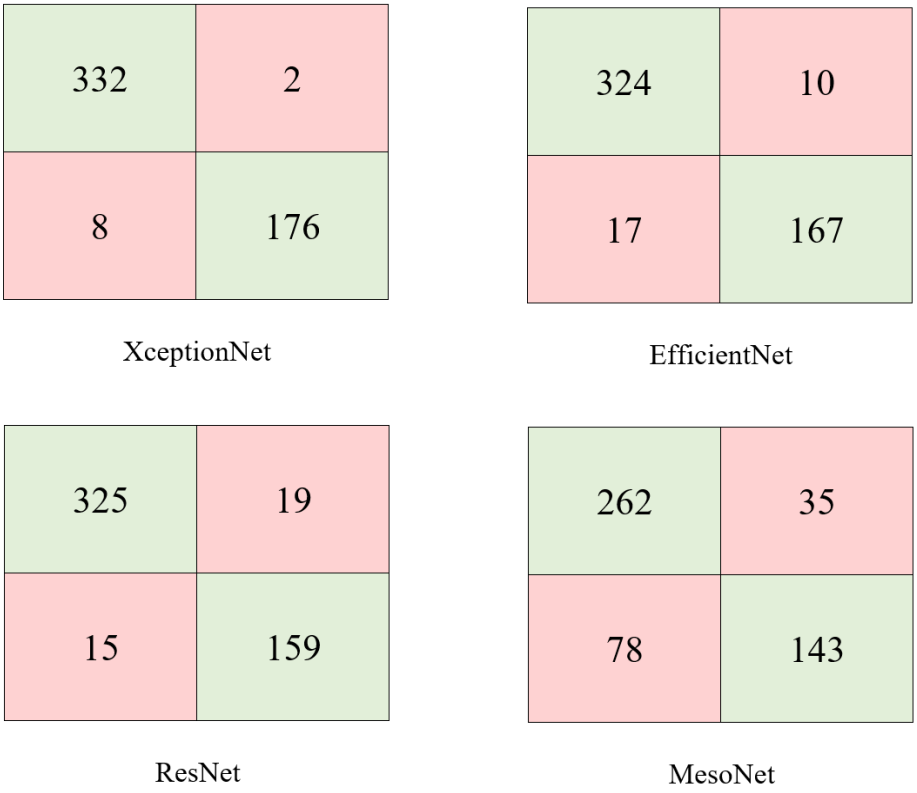

Рисунок 3.2 – Матриці помилок

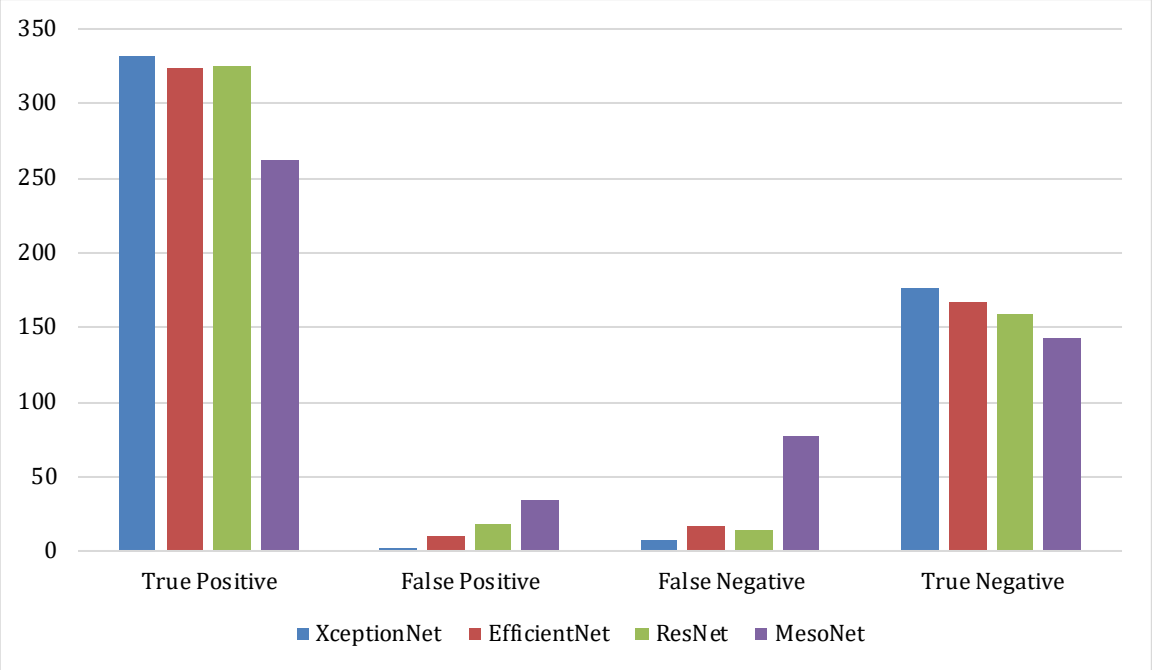

Рисунок 3.3 – Порівняння матриць помилок

Результати тестування на точність, чутливість, специфічність, криву AUC та рівень втрат для моделей виявлення Deepfake наведені у таблиці 3.3 та представлені у рисунках 3.4-3.6 для кращої візуалізації.

Таблиця 3.3

Результати тестування для моделей виявлення Deepfake

Модель	Точність, %	Чутливість	Специфічність	AUC	Втрати
XceptionNet	**98.06%**	**0,9764**	**0,9887**	**0,99835**	**0,1504**
EfficientNet	94,78%	0,9501	0,9435	0,9853	0,1812
ResNet	93,44%	0,9558	0,8932	0,9689	0,1987
MesoNet	78,18%	0,7705	0,8033	0,87507	0,56423

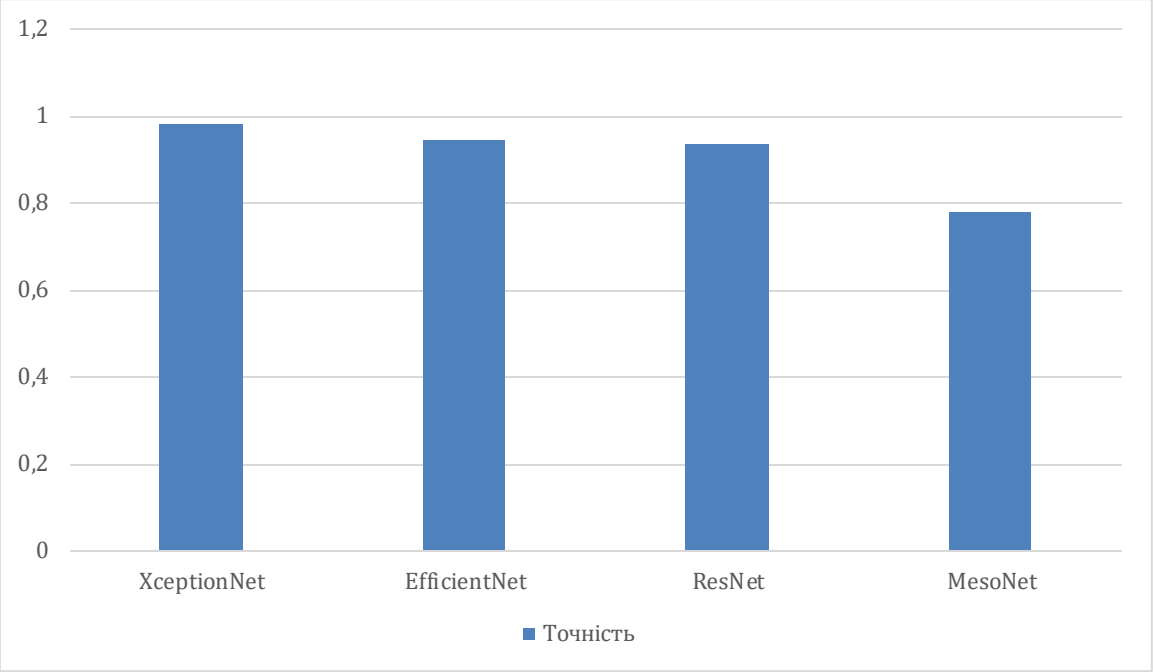

Рисунок 3.4 – Порівняння точності моделей

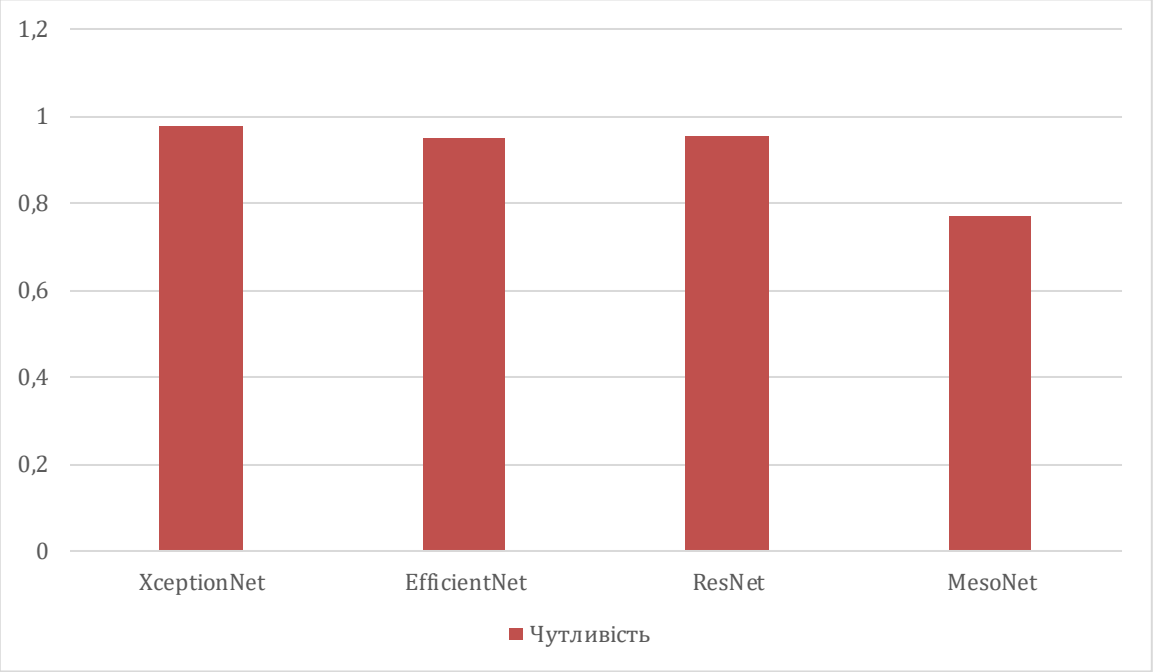

Рисунок 3.5 – Порівняння чутливості моделей

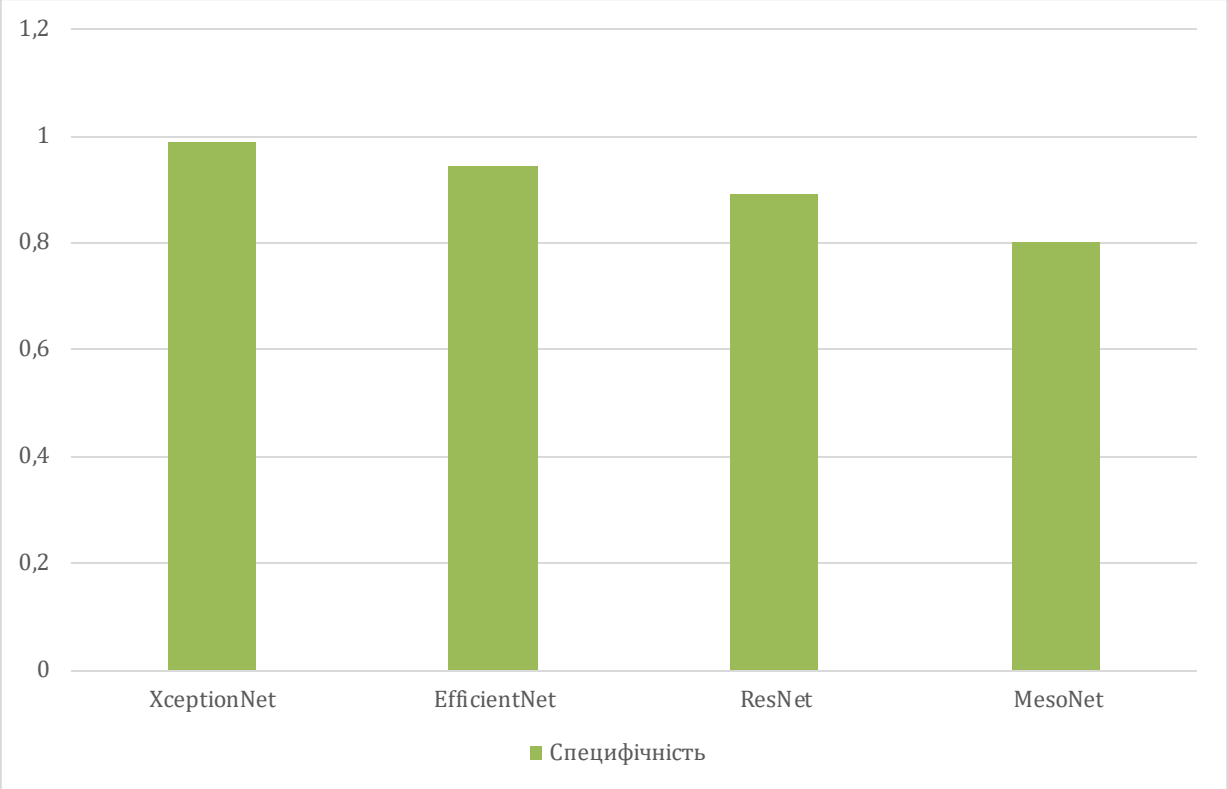

Рисунок 3.6 – Порівняння специфічності моделей

Із отриманих результатів можна побачити, що модель XceptionNet забезпечує найкращі результати порівняно з іншими. Вона може правильно знаходити маніпульовану область на підроблених відео, тоді як на реальних відео ми навряд чи помічаємо помилкові спрацьовування. Однак моделі EfficientNet та ResNet продемонстрували не набагато гірші результати.

В свою чергу модель MesoNet відстає по точності розпізнавання на 20%, що є дуже великим відхиленням. З цього можна зробити висновок, що CNN моделі із занадто малою кількістю шарів можуть бути непридатними для якісного виявлення підробок у відеоматеріалах.

Але з алгоритмом XceptionNet також є застереження – він може бути використаний і для поліпшення якості заміни обличчя, що ускладнить виявлення підробок. Крім того, як тільки запускається алгоритм виявлення підробки, шахраї завжди намагаються вдосконалити свою модель, щоб залишатися на крок попереду.

3.4 Рекомендації щодо захисту від Deepfake у корпоративному середовищі

Оскільки на сьогоднішній день ринок інформаційної безпеки не пропонує спеціальних рішень для захисту від Deepfake, для того щоб мінімізувати ризики таргетованих атак за допомогою використання цієї технології в корпоративному середовищі, є доречним виробити наступні рекомендації:

- використовувати багатофакторну автентифікацію співробітників,
- використовувати електронний підпис для захисту листів електронної пошти;
- здійснювати відстеження наявності програм для створення Deepfake на персональних комп'ютерах працівників компанії, а також спроби пошуку таких додатків в мережі Інтернет, звертати особливу увагу на подібних працівників і проводити в їх відношенні внутрішні перевірки;
- мінімізувати число комунікаційних каналів компанії;
- забезпечити узгоджене поширення інформації;
- обмежити фото- і відеоконтент за участю осіб, що займають керівні позиції на підприємстві;
- розробити план реагування на дезінформацію;
- організувати централізований моніторинг каналів і звітність;
- всередині компанії ввести практику використання усних паролів та кодових слів або контрольних питань, відповідь на які відома лише двом сторонам;
- стежити за новими способами виявлення Deepfake і методами боротьби з ними.

Висновки за розділом 3

У даному розділі було проведено дослідження із тестування та порівняння роботи чотирьох моделей виявлення Deepfake, що базуються на згорткових нейронних мережах: XceptionNet, EfficientNet, ResNet, MesoNet. Було досліджено такі параметри, як точність, чутливість та специфічність виявлення, а також матриці

помилок. В результаті проведеного тестування було виявлено, що найкращі результати показала мережа XceptionNet.

Останні підходи до виявлення підробок у відеоматріалах, як правило, стосуються проблеми заміни обличчя, і більшість завантажених підроблених відео належать до цієї категорії. Основні вдосконалення алгоритмів виявлення включають:

- ідентифікацію артефактів, залишених у процесі генерації, таких як невідповідність позі голови, відсутність моргання очей, кольорові варіації текстури обличчя та вирівнювання зубів;
- виявлення невидимих згенерованих зразків GAN;
- просторово-часові особливості;
- психологічні сигнали, такі як частота серцевих скорочень та моделі поведінки індивіда.

Незважаючи на те, що була представлена велика робота щодо автоматизованого виявлення, все ще потрібно вдосконалення:

- Існуючі методи не є надійними для таких операцій, як стиснення, шумові ефекти, світлові варіації тощо. Крім того, представлена обмежена кількість робіт, які стосуються виявлення як аудіо-, так і відео-Deepfake.
- Останнім часом більшість методів зосереджуються на виявленні заміни обличчя шляхом експлуатації обмежень цього методу, таких як видимі артефакти. Однак, з величезними технологічними досягненнями, найближче майбутнє призведе до більш складних замін обличчя. Окрім цього, інші типи глибоких підробок, такі як реконструкція обличчя та синхронізація губ, з кожним днем прогресують.
- Антикриміналістичні методи можуть бути використані для позначення оригінального відео як фальшивого за рахунок додавання імітованих ключових точок рівня сигналу, що використовуються існуючими методами ідентифікації, стан, який називають «Fake Deepfake»
- У існуючих наборах даних Deepfake відсутні потенційні атрибути, необхідні для оцінки ефективності більш надійних методів виявлення Deepfake. Дослідницьке співтовариство проігнорувало той факт, що відеоролики з глибокими підробками

містять не лише візуальні підробки, але й звукові маніпуляції. Існуючі масиви фальшивих даних не враховують підробку звуку, а зосереджуються лише на візуальних підробках.

ВИСНОВКИ

У дипломній роботі було висвітлено актуальність такої проблеми як активне поширення в інтернеті відео- та аудіоматеріалів, створених за допомогою технології Deepfake. Генеративні алгоритми глибокого навчання розвинулися до такого ступеня, що стало важко відрізнити реальне зображення від підробленого. У 2017 році було виявлено, наскільки легко використовувати цю технологію для неетичних та зловмисних цілей, таких як поширення дезінформації, уособлення політичних лідерів та наклеп на невинних людей. З тих пір технологія Deepfake значно просунулась.

Зважаючи на це, Було проаналізовано еволюціїю Deepfake, основні методи, які використовує ця технологія, та проблеми, що виникають при створенні підробленого мультимедійного контенту.

Було проаналізовано основні методи виявлення Deepfake, які базуються насамперед на виявленні невідповідностей у підроблених відеоматеріалах. Також було зазначено проблемні моменти, що виникають при спробах виявити Deepfake.

На основі аналізу було вирішено дослідити ефективність дії методів виявлення Deepfake, заснованих на згорткових нейронних мережах. Для досягнення поставленої мети роботи було обрано датасет із високоякісними підробками та проаналізовано чотири моделі виявлення.

За результатами дослідження можна зробити висновок, що на сьогоднішній день більшість алгоритмів із виявлення Deepfake зазнала суттєвих вдосконалень, наприклад у виявленні артефактів, просторово-часових невідповідностей тощо. Однак слід зважати, що досліджувані алгоритми можуть бути використані і для поліпшення якості заміни обличчя, що ускладнить виявлення підробок у майбутньому.

Практичне значення роботи полягає у полягає у аналітичному огляді та систематизації методів виявлення Deepfake із зазначенням проблемних моментів у їхньому функціонуванні, а також дослідженні та обгрунтованому запропонуванні

найкращого методу для виявлення підробок у відеоматеріалах. До того ж, з огляду на відсутність конкретних рішень на ринку ІБ по захисту від Deepfake, було вироблено рекомендації для захисту у корпоративному середовищі.

Таким чином, мету роботи досягнуто, поставлені задачі виконано.

Основні результати дипломної роботи доповідалися та обговорювалися на VII Міжнародній науково-практичній конференції «Information Technology and Interactions» (IT&I-2020) (Київ, 2020); IV Міжнародній науково-практичній конференції «Проблеми кібербезпеки інформаційно-телекомунікаційних систем» (Київ, 2021).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. WHAT YOU NEED TO KNOW ABOUT AIPOWERED “DEEP FAKES” IN ART CULTURE 17 December, 2019 [Електронний ресурс]. – Режим доступу: https://cuseum.com/blog/2019/12/17/3-things-you-need-toknow-about-ai-powered-deep-fakes-in-art-ampculture.

2. Dali lives (via artificial intelligence) [Електронний ресурс]. – Режим доступу: https://thedali.org/exhibit/dalilives/.

3. Positive Applications for Deepfake Technology [Електронний ресурс]. – Режим доступу: https://hackernoon.com/thelight-side-of-deepfakes-how-the-technology-can-be-usedfor-good-4hr32pp.

4. 10 deepfake usage examples that scared the Internet [Електронний ресурс]. – Режим доступу: https://www.creativebloq.com/features/deepfake- videos examples

5. Deepfake app is a reason for major concerns [Електронний ресурс]. – Режим доступу: bbc.com/news/technology-49570418

6. K. M. Malik, H. Malik, and R. Baumann. Analysis of susceptibility of voice-driven interface. *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 523-528: IEEE.

7. CEO was scammed for $243 000 using a voice deepfake [Електронний ресурс]. – Режим доступу: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=2e949f202241

8. Abraham Lincoln – Library of Congress [Електронний ресурс]. – Режим доступу: https://www.loc.gov/pictures/item/2003654314/

9. I. Goodfellow, Y. Bengio and A. Courville. Deep Learning. *Cambridge, MA: MIT Press*, 2016.

10. N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. *The Generalization Gap and Sharp Minima in Large-Batch Deep Learning Training* arXiv preprint arXiv:1609.04836, 2017

11. R. Caldeli, A. Bimbo. Deepfake video detection through optical flow based cnn. In The IEEE International Conference on Computer Vision (ICCV) Workshops, 2019.

12. J. Vincent. New AI deepfake app creates nude images of women. [Електронний ресурс]. – Режим доступу: https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnudenon-consensual-pornography

13. S. Arik et al. Neural text-to-speech in real time arXiv preprint arXiv:1702.07825, 2017.

14. Every six months, the amount of deepfake content doubles. [Електронний ресурс]. – Режим доступу: https://www.thehindu.com/sci-tech/technology/deepfake-content-is-doubling-every-six-months-trend-to-grow-in-asia/article32064299.ece

15. I. Goodfellow et al., Generative adversarial networks in Advances in neural information processing systems, 2014, pp. 2672-2680

16. B. Jan, L. Wong, and Y. Xiao. Deep bidirectional LSTM. in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4884-4888: IEEE.

17. J. Charles, D. Maggie, and D. Hopps. Characters from television series are being resurrected to achieve virtual immortality. in European Conference on Computer Vision, 2016, pp. 879-886: Springer.

18. S. Seitz, and I. Shlizerman. Learning lip sync from audio and synthesizing Obama. ACM Trans. Graph., vol. 36, no. 4, pp. 95:1-95:13, 2017.

19. A. Jamaludin, J. S. Chung, and A. Zisserman. Synthesising talking faces from audio. International Journal of Computer Vision, pp. 1-13, 2019.

20. M. Pantic. Temporal GANs for End-to-End Speech-Driven Realistic Facial Animation in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 37-40.

21. P. Garrido et al., Face video of performers is modified for realistic visual alignment to a dubbed audio track using Vdub. Computer graphics forum, 2015, vol. 34, no. 2, pp. 193-204: Wiley Online Library.

22. A. Jha, C. Jawahar. In the direction of automatic face-to-face translation. Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1428-1436.

23. R. Mukhopadhyay, C. Jawahar, For Speech to Lip Generation in the Wild, All You Need Is A Lip Sync Expert. Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484-492.

24. H. Kim et al., "Deepfake Portraits" ACM Transactions on Graphics (TOG), vol. 37, no. 4, p. 163, 2018.

25. C. Li, C. Loy, Reenactgan: Using boundary transfer to learn to reenact faces. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 603-619.

26. C. Theobalt, and M. Nießner. Face2face: Real-time face capture and rgb video reenactment. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2387-2395.

27. M. Liu, A. Tao, J. Kautz Conditional GAN for high-resolution picture creation and semantic modification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798-8807.

28. H. Nguyen et al. Video portraits with depth. ACM Graphics Transactions (TOG), vol. 37, no. 4, p. 163, 2018.

29. O. Wiles, A. Koepke. X2face is a network that uses photos, audio, and position codes to govern face generation. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 670-686.

30. A. Shysheya, V. Lempitsky. Learning realistic neural talking head models with a few shots of adversarial learning. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9459-9468

31. C. Zhang, C. Loy, Z. Liu. Reenactment of a single face. arXiv preprint arXiv:1908.03251, 2019.

32. K. Hao, A. R. Reibman, J. Derp. FaR-GAN Reenactment of a single face. arXiv preprint arXiv:2005.06402, 2020.

33. S. Ha S. Seo, B. Kim. Marionette: a recreation of a few shots of a face that preserves the identification of unseen targets. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, no. 07, pp. 10893-10900.

34. V. Blanz, T. Vetter. A morphable model for the creation of three-dimensional faces. Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999, pp. 187-194.

35. D. P. Kingma and M. Welling. Variational Bayes with auto-encoding. arXiv preprint arXiv:1312.6114, 2013

36. A. Radford, S. Chintala. Deep convolutional generative adversarial networks for unsupervised representation learning. arXiv preprint arXiv:1511.06434, 2015.

37. M. Liu, O. Tuzel. Paired adversarial generative networks. Advances in neural information processing systems, 2016, pp. 469-477.

38. T. Karras, S. Llaine. Gans are being grown in stages to improve quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

39. B. Karras, D. Aila. For generative adversarial networks, a style-based generator architecture is proposed. Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4401-4410

40. S. Llaine, J. HelstanT. Aila. Analyzing and improving stylegan's image quality. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110-8119.

41. J. Trueba, X. Wang, T. Kinnen. Can we clone your voice using GAN, WaveNet, and low-quality found data? An first inquiry on cloning Obama's voice using GAN, WaveNet, and low-quality found data. arXiv preprint arXiv:1803.00860, 2018.

42. G. Mysore, J. Liu. Voco: a text-based insertion and replacement system ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1-13, 2017.

43. B. Silsman, L. Zao, and P. Lio. DeepConversion: Voice conversion using only a small amount of parallel training data. Speech Communication, 2020.

44. P. Partila, J. Tovarek, G. H. Ilk, J. Rozhon, and M. Voznak. How Vulnerable Are Automatic Speaker Verification Systems to Spoofing Trials? Deep Learning Serves

Voice Cloning: How Vulnerable Are Automatic Speaker Verification Systems to Spoofing Trials? IEEE Communications Magazine, vol. 58, no. 2, pp. 100-105, 2020.

45. A. Gokale, P. Mulay, D. Pramod, and R. Kulkarni. A bibliometric analysis of digital image forensics. Science & Technology Libraries, 39(1):96–113, 2020

46. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: a large-scale video dataset for forgery detection in human faces, 2018. arXiv: 1803.09179 [cs.CV].

47. N. Neverova, R. Alp Guler, and I. Kokkinos, "Transfering of dense poses" in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 123-138.

48. D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping is the process of replacing people's faces in images mechanically" in ACM Transactions on Graphics (TOG), 2008, vol. 27, no. 3, p. 39: ACM.

49. E. A. AlBadawy, S. Lyu, and H. Farid, "Using Bispectral Analysis to Detect AI-Synthesized Speech" in CVPR Workshops, 2019, pp. 104-109.

50. P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Denoising autoencoders are used to extract and compose strong features" in Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096-1103.

51. R. Faris, H. Roberts, B. Etling, N. Bourassa, E. Zuckerman, and Y. Benkler, "Online media and the 2016 US presidential election: partisanship, manipulation, and disinformation" Berkman Klein Center Research Publication, vol. 6, 2017

52. E. Sanchez and M. Valstar, "In GAN-based face synthesis, there is a triple consistency loss for pairing distributions" arXiv preprint arXiv:1811.03492, 2018.

53. G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez. Image editing with invertible conditional gans. arXiv preprint arXiv:1611.06355, 2016.

54. T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Deepfake Detection: Learning to Recognize Patch-Wise Consistency arXiv preprint arXiv:.09311, 2020.

55. G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. A. Ranzato, "Images are manipulated using fader networks, which use sliding properties to

manipulate them" in Advances in neural information processing systems, 2017, pp. 5967-5976.

56. Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Image synthesis in a variety of disciplines in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188-8197

57. E. A. AlBadawy, S. Lyu, and H. Farid. Bispectral Analysis for Detecting AI-Synthesized Speech in CVPR Workshops, 2019, pp. 104-109.

58. X. Yang, Y. Li, and S. Lyu. Using inconsistencies in head positions to expose deep fakes in ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8261-8265: IEEE.

59. N. Yu, L. S. Davis, and M. Fritz. Learning and evaluating GAN fingerprints to attribute false pictures to GANs in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7556-7566.

60. Y. Lin, Q. Lin, F. Tang, and S. Wang. Replacement of the face with large-pose differences in Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 1249-1250: ACM

61. F. Perez, S. Avila, E. Valle. Is it better to perform solo or in an ensemble? Melanoma Classification: Choosing a CNN Architecture, arXiv preprint arXiv:1904.12724, 2019.

62. L. N. Smith. Training Neural Networks with Cyclical Learning Rates arXiv preprint arXiv:1506.01186, 2017.

63. K. Simonyan and A. Zisserman. Large-Scale Image Recognition using Very Deep Convolutional Networks arXiv preprint arXiv:1409.1556, 2015

64. A. Zheng. Evaluating Machine Learning Models. O'Reilly Media, Inc., 2015.

65. F. Hirumare (2017) Deep learning with depthwise separable convolutions (Xception). *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1251–1258

66. B. Bayar and M. C. Stamm. Using a new convolutional layer, a deep learning technique to universal picture modification detection has been developed. *Proceedings of*

the 4th ACM Workshop, *IH&MMSec* '16, 5–10, Vigo, Galicia, Spain. Association for Computing Machinery, 2016.

67. M. Tan and Q. Le. Rethinking model scaling for convolutional neural networks with EfficientNet. *Proceedings of Machine Learning Research*, pages 6203–6105, Long Beach, California, USA, 2019.

68. D. Althar, V. Vorick,. Mesonet: aface video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1– 7, 2018.

69. Що таке матриця помилок [Електронний ресурс]. – Режим доступу: https://learnmachinelearning.wikia.org/ru/wiki/%D0%9C%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0_%D0%BE%D1%88%D0%B8%D0%B1%D0%BE%D0%BA_(Confusion_matrix)

70. M. Hossin & S. M.N. Evaluation Metrics for Data Classification Evaluations : A Review 2015 *International Journal of Data Mining & Knowledge Management Process*. 5. 01-11. 10.5121/ijdkp.2015.5201.

71. A. Zheng. Evaluating Machine Learning Models. *O'Reilly Media, Inc.,* 2015.

72. P. Sun, H. Qi, and S. Lyu, Celeb-df: A new dataset for deepfake forensics. arXiv preprint arXiv:1709.12373, 2019.

73. R. Wang, L. Ma, F. Xu, . 2019. Fakespotter: detecting artificial intelligence-generated phony faces. arXiv preprint arXiv:1909.06122 (2019).