Харківський національний університет імені В.Н. Каразіна

Факультет комп'ютерних наук

Безпека інформаційних систем і технологій

«Допущено до захисту»

Зав.кафедрою БІСТ

Сватовський І.І.

« » червня 2023р.

Пояснювальна записка

до кваліфікаційної роботи бакалавра

спеціальність: 125 Кібербезпека

на тему: «Дослідження технології deepfake та методів захисту від підробки відеозображень»

оцінка « »

Голова ЕК

Лемешко О.В._____

Керівник Сватовський І.І.
(прізвище та ініціали/підпис)

Рецензент Бакуменко Н.С.
(прізвище та ініціали/підпис)

Виконавець студент групи КБ-42

Чепель Д.О.
(прізвище та ініціали/підпис)

Харків – 2023

РЕФЕРАТ

Пояснювальна записка містить: 50 сторінок, 2 рисунка
17 використаних джерел.

Метою дипломної роботи є узагальнення відомостей, щодо технології діпфейк та методів захисту від підробки відеозображень за допомогою цієї технології та створення рекомендацій щодо забезпечення захисту від поширення дезінформації шляхом використання діпфейків.

Об’єктом дослідження дипломної роботи є технології і засоби протидії поширенню дезінформації за допомогою технології діпфейк.

Предметом дослідження є особливості інтеграції та використання методів та заходів із боротьби із діпфейками.

Методи дослідження: аналіз та порівняння.

В роботі досліджено питання поширення дезінформації за допомогою технології діпфейк. Проведено аналіз існуючих технологій та методів виявлення діпфейків, автентифікації контенту та перешкоджання поширенню діпфейків. Визначено, що захист від поширення дезінформації за допомогою діпфейків можна забезпечувати різними способами, такими як використання інструментів виявлення діпфейків, встановлення методів визначення автентичності відео та перешкоджання розповсюдженню діпфейків.

Результатами проведеної роботи є рекомендації щодо боротьби із поширенням дезінформації за допомогою діпфейк можуть бути використані в освітніх цілях та у якості довідкового матеріалу.

Ключові слова: ДІПФЕЙК, НЕЙРОННІ МЕРЕЖІ, ДЕЗІНФОРМАЦІЯ, АВТЕНТИФІКАЦІЯ КОНТЕНТУ, ЗАХИСТ ВІД ПІДРОБКИ ВІДЕОЗОБРАЖЕНЬ.

ABSTRACT

The explanatory note to the mast project contains: 50 pages, 2 figures, 17 source references.

The purpose of the work is a generalization of information about deepfake technology and methods of protection against forgery of video images using this technology and to create recommendations for ensuring protection against the spread of disinformation through the use of deepfakes.

The object of research is technologies and means of countering the spread of disinformation using deepfake technology.

Subject of research is the features of the integration and use of methods and measures to combat deepfakes.

The main research methods are analysis and comparison.

The paper examines the issue of spreading disinformation using deepfake technology. An analysis of existing technologies and methods for detecting deepfakes, authenticating content and preventing the spread of deepfakes was carried out. It has been determined that protection against the spread of misinformation by means of deepfakes can be ensured in various ways, such as using deepfake detection tools, establishing methods for determining the authenticity of videos, and preventing the spread of deepfakes.

The results of the work are recommendations for combating the spread of disinformation with the help of deepfakes, which can be used for educational purposes and as reference material.

Keywords: DEEPFAKE, NEURAL NETWORKS, DISINFORMATION, CONTENT AUTHENTICATION, PROTECTION AGAINST FORGERY OF VIDEO IMAGES.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ ТА СИМВОЛІВ

ЗНМ	-	Згорткова нейронна мережа
РНМ	-	Рекурентна нейронна мережа
ГЗМ	-	Генеративно-змагальна мережа
CapsNet	-	Капсульна мережа
VGGNet	-	Visual Geometry Group Network
ResNet	-	Residual Neural Network
LSTM	-	Long Short-Term Memory
RSA	-	Rivest-Shamir-Adleman
ECDSA	-	Elliptic Curve Digital Signature Algorithm
EdDSA	-	Edwards-curve Digital Signature Algorithm
SHA-256	-	Secure Hash Algorithm 256-bit
MD5	-	Message Digest Algorithm 5
SHA-3	-	Secure Hash Algorithm 3

ВСТУП

У сучасному світі технології стали невід'ємною частиною нашого життя. Велику роль грає отримання інформації про події у світі через інтернет-медіа, будь то новини чи інформація, що поширюється соціальними мережами. Традиційно надійним доказом автентичності подій були відеозображення, але тепер вони поставлені під сумнів завдяки новим технологіям у сфері редагування відео.

Розробки в галузі штучного інтелекту, комп'ютерного зору та машинного навчання призвели до появи технології діпфейк. Сьогодні за допомогою різних сервісів та програм будь-хто може згенерувати контент, який імітує написаний людиною текст, фіктивні зображення та відео, які складно відрізнити від справжніх або повністю скопіювати голос людини та сказати ним своє повідомлення. Найбільш небезпечним типом діпфейків є діпфейк-відео, яке складно відрізнити від справжнього запису людини та яке знаходиться у форматі, який традиційно було підробити найважче – відео. Діпфейки можуть створювати загрози демократії, міжнародним відносинам, національній безпеці та честі людей, тому питання методів засобів виявлення діпфейків та захисту від підробки відеозображень є релевантним. Потрібно розуміти, як працює ця технологія та які є способи захисту від неї, щоб забезпечити довіру до відеодоказів у судових процесах, новинах та інших галузях життя.

Предметом вивчення цієї роботи є методи захисту від підробки відеозображень за допомогою технології діпфейк та методи перешкоджання поширенню діпфейків. Крім того будуть розглянуті застосування цієї технології та загрози, що можуть викликати діпфейки, їх типи і методи створення.

Метою роботи є створення рекомендацій щодо забезпечення захисту від поширення дезінформації шляхом використання діпфейків

1 АНАЛІЗ ТЕХНОЛОГІЇ ДІПФЕЙК

Діпфейк - це технологія створення синтетичного контенту, який був змінений за допомогою цифрової обробки. Термін "deepfake" походить від англійських слів "deep learning" і "fake". За допомогою діпфейк можна створити відео або інший контент з участю будь-якої людини, навіть якщо ця людина не була присутня під час запису. Контент, що створюється може бути використаний із благим наміром, так і зі злочинним, наприклад поширення фальшивих новин, дезінформації або дискредитації.

Технологія діпфейк базується на глибокому навчанні, а саме на нейронних мережах, що дає можливість створювати реалістичні образи, звуки та текст. Для створення діпфейків зазвичай використовуються відкриті програмні бібліотеки, такі як TensorFlow, Keras або PyTorch, що дозволяє багато розробникам створювати власні діпфейк-моделі.

Однією з особливостей діпфейків є їхня здатність швидко еволюціонувати та покращуватись завдяки постійному вдосконаленню навчання та розробці нових моделей. Діпфейки є продуктом Генеративно-змагальних мереж (Generative Adversarial Networks, GAN), тобто двох нейронних мереж, які працюють разом, щоб створити реалістичне медіа. Ці дві мережі, які називаються «генератор» і «дискримінатор», навчаються на одному наборі даних зображень, відео або звуків. Перша намагається створити нові зразки медіа з ціллю обдурити другу мережу, яка намагається визначити реальність наданої їй медіа. Таким чином вони спонукають один одного до вдосконалення. Одним з ключових етапів при створенні діпфейків є навчання моделі на великій кількості даних, які складаються з зображень, звуків або текстів. Потім модель використовують для генерації нового контенту. Якість діпфейків значно залежить від якості вхідних даних, на основі яких модель навчалась [1].

Технологія діпфейк з'явилася досить недавно, у 2017 році, коли користувачі Інтернету почали створювати відео, де обличчя знаменитостей були замінені на обличчя інших людей за допомогою інструментів машинного навчання, як на рисунку 1.1. Пізніше у тому ж році, почали застосовувати нейронні мережі для створення більш складних діпфейк-відео, де зображення знаменитостей були замінені на зображення інших знаменитостей або на вигадані образи. Це був перший крок в розвитку технології діпфейк.

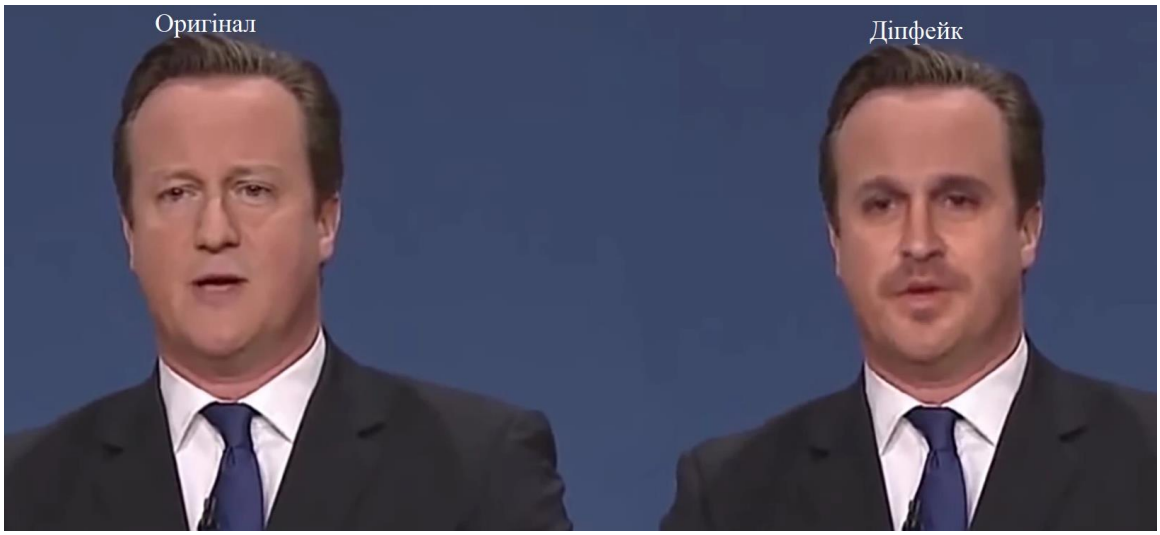

Рисунок 1.1 – Змінення обличчя за допомогою технології діфпфейк

Найбільш поширеним способом поширення діпфейків є їх розміщення у соціальних мережах. Користувачі діляться такими відео між собою, що призводить до їх швидкого поширення. Крім того, зловмисники можуть використовувати спеціальні програми для автоматичного поширення діпфейків через багато профілів, що підвищує їх популярність та вплив.

Нині технологія діпфейк продовжує розвиватися, і все більше компаній та організацій займаються розробкою інструментів для виявлення діпфейк контенту. Також багато держав прийняли закони, націлені на боротьбу із використанням технології діпфейк для створення фальшивих відео та іншого контенту для використання із злим наміром.

1.1 Переваги технології діпфейк

Діпфейк-технологія має позитивні застосування в багатьох галузях, включаючи кіно, навчальні та соціальні фільми, ігри та різні сфери бізнесу, такі як мода та електронна комерція. У кіноіндустрії технологія діпфейк може допомогти створити цифрові голоси для акторів, які втратили свій голос, покращити якість звуку у старих фільмах, створювати нові фільми із померлими акторами, спростити створення спецефектів, а також покращувати аматорські відео до професійної якості. Також згенеровані технологією діпфейк зображення можуть прискорити та здешевити розробку відеоігор.

За допомогою технології діпфейк можна дешево та за короткий термін створювати реалістичні дубляжі для фільмів, аудіокниги, озвучування для відеоігор та анімаційних фільмів будь-якою мовою. За допомогою діпфейк-технології можуть бути створені розумні помічники, що мають природне звучання та зовнішній вигляд.

Діпфейки можуть використовуватися у модельному бізнесі, дозволяючи брати участь людям із неідеальними зовнішніми характеристиками, або відтворити велику кількість типів зовнішності, використовуючи невелику кількість людей. Також діпфейки можуть використовуватися у електронній комерції одягу як спосіб дати покупцю можливість примірити товар перед його придбанням.

Текстові діпфейки можуть бути використані для швидкого та ефективного створення контенту або автоматизації процесу написання тексту. Замість того, щоб витрачати час на написання тексту з нуля, письменник може використовувати текстові діпфейки, щоб отримати базовий текст та який він потім редагує під свої потреби. Або якщо компанія має великий обсяг рутинних завдань з написання тексту, використання текстових діпфейків може допомогти прискорити процес та зменшити ризик помилок.

1.2 Загрози злонамірeного використання технології діпфейк

Діпфейки є серйозною загрозою для суспільства, політичної системи та бізнесу, тому що вони чинять тиск на журналістів та користувачів Інтернету, які намагаються відфільтрувати справжні новини від фальшивих, загрожують національній безпеці, поширюючи пропаганду та втручаючись у вибори, перешкоджають довірі громадян до інформації з боку влади та порушують питання кібербезпеки.

Діпфейки становлять більшу загрозу, ніж «традиційні» фальшиві новини, оскільки їх важче помітити. Технологія дозволяє створювати, здавалося б, справжні новинні відео, які ставлять під загрозу репутацію журналістів і ЗМІ, бо гонка серед ЗМІ за доступ до відеоматеріалів та бажання бути першими, хто висвітлив матеріал може призвести до недостатньої перевірки матеріалів на достовірність.

Діпфейки можуть бути використані з ціллю загрози національній безпеці шляхом поширення політичної пропаганди та зриву виборчих кампаній. Вірусне відео, у якому кандидат говорить словами зловмисника або бере хабар, може спотворити думку виборців. Іноземні агентства можуть виготовити фальшиве відео скоєння військових злочинів або політика, що визнає план змови. Такі фальшиві відео не тільки можуть викликати внутрішні заворушення або зірвати вибори, але й призвести до міжнародних конфліктів.

Діпфейки перешкоджають цифровій грамотності громадян, підривають довіру до наданої владою інформації та інформації в Інтернеті в цілому. Найбільш шкідливим аспектом діпфейків може бути не дезінформація сама по собі, а те, як постійний контакт з дезінформацією змушує людей відчувати, що більшості інформації, включно з відео, просто не можна довіряти, що призводить до явища, яке називають «інформаційним апокаліпсисом» або «апатією до реальності».

Технологія діпфейк може бути використана для шахрайства. Наприклад для маніпулювання ринками та акціями, зображуючи керівника фірми, який робить неправдиві заяви про банкрутство. Діпфейки можуть використовуватися для

виготовлення матеріалів для шантажу чи дискредитації. Технологія дозволяє видавати себе за іншу особу у реальному часі, що може бути використано з метою попросити знайомого жертви зробити грошовий переказ або розголосити конфіденціальну інформацію [2].

1.3 Типи діпфейків

Найбільш відомим типом діпфейків є відео, але діпфейки можуть також бути використані в інших контекстах [3]. У цьому розділі роздивимося типи діпфейків та їх особливості.

- Текстові діпфейки

Технології штучного інтелекту можуть складати текст, створювати статті, вірші, блоги та інші твори. GPT від OpenAI є прикладом системи генерування тексту, яка дозволяє користувачам зручно створювати масиви тексту шляхом введення заголовка чи теми. Технологія глибокого навчання дозволяє цій системі створювати схожий на написаний людиною текст, розуміти мовні моделі та адаптуватися під потреби користувача.

- Відео діпфейки

Серед усіх типів найбільш поширеними є дипфейк-відео, це реалістичні відео, що згенеровані за допомогою штучного інтелекту та технології редагування відео. Маючи лише фотографію, різні програми можуть дозволити користувачам замінити існуючу особу у відео обличчям іншої людини або накласти рухи із відео на фотографію.

- Діпфейк зображення

Діпфейк зображення можуть бути використані для автоматизації процесу фотомонтажу, або для виконання основної частини роботи. Вони можуть бути використані з тими ж цілями, що і діпфейк-відео, або для полегшення роботи обробників фотографій.

- Діпфейк аудіо

Діпфейк-аудіо - це штучно створені аудіофайли, які звучать, як справжні записані голоси, але насправді вони створені технологією діпфейк. Алгоритм працює, створюючи штучні звукові хвилі, які потім обробляються для створення реалістичної мови.

У цьому розділі була розглянута сутність технології діпфейк, її застосування та ризики, що пов’язані з її злонамірeним застосуванням. Також були розглянуті існуючі типи діпфейків та їх застосування. Було встановлено, що технологія може бути корисна при виконанні різноманітних завдань, але вона може нести небезпеку при її злонамірeному використанні

1 АНАЛІЗ МЕТОДІВ ВИЯВЛЕННЯ ДІПФЕЙКІВ

Одним із основних методів захисту від діпфейків є їх виявлення. Існує різниця у складності виявлення між діпфейками різної якості. Якість діпфейків може змінюватись в залежності від різних факторів, таких як використовувані алгоритми, вхідні дані, навички та технічні можливості. Діпфейки високої якості, створені з використанням передових алгоритмів та великої кількості навчальних даних, можуть бути дуже реалістичним. Вони можуть імітувати міміку, освітлення, текстури та інші деталі, що робить їх візуально практично відмінними від оригінального матеріалу. З іншого боку, діпфейки низької якості, створені з використанням менш точних алгоритмів або обмеженої кількості навчальних даних можуть мати явні артефакти, невідповідності в рухах або інші візуальні особливості. У цьому розділі роздивимося методи виявлення діпфейків.

2.1 Традиційні методи виявлення обробки відео

Хоча деякі методи, які були ефективні для виявлення підроблених відео в минулому, є менш ефективними в контексті сучасних діпфейків, це не означає, що традиційні методи виявлення перестають бути корисними. У деяких випадках, особливо якщо для створення діпфейків були використані низькоякісні інструменти та техніки підробки, традиційні методи можуть виявляти підроблені відеозаписи. На таких діпфейках низької якості можуть з’являтися артефакти, помітні неозброєному оку, як на рисунку 2.1. Крім того, комбінування традиційних методів із сучасними алгоритмами машинного навчання може покращити ефективність виявлення. У цьому розділі розглянемо традиційні методи виявлення обробки відео, які можуть бути використані при виявлення діпфейків.

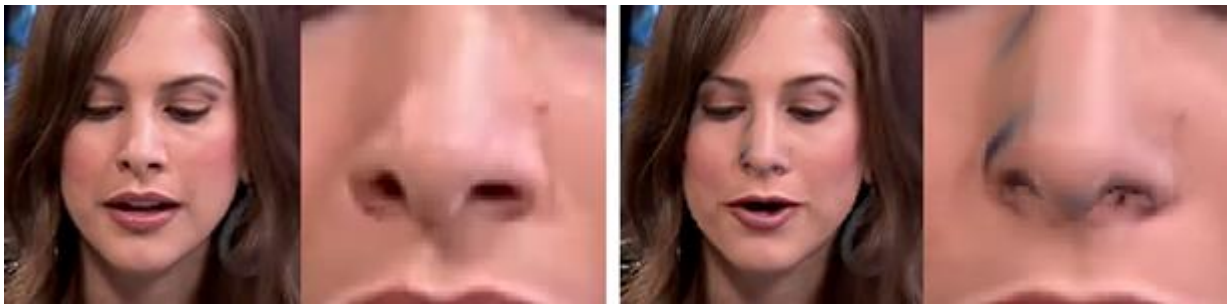

Рисунок 2.1 – Видимі артефакти у діпфейк зображенні (зліва – оригінал)

- Аналіз артефактів стиснення

Сучасні алгоритми стиснення, такі як JPEG, поділяють зображення на блоки та застосовують стиск до кожного блоку. Якщо зображення було підроблено, деякі блоки можуть виглядати інакше, ніж решта зображення. При стисненні зображень можуть виникати артефакти у колірному просторі та артефакти шуму, тому підроблені області можуть мати незвичайні зміни кольору, насиченості, яскравості, характеристик шуму. Якщо фрагмент зображення має нижчу роздільну здатність або менш різкі краї порівняно з рештою зображення, це також може також свідчити про його фальсифікацію.

- Аналіз метаданих

Відеофайли зазвичай містять метадані, такі як дата та час створення, пристрій створення, роздільна здатність та інша інформація. Наприклад, якщо метадані дати відео не відповідають тому, коли відео буди нібито знято, це може вказувати на підробку. Однак метадані можуть бути змінені або видалені, тому надійність цього методу мала.

- Аналіз розбіжностей візуальних властивостей

Діпфейк-відео можуть мати невідповідності у візуальних властивостях, таких як освітлення, тіні, перспектива або різкість, які не узгоджуються з оригінальними сценами чи контекстом. Аналіз цих розбіжностей допоможе виявити підроблене відео.

- Аналіз руху

Візуальні алгоритми, які використовуються створення діпфейків, можуть мати обмеження у точності моделювання природних рухів. Підроблені відео можуть виявляти аномальний або неприродний рух об'єктів або осіб, які можна виявити під час аналізу змін позицій об'єктів.

- Аналіз акустичних характеристик

Якщо відеозаписи містять мову або інші звукові ефекти, можуть мати місце звукові аномалії, або невідповідності між звуками та зображенням, що може вказувати на можливість підробки.

2.2 Методи виявлення діпфейків за допомогою глибокого навчання

У цьому підрозділі будуть розглянуті методи виявлення діпфейків, що ґрунтуються на глибокому навчанні – галузі штучного інтелекту, яка займається розвитком алгоритмів та моделей, здатних автоматично вивчати представлені дані та робити прогнози. Використання глибокого навчання корисне при виявленні діпфейків, завдяки його здатності автоматично виявляти складні закономірності у зображеннях або відео [4][5][6][7].

2.2.1 Згорткові нейронні мережі

Згорткові нейронні мережі (ЗНМ) - це клас моделей глибокого навчання, які демонструють високу ефективність у різних завданнях комп'ютерного зору, таких як класифікація зображень, виявлення об'єктів та сегментація зображень. ЗНМ спеціально розроблені для обробки структурованих сітчастих даних, таких як зображення, шляхом використання процесу згортки.

ЗНМ складаються з декількох шарів, включаючи згортковий шар, агрегувальний шар, повноз'єднаний шар та шар втрат. Згорткові шари відповідають за навчання та витягування важливих ознак із зображень. Ці шари застосовують навчальні фільтри, також відомі як детектори ознак, до вхідного зображення за допомогою операцій згортки. Шляхом згортки фільтрів по зображенню ЗНМ можуть захоплювати локальні шаблони та просторові зв'язки між сусідніми пікселями.

Агрегувальні шари, як правило, вставляються після згорткових шарів, щоб зменшити просторові розміри вивчених ознак. Вони виконують операції зменшення розміру, такі як максимізаційне агрегування, яке вибирає найрепрезентативніші ознаки з регіону та відкидає решту. Це допомагає зменшити обчислювальну складність та робить вивчені ознаки більш незалежними від невеликих просторових змін.

Повністю з'єднані шари в кінці ЗНМ відповідають за прийняття рішень на основі вивчених ознак. Ці шари беруть ознаки, отримані зі згорткових та агрегувальних шарів, зважують їх та видають результат.

Шар втрат є останнім шаром моделі, який відповідає за обчислення функції втрати. Функція втрати визначає, наскільки точні прогнози моделі порівняно зі справжніми значеннями цілей, тобто оцінює відхилення між передбаченими та очікуваними результатами. Ціллю функції втрати є покарання моделі за некоректні прогнози та стимулювання її до кращих результатів.

ЗНМ можуть використовуватися для виявлення діпфейків завдяки їх здатності вивчати та розпізнавати шаблони. Один з підходів - навчити ЗНМ на великих наборах даних, які включають як реальні, так і фальшиві відео. Навчаючи модель на різноманітних автентичних та підроблених відео, вона може навчитися впізнавати характеристичні риси та артефакти, пов'язані з діпфейками. ЗНМ може виявляти аномалії в освітленні, текстурі, розмірі або взаємному розташуванні об'єктів на зображенні, які можуть бути ознаками діпфейку [8].

Також ЗНМ можуть використовуватися для аналізу специфічних артефактів, які можуть виникати в результаті обробки або маніпуляцій зображення. Наприклад, певні операції маніпуляції, такі як розмиття, клонування або замазування, можуть залишати сліди, які можна виявити за допомогою ЗНМ.

2.2.2 Рекурентні нейронні мережі

Рекурентні нейронні мережі (РНМ) - це тип штучних нейромереж, які використовуються для аналізу даних з урахуванням послідовного контексту. РНМ можуть запам'ятовувати попередні стани та використовувати їх для обробки нових вхідних даних. Це робить їх ефективними для роботи з послідовними, такими як мова, музика та інше.

Основний елемент РНМ - це рекурентний шар, який дозволяє передавати інформацію з одного часового кроку до наступного. На кожному кроці РНМ обчислює новий стан, використовуючи поточний вхідний елемент і попередній стан. Цей новий стан стає вхідним станом для наступного кроку. Для обчислення нового стану використовуються алгоритми, які навчаються під час тренування моделі. Це дозволяє РНМ враховувати контекстуальну інформацію з попередніх елементів послідовності.

Крім обчислення нового стану, РНМ також генерує вихідні значення на кожному кроці, які можуть бути використані для вирішення конкретних задач, наприклад, класифікації або прогнозування. Вихідні значення залежать від поточного стану та можуть передаватись на наступні кроки або використовуватись для оцінки результатів.

Одна з особливостей РНМ полягає у використанні зворотного зв'язку, що дозволяє передавати інформацію від кінця до початку послідовності. Це особливо корисно при аналізі послідовностей, оскільки дозволяє моделі враховувати всю доступну історію для прийняття рішень.

РНМ може бути навчений розпізнавати неприродні зміни у русі та вигляді обличчя. Він може аналізувати послідовність кадрів і порівнювати їх зі знанням про реальний рух та зміну вигляду обличчя. Наприклад, якщо обличчя на відео проявляє незвичайні зміни форми або виразу, це може свідчити про наявність синтезованого відео.

Також РНМ може бути використано для аналізу аудіо та виявлення незвичайних шаблонів або аномалій, що вказують на можливість присутності підробки. Мережа може аналізувати зміни звукових даних і виявляти аномальні шуми, артефакти або несумісні з реальними записами звуки [9].

2.2.3 Генеративно-змагальні мережі

Генеративно-змагальна мережа (ГЗМ) - це модель машинного навчання, що складається з двох нейромереж: генератора і дискримінатора. Метою генератора є створення реалістичних даних, тоді як дискримінатор намагається розрізнити згенеровані дані від реальних даних тренувального набору. Генератор і дискримінатор тренуються разом у конкурентному процесі. Генератор спочатку створює випадкові зразки, а дискримінатор навчається класифікувати їх як реальні чи фальшиві. По мірі тренування генератор покращує свою здатність генерувати реалістичні зразки, отримуючи відгук від дискримінатора. Процес триває до тих пір, поки генератор не зможе створювати зразки, які важко відрізнити від реальних даних.

Якщо розглядати ГЗМ як засіб створення діпфейків, то інтерес має мережа-генератор, а при обговоренні методів виявлення діпфейків головна роль відводиться дискримінатору. Дискримінатор відіграє ключову роль у виявленні діпфейків за допомогою ГЗМ. Він навчається розрізняти оригінальні зображення від згенерованих генератором. Аналізуючи відмінності між двома типами зображень, дискримінатор може виявити характеристики, які є характерними лише для діпфейків. Оригінальні та згенеровані зображення мають різні набори ознак, такі як текстури, кольори і форми. Аналіз цих ознак дозволяє виявити аномалії, які можуть вказувати на наявність діпфейків. Деякі методи виявлення діпфейків комбінують інформацію від кількох дискримінаторів, використовуючи різні архітектури та ознаки для більш точного виявлення [10][11].

Щоб забезпечити ефективність виявлення діпфейків, важливо постійно оновлювати навчальний набір даних з реальними відео та новими типами

діпфейків, що постійно з'являються. Це дозволяє моделі підтримувати актуальність та ефективність.

2.2.4 Механізм уваги

Механізм уваги є концепцією, широко використовуваною в глибокому навчанні та обробці природної мови. Він був розроблений для покращення продуктивності моделей sequence-to-sequence, таких як машинний переклад, дозволяючи моделі фокусуватися на важливих частинах послідовності входу під час генерації виходу. Механізм уваги дозволяє моделі приділяти різні рівні важливості або уваги різним частинам послідовності вводу. Замість рівномірної обробки усієї вхідної інформації, модель вчиться надавати вагу кожному елементу вводу на основі його важливості для поточного кроку генерації виводу. Ця вага вказує на важливість або увагу, надану кожному елементу вводу, і вони використовуються для обчислення суми ваги елементів вводу. Ця сума потім поєднується зі внутрішнім станом моделі для генерації виходу.

Механізм уваги довів свою ефективність у різних завданнях обробки природної мови, включаючи машинний переклад, підсумовування тексту та аналіз тональності. Він дозволяє моделі уловлювати довгострокові залежності та фокусуватися на найважливішій інформації для виконання завдання. Завдяки увазі модель стає здатною краще розуміти контекст і генерувати точні та зв'язні виходи.

У контексті боротьби з діпфейками механізми уваги можуть застосовуватися декількома способами. Наприклад, для виявлення діпфейків на відео можна використовувати механізм уваги для аналізу певних ознак обличчя, таких як рухи очей, синхронізація рухів губ або текстури шкіри, які важко точно відтворити у діпфейках. Аналогічно, при виявленні діпфейків механізми уваги можуть застосовуватися для дослідження деталей, таких як текстурні візерунки, нерегулярності на рівні пікселів або некоректності в освітленні та тінях. Звертаючи увагу на ці важливі візуальні ознаки, модель може виявити сліди маніпуляції і надати надійну оцінку автентичності зображення [12].

2.2.5 Капсульні нейронні мережі

Капсульні мережі, також відомі як CapsNets, є типом архітектури глибинного навчання, запропонованої як альтернатива традиційним згортковим нейронним мережам (ЗНМ). Капсульні мережі спрямовані на вирішення деяких обмежень ЗНМ, зокрема їх нездатності ефективно ефективно вловлювати просторові ієрархії та обробляти варіації точок зору.

У CapsNet базовим будівельним блоком є капсула, яку можна розглядати як групу нейронів, які представляють певну сутність або концепцію. Кожна капсула відповідає за виявлення та представлення присутності сутності у вхідних даних. На відміну від нейронів у ЗНМ, капсули мають векторні значення та видають не лише активацію, але й параметри екземплярів, такі як положення, масштаб та орієнтація. Це робить їх більш стійкими до варіацій у вхідних даних.

Капсули організовані у ієрархічні шари, де капсули більш високого рівня представляють складніші концепції, що складаються з капсул нижчого рівня. Зв'язок між капсулами є динамічним і визначається механізмом маршрутизації, який ґрунтується на згоді між капсулами нижнього і вищого рівнів. Ця динамічна маршрутизація дозволяє мережі вивчити зв'язки між сутностями та їх просторовими конфігураціями.

Використовуючи здатність мережі відтворювати просторові ієрархії та моделювати зв'язки між сутностями, капсульні мережі можуть навчитися виявляти основні структури та патерни у реальних і фальшивих зображеннях. Вони можуть фіксувати внутрішні властивості різних складових обличчя, таких як очі, ніс та рот, і вивчати їх просторові відношення [13].

Застосування капсульних мереж для виявлення діпфейків вимагає великої кількості навчальних даних, які включають як реальні, так і синтетичні зображення. Мережа може бути постійно вдосконалюваною, здатною адаптуватися до нових типів діпфейків, які постійно з'являються.

2.2.6 Автокодувальник

Автокодувальник є типом архітектури нейронних мереж, які часто використовуються для навчання без вчителя. Основною метою автокодувальника є навчання стискання вхідних даних, шляхом кодування їх в простір меншої розмірності і подальшого відтворення початкових даних з цього стислого представлення.

Автокодувальник складається з двох основних компонентів: Кодувальника та декодувальника. Кодувальник бере вхідні дані і відображає їх у простір меншої розмірності, також відомий як шар затримки. Цей шар затримки містить стисле представлення вхідних даних. Декодувальник бере це стисле представлення і відтворює початкові вхідні дані з нього.

Під час процесу навчання автокодувальник прагне мінімізувати помилку відтворення, яка є різницею між початковими вхідними даними та відтвореним виводом. Таким чином, автокодувальник навчається захоплювати найважливіші риси вхідних даних у стислому представленні.

Один із підходів до використання автокодувальника для виявлення діпфейків полягає в навчанні на великому наборі аутентичних зображень або відео. Автокодувальник вчиться кодувати і відтворювати ці аутентичні вхідні дані. Після того, як автокодувальник навчений, його можна використовувати для виявлення діпфейків. Якщо автокодувальник отримує діпфейкове зображення або відео, він може мати проблеми з точним відтворенням через різницю в навчених представленнях між аутентичним і зміненим контентом. Ця помилка відтворення може служити індикатором наявності дипфейку [14].

Автокодувальник також може застосовуватися разом з іншими техніками, такими як змагальне навчання, для поліпшення виявлення діпфейків. Змагальне навчання передбачає навчання автокодувальник як на аутентичних, так і на діпфейкових даних, одночасно навчаючи мережу дискримінатора, яка розрізнятиме аутентичний і змінений контент. Цей спільний процес навчання може покращити здатність автокодувальника розрізняти реальний і фейковий контент.

2.2.7 Передавальне навчання

Передавальне навчання (англ. Transfer learning) - це техніка машинного навчання, яка полягає в використанні знань, отриманих під час навчання однієї моделі для конкретного завдання, для покращення результатів в іншому, але пов'язаному завданні. Замість навчання моделі з нуля, передавальне навчання дозволяє використовувати попередньо навчені моделі, які були натреновані на великих наборах даних для схожих завдань.

Завдяки використанню передавального навчання, можливо використовувати попередньо навчені моделі, які навчилися розпізнавати та розуміти людські обличчя, емоції та інші візуальні ознаки. Попередньо навчена модель може бути навчена далі на невеликому наборі даних, що містить як реальний, так і діпфейковий контент. Таким чином, модель може навчитися відрізняти між справжнім та фальсифікованим контентом, ідентифікуючи патерни, розбіжності або артефакти, характерні для діпфейків.

Потенційними кандидатами для перенесення навчання із метою виявлення діпфейків є:

- VGGNet - це згорткова нейронна мережа, яка була навчена на великому наборі даних зображень ImageNet. Вона може бути використана для вилучення ознак з облич, текстур та інших візуальних аспектів зображень, що може допомогти виявити артефакти та аномалії, характерні для діпфейків.

- ResNet (Residual Neural Network) - це глибока нейронна мережа, яка навчається на зображеннях для класифікації та виявлення об'єктів. Її глибока архітектура з використанням залишкових блоків дозволяє виявляти складні шаблони та особливості в зображеннях, що може бути корисним для виявлення діпфейків.

- FaceNet - це нейронна мережа, спеціалізована на розпізнаванні облич. Вона використовує глибокі згорткові шари для вилучення унікальних ознак облич та створення компактних представлень. FaceNet може бути використана для

порівняння та зіставлення облич на зображеннях, допомагаючи виявити ідентичність облич та виявити можливі діпфейки.

- LSTM (Long Short-Term Memory) - це рекурентна нейронна мережа, яка добре підходить для аналізу послідовностей даних, таких як звукові сигнали та мовлення. Вона може бути використана для виявлення аномалій у мовленні або артефактів, пов'язаних із синтезованими голосами в дипфейках, та допомогти ідентифікувати підозрілі аудіофрагменти.

2.2.8 Ансамблі моделей

Ансамбль моделей - це поєднання декількох моделей або алгоритмів для прогнозування або прийняття рішень. У контексті виявлення діпфейків ансамблеві моделі можуть бути використані для підвищення точності та надійності процесу виявлення.

Ідея використання ансамблю моделей полягає в тому, щоб навчати та поєднувати кілька моделей машинного навчання, кожна з яких має свої сильні та слабкі сторони, для створення більш надійної системи виявлення. Кожна модель у складі ансамблю може використовувати різні ознаки або техніки для виявлення ознак маніпуляції, такі як розбіжності у обличчі, неприродні рухи очей або аномалії в аудіовізуальних шаблонах.

Ансамблеві моделі можуть мати різні конструкції. Один з підходів полягає у навчанні різних моделей на різних підмножинах набору даних, що дозволяє кожній моделі спеціалізуватися на визначенні певних типів діпфейків. Інший підхід полягає у використанні різних алгоритмів, таких як згорткові нейронні мережі, рекурентні нейронні мережі або інші мережі, і комбінуванні їх прогнози за допомогою голосування або усереднення.

Під час фази виявлення ансамбль моделей аналізує задані вхідні дані і робить індивідуальні прогнози. Ці прогнози потім агрегуються, і остаточне рішення приймається на основі згоди або рівня впевненості ансамблю. За допомогою

спільного інтелекту декількох моделей метод ансамблю покращує точність виявлення діпфейків та зменшує ризик помилкових результатів.

У цьому розділі були розглянуті методи, з використанням яких можливо виявляти діпфейки. Були розглянуті традиційні методи виявлення обробки відео та методи виявлення діпфейків за допомогою технологій глибокого навчання, їх типи та можливості. Було встановлено, що традиційні методи виявлення підробок у відео можуть бути неефективними у випадках, якщо діпфейк було створено з використанням якісних алгоритмів та навичок. Крім того було встановлено, що найбільш ефективним методом виявлення діпфейків за допомогою глибокого навчання є метод ансамблів моделей, який дозволяє використовувати різні сильні сторони різних моделей глибокого навчання та збільшити варіативність діпфейків, що можуть бути виявлені.

3 АНАЛІЗ ПРЕВЕНТИВНИХ МЕТОДІВ БОРОТЬБИ З ДІПФЕЙКАМИ

У забезпеченні захисту від діпфейків можна приймати превентивні заходи, які дозволять запобігти появі та поширенню штучно створених відео та зображень. Превентивні підходи зосереджені на виявленні та запобіганні діпфейків заздалегідь, замість реактивного виявлення після їх поширення.

У цьому розділі ми розглянемо різноманітні превентивні методи боротьби з діпфейками, що базуються на технологіях автентифікації та захисту контенту від змін.

3.1 Використання технології блокчейн

Технологія блокчейн має потенціал бути дуже корисним інструментом у боротьбі з діпфейками, забезпечуючи безпечну і недоступну для втручань платформу для перевірки автентичності цифрового контенту. Одна з ключових особливостей блокчейну - це його децентралізована природа. Він не має центральної влади або сервера, який контролює усю мережу. Використовуючи децентралізований характер блокчейну, кілька вузлів можуть брати участь у процесі перевірки. Цей механізм згоди забезпечує, що перевірка не контролюється однією стороною і зменшує ризик маніпуляції або цензури. Прозорість блокчейну дозволяє будь-кому отримати доступ і провести аудит зареєстрованого контенту. Це робить блокчейн ефективним інструментом у боротьбі з дипфейками та зміненим цифровим контентом, сприяючи створенню надійних та безпечних цифрових середовищ.

Автентичний контент, такий як фотографії, відео або аудіозаписи, можна зареєструвати на блокчейні. Процес реєстрації передбачає створення унікального цифрового відбитка, або хешу, контенту і збереження його на блокчейні. Цей відбиток служить цифровим підписом, що унікально ідентифікує контент і не може бути змінений. При зустрічі з медіа користувачі можуть запитати блокчейн для перевірки його автентичності. Хеш медіа можна порівняти з зареєстрованими хешами на блокчейні. Якщо знайдено відповідність, це підтверджує, що медіа є автентичним і не піддавалося втручанню. Функцію встановлення часу блокчейну

можна використовувати для встановлення хронологічного порядку реєстрації контенту. Це допомагає визначити початкове джерело медіа і запобігає зміні міток часу, що додає додатковий рівень довіри [15].

Впровадження рішень на основі блокчейну для виявлення та перевірки діпфейків може значно посилити можливість виявляти змінений медіаконтент та захищати від його шкідливого впливу.

3.2 Криптографічний захист

Криптографічний захист від діпфейків полягає у використанні криптографічних технік для захисту від створення та маніпулювання контентом діпфейків.

Один з способів використання криптографічного захисту від діпфейків - це цифровий підпис. Цифровий підпис є математичною схемою, яка перевіряє автентичність та цілісність цифрових повідомлень або документів. Застосовуючи цифровий підпис до зображення або відео, стає можливим виявлення будь-яких змін або спроб втручання в контент. При створенні діпфейку та зміні оригінального контенту, цифровий підпис втрачає свою дійсність, що свідчить про можливість підробки.

Деякі можливі типи підписів:

- RSA (Rivest-Shamir-Adleman)

RSA є одним з найпоширеніших асиметричних криптографічних алгоритмів для цифрових підписів. Він базується на проблемі факторизації великих простих чисел і забезпечує високий рівень безпеки.

- ECDSA (Elliptic Curve Digital Signature Algorithm)

ECDSA є іншим асиметричним алгоритмом для цифрових підписів. Він базується на математичних властивостях еліптичних кривих і забезпечує ефективне виконання та міцну безпеку.

- EdDSA (Edwards-curve Digital Signature Algorithm)

EdDSA є модифікованою версією ECDSA, яка використовує іншу форму еліптичної кривої (крива Едвардса). Він пропонує покращену швидкість та безпеку порівняно з ECDSA.

Крім того, криптографічна хеш-функція може відігравати роль у захисті від діпфейків. Хеш-функція генерує унікальне хеш-значення для заданого вводу, такого як файл зображення або відео. Якщо хеш-значення відрізняється від очікуваного, це може свідчити про факт втручання у медіа.

Деякі можливі типи хеш-функцій:

- SHA-256 (Secure Hash Algorithm 256-bit)

SHA-256 є однією з найпоширеніших криптографічних хеш-функцій. Вона приймає вхідні дані будь-якого розміру і генерує хеш-значення довжиною 256 бітів.

- MD5 (Message Digest Algorithm 5)

MD5 є старішою хеш-функцією, яка генерує 128-бітове хеш-значення. Однак, через підвищену вразливість до колізійних атак, MD5 не рекомендується для захисту від сучасних загроз.

- SHA-3 (Secure Hash Algorithm 3)

SHA-3 є новішою версією криптографічної хеш-функції, розробленої на основі конкурсу NIST. Вона пропонує високу безпеку та стійкість до колізійних атак, а також може мати різні розміри вихідних хеш-значень (наприклад, SHA-3-256 або SHA-3-512).

Ці криптографічні техніки надають засоби для перевірки цілісності та автентичності цифрового контенту, сприяючи боротьбі з діпфейками. Застосовуючи цифрові підписи та хеш-функції, стає можливим виявити та запобігти поширенню зміненого або підробленого медіа.

3.3 Водяні знаки

Водяні знаки - це цифрові маркери або шаблони, які вбудовуються в саме відео і не легко помітні для людського ока. Вони служать формою автентифікації та можуть допомогти виявити початкове джерело відео.

Існують різні типи водяних знаків, які можуть бути застосовані до відео. Видимі водяні знаки, як правило, це напівпрозорі логотипи або текст, що накладаються на відео та вказують на авторські права. Невидимі водяні знаки, навпаки, непомітні для глядача і вбудовуються в дані відео, часто у вигляді незначних змін у значеннях пікселів або модифікацій спектру звуку відео.

У контексті боротьби з діпфейками, водяні знаки можуть відігравати важливу роль у збереженні цілісності відеоконтенту. Завдяки вбудованим у початкові відео унікальним водяним знакам стає можливим перевірити автентичність контенту та виявити будь-які несанкціоновані модифікації або спроби підробки. Водяні знаки можуть слугувати стримуванням для тих, хто може бути схильний створювати або поширювати діпфейки, оскільки наявність видимого або невидимого водяного знака ускладнює підробку.

Видимі водяні знаки, такі як логотипи або текст, можуть бути змінені або видалені, що дозволяє зловмисникам створити ілюзію оригінальності відео. Таке спотворення може вказувати на підробку та несанкціоновану модифікацію контенту.

Невидимі водяні знаки, які вбудовуються безпосередньо в дані відео можуть схильні до спотворення. Якщо в результаті обробки відео або інших маніпуляцій водяні знаки змінено або пошкоджено, це може бути пов'язано зі спробами видалити водяні знаки та може бути сигналом про підробку.

Спотворення водяних знаків може бути виявлено шляхом візуального аналізу відео або за допомогою спеціалізованих алгоритмів обробки зображень. Якщо спотворення помітне або є аномалії, це може бути підставою для подальшого розслідування та підозр у підробці.

Водяні знаки можуть мати різні властивості та характеристики, які роблять їх надійними засобами захисту від діпфейків, наприклад:

- Кожен водяний знак може бути унікальним та пов'язаним із конкретним автором, організацією або джерелом контенту. Це дозволяє легко визначити справжність відео та ідентифікувати автора.
- Деякі водяні знаки можуть бути візуально непомітними для звичайного спостерігача. Вони інтегруються у відео за допомогою різних методів, таких як зміна яскравості або колірних каналів, щоб бути практично непомітними.
- У разі виникнення спотворень або пошкоджень відео деякі методи водяного маркування дозволяють відновити водяні знаки та перевірити справжність контенту.
- Водяні знаки можуть бути вбудовані в різні компоненти відео, такі як зображення, звук або метадані. Це дозволяє створювати складні та багаторівневі водяні знаки, які важко виявити чи видалити.
- Різні алгоритми та методи обробки сигналів використовуються для створення та впровадження водяних знаків. Це може включати частотне перетворення, стеганографію або цифрові підписи, які роблять водяні знаки надійними та стійкими до атак.

3.4 Захист метаданих

Метадані - це інформація, вбудована в файл, яка надає деталі про його походження, створення, внесені зміни та інші атрибути. Застосування методів захисту метаданих ускладнює змінення їх зловмисниками з метою приховання маніпуляцій із відео.

Метадані можуть бути захищені наступними способами:

- Зберігання метаданих на блокчейні забезпечує їхню безпеку і незмінність. Технологія блокчейн гарантує, що метадані залишаються незмінними, оскільки будь-які зміни вимагають згоди учасників мережі.

- Застосування цифрових підписів до метаданих забезпечує їх автентичність та цілісність. Перевірка цифрового підпису метаданих допомагає встановити надійність відеоконтенту.
- Вбудовування невидимих або видимих водяних знаків в метадані може служити додатковим рівнем захисту. Водяні знаки можуть містити унікальні ідентифікатори або коди автентифікації, які допомагають перевірити автентичність метаданих. Зміна або видалення метаданих з водяними знаками призведе до помітних змін, що можуть вказувати на факт підробки відео.
- Шифрування метаданих допомагає захистити їх від несанкціонованого доступу та змін. Алгоритми шифрування забезпечують, що лише авторизовані сторони з доступом до ключів шифрування можуть отримувати доступ до метаданих та вносити в них зміни. Із зашифрованими метаданими, ризик маніпуляцій несанкціонованими особами значно зменшується.

3.5 Використання фізичних маркерів

Фізичні маркери - це конкретні об'єкти або шаблони, які розміщуються візуально на кадрі відео для автентифікації та перевірки. Вони можуть бути унікальними символами, логотипами, кольоровими плямами або візерунками, які важко точно відтворити або змінити.

Наявність фізичних маркерів у відео значно ускладнює завдання створення діпфейків і підробки контенту. Фізичні маркери ускладнюють створення діпфейків наступними шляхами:

- Фізичні маркери можуть бути унікальними символами або візерунками, що дуже важко точно відтворити або змінити. Це робить завдання створення вірогідних діпфейків з наявністю маркерів надзвичайно складним.
- Алгоритми виявлення діпфейків можуть аналізувати наявність та цілісність фізичних маркерів. Якщо діпфейк спробує замаскувати або видалити

маркери, це може призвести до помітних відхилень або аномалій, що розкриють факт маніпуляції з відео.

- Фізичні маркери можуть бути використані як посилання на інші елементи відео. Якщо діпфейк намагається змінити або перемістити маркери, це може призвести до помітних змін у візуальному контенті, які можуть бути виявлені алгоритмами аналізу відео.

Наявність фізичних маркерів робить підробку відео набагато складнішою та ризикованою для зловмисників, збільшуючи безпеку та довіру до автентичності відеоконтенту.

У цьому розділі були розглянуті превентивні методи боротьби з діпфейками. Було встановлено, що можливо прийняти завчасні міри, які можуть перешкоджати створенню реалістичних діпфейків або видаванню їх за автентичний контент.

4 ЗАСТОСУВАННЯ МЕТОДІВ ЗАПОБІГАННЯ ПОШИРЕННЮ ДІПФЕЙКІВ

Однією з основних небезпек діпфейків є їх вплив на громадську думку шляхом поширення через мережу Інтернет. За допомогою вжиття заходів, що перешкоджають їх поширенню можна значно зменшити цей ефект.

У цьому розділу будуть розглянуті можливі рішення для боротьби з поширенням дезінформації шляхом діпфейків.

4.1 Верифікація користувачів

Для боротьби з поширенням діпфейків, соціальні мережі та інші платформи, що можуть слугувати як шляхи поширення діпфейків, можуть впровадити системи верифікації користувачів. Це передбачає пов'язання облікового запису користувача з відеозаписами, щоб перевірити автентичність та достовірність вмісту. Коли користувач завантажує відео, платформа перевіряє його особу шляхом перевірки інформації облікового запису, такої як історія активності та загальна взаємодія на платформі. Якщо користувач вважається автентичним, поряд з відео може бути відображена знак верифікації або мітка, що підтверджує його достовірність.

Впровадження верифікації користувачів у боротьбі з діпфейками має кілька переваг. По-перше, це додає додатковий рівень довіри та відповідальності до вмісту, яким користувачі діляться на соціальних мережах та інших платформах. Користувачі більш ймовірно будуть довіряти та покладатися на відео, пов'язані з верифікованими обліковими записами, зменшуючи потенційне поширення шкідливих діпфейків. Крім того, це діє як стримування для тих, хто може намагатися створювати і розповсюджувати діпфейки, оскільки вони можуть втратити статус довіреного користувача.

Хоча верифікація користувачів може бути ефективним інструментом у боротьбі з діпфейками, важливо визнати, що вона не є надійним рішенням і може мати свої

недоліки. Один з можливих ризиків полягає в тому, що верифікований обліковий запис може бути скомпрометований або зламаний. Якщо верифікований обліковий запис скомпрометований, довіра, пов'язана з міткою верифікації, може бути використана для обману інших користувачів. Діпфейки, розповсюджені зловмисниками за допомогою скомпрометованого верифікованого облікового запису, можуть бути ще переконливішими та шкідливими, оскільки більш ймовірно, що їм будуть довіряти аудиторії

Системи моніторингу верифікованих облікових записів можуть допомогти виявити будь-яку незвичайну або підозрілу діяльність, що дозволить вчасно здійснити втручання та зменшити можливі ризики. Платформи також повинні мати чіткі політики та процедури для реагування на скомпрометовані облікові записи та прийняття відповідних заходів, таких як тимчасове призупинення або анулювання статусу верифікації, поки не буде відновлена безпека облікового запису.

4.2 Флагування підозрілого контенту

Флагування передбачає позначення або маркування відео, які підозрюються або підтверджуються як діпфейки. Ці прапорці слугують попередженням для глядачів і платформ, що відео може бути недостовірним або ненадійним.

У контексті боротьби з діпфейками флагування підробленого контенту може бути впроваджено кількома способами. Користувачі можуть бути закликані повідомляти підозрілі відео, як через функцію скарг на платформі. Крім того, можуть використовуватись автоматизовані системи на основі алгоритмів штучного інтелекту для аналізу відео та позначення тих, що проявляють характеристики, характерні для діпфейків.

Флагування підробленого контенту має кілька переваг. По-перше, воно допомагає підвищити усвідомлення серед глядачів про наявність потенційних діпфейків, що дозволяє їм критично ставитися до контенту і бути обережними. Крім того, воно допомагає соціальним мережам швидше виявляти і перевіряти потенційно шкідливі відео, щоб вжити відповідних заходів, таких як видалення контенту або

подальше розслідування. Крім того, флагування може сприяти розвитку баз даних та навчальних наборів для алгоритмів виявлення діпфейків, покращуючи їх точність та ефективність з часом.

Однак важливо враховувати складнощі, пов'язані з флагуванням підробленого контенту. Визначення автентичності відео лише на підставі повідомлень користувачів або автоматизованих систем може бути складним та схильним до помилок. Можливі помилки, що призводять до неправильної ідентифікації справжнього контенту або невдалого позначення складних діпфейків. Тому важливим є поєднання повідомлень користувачів, технологій штучного інтелекту та процесів перегляду людиною для ефективної системи флагування.

4.3 Посилення захисту акаунтів

Посилення захисту облікових записів у соціальних мережах та інших платформах може допомогти при боротьбі із поширенням діпфейків. Хоча зловмисники можуть створювати безліч облікових записів з метою розповсюдження дезінформації, користувачі платформ з більшою ймовірністю повірять у справжність контенту якщо його опублікував не новий запис, створений учора, а запис, який виглядає як той, якому можна довіряти. Це можуть бути як і облікові записи звичайних людей, які зламали з метою надання контенту автентичного виду, так і облікові записи брендів, знаменитостей, інфлюенсерів, політиків та державних установ. Потенційні наслідки зламу облікового запису впливового або довіреного користувача можуть бути дуже серйозними. Тому важливо мати адекватні засоби захисту акаунтів у соціальних мережах та інших онлайн-платформах, щоб знизити ризики, пов'язані з діпфейками.

Можливими методами захисту можуть бути:

- Двофакторна автентифікація

Двофакторна автентифікація - це метод підвищення безпеки акаунту, який вимагає введення двох незалежних способів підтвердження особи користувача перед отриманням доступу до акаунту. Найпоширенішим методом двофакторної

автентифікації є одноразові паролі, які надсилаються на мобільний пристрій користувача. Також можуть бути використана перевірка біометричних даних, таких як відбитки пальців або розпізнавання обличчя.

- Вимога до сильних паролів

Платформа може вимагати від користувачів створювати складні паролі із комбінацій букв, цифр та спеціальних символів. При цьому платформа може запропонувати ресурси для оцінки складності паролів та перевірки їх стійкості.

- Обмеження спроб входу

Обмеження кількості невдалих спроб входу до облікового запису. Після певної кількості невдалих спроб входу, обліковий запис може бути заблокований або вимагати додаткову автентифікацію. Цей метод ефективний при захисті від перебору паролів.

- Обмеження доступу з непідтримуваних пристроїв

Заборона доступу до облікового запису з непідтримуваних пристроїв або з неактуальних версій програм, які можуть бути вразливими до атак через свою застарілість або відсутність необхідних оновлень безпеки.

- Моніторинг активності акаунтів

Моніторинг активності акаунтів є важливим методом забезпечення їх безпеки. Цей процес включає постійне спостереження за діяльністю користувача та виявлення будь-яких підозрілих або нестандартних змін.

Платформа може зберігати історію входу користувача, включаючи дату, час і місце входу. Це дозволяє користувачу перевірити активність на своєму обліковому записі та виявити будь-які незвичні або невідомі входи. Крім того може отримувати сповіщення про будь-яку підозрілу активність на своєму акаунті.

Платформа може реагувати на незвичайну активність на акаунті. Це можуть бути спроби входу з незвичайних місць, незвичайний обсяг активності або незвичайні

запити на доступ до даних. Якщо виявляються підозрілі або небезпечні активності, платформа може приймати заходи для блокування таких дій. Наприклад, акаунт може бути тимчасово заблокований або вимагати додаткової автентифікації для підтвердження користувача.

- Шифрування даних

Комунікація між користувачем і сервером може бути зашифрована за допомогою протоколів шифрування з метою забезпечення конфіденційності даних під час їх передачі через мережу Інтернет.

Дані, які зберігаються на серверах платформи, можуть бути зашифровані, щоб забезпечити їх конфіденційність. Шифрування може застосовуватися до усіх баз даних, або конкретно до чутливої інформації. Це дозволяє зберігати дані в зашифрованому вигляді, навіть якщо зловмисник отримає прямий доступ до внутрішньої інформації платформи.

- Регулярні аудити безпеки

Аудит безпеки є процесом систематичного перегляду та оцінки заходів безпеки для визначення потенційних слабких місць і виявлення потенційних загроз безпеці. Основна мета аудиту безпеки полягає в забезпеченні відповідності з принципами безпеки, виявленні потенційних ризиків і наданні рекомендацій щодо поліпшення системи безпеки. Аудит безпеки оцінює велику кількість систем, наприклад: ефективність методів ідентифікації та автентифікації, захисту даних та мережі, фізичної безпеки, безпеку додатків тощо.

- Навчання користувачів

Навчання користувачів про захист акаунтів є важливою складовою частиною забезпечення безпеки. Це допомагає користувачам розуміти потенційні загрози та навчає їх кращим практикам забезпечення безпеки акаунтів. Можливі аспекти, з якими можна ознайомлювати користувачів:

- Навчання користувачів про створення міцних паролів.

- Пояснення принципу роботи та важливості двофакторної автентифікації. Навчання користувачів про налаштування двофакторної автентифікації для своїх акаунтів.
- Навчання користувачів розпізнавати спроби фішингу, уникати небезпечних посилань, не надавати конфіденційної інформації на підозрілих веб-сайтах.
- Пояснення важливості регулярного оновлення програмного забезпечення, включаючи операційну систему, браузери та інші програми.
- Вказівки щодо безпечного обміну інформацією.
- Навчання користувачів про важливість перевірки активності свого акаунту, та сповіщень про неї.
- Пояснення необхідності встановлення та оновлення антивірусного програмного забезпечення на пристроях користувачів для запобігання вірусам та іншим шкідливим програмам.

4.4 Програми цифрової грамотності

Цифрова грамотність означає здатність ефективно взаємодіяти з цифровими технологіями, зокрема розуміти та оцінювати автентичність та надійність цифрового контенту. У контексті боротьби з дІпфейками програми цифрової грамотності спрямовані на забезпечення особам необхідних навичок і знань для виявлення та зменшення ризиків, пов'язаних із підробленими відео.

Ці програми навчають людей різним аспектам, що пов'язані з дІпфейками, таких як технології, що використовуються для їх створення, типові візуальні артефакти та аномалії, що зустрічаються в змінених відео, а також методи перевірки автентичності відеоконтенту. Учасники вчаться аналізувати метадані відео, оцінювати достовірність джерел і використовувати інструменти перевірки фактів для перевірки інформації.

Завдяки розвитку навичок цифрової грамотності, особи можуть стати більш обізнаними споживачами відеоконтенту. Вони можуть розпізнавати ознаки маніпуляції, такі як непослідовність виразів обличчя, ненатуральні рухи або розбіжності між аудіо та відео. Програми цифрової грамотності також сприяють

критичному мисленню, заохочуючи осіб перевіряти інформацію, перш ніж приймати її за достовірну. Підвищуючи розуміння технік, використовуваних для створення діпфейків, особи можуть краще захистити себе від дезінформації. Крім того, ті, хто володіє навичками цифрової грамотності, більш ймовірно виявлять діпфейки і повідомлять про них відповідним органам або платформам, що допомагає стримати їх поширення.

У цьому були розглянуті різноманітні методи, спрямовані на запобігання поширенню діпфейків. Було встановлено, що важливо надавати користувачам навички та інструменти для прийняття рішень щодо автентичності контенту в Інтернеті та встановлювати і захищати механізми, що породжують довіру в перевірені джерела.

5 КОМПАНІЇ ТА ПРОГРАМИ, ЩО НАДАЮТЬ ПОСЛУГИ З БОРОТЬБИ З ДІПФЕЙКАМИ

У цьому розділі приведені програми та компанії, що пропонують послуги захисту від діпфейків. Ці організації працюють над розробкою та вдосконаленням технологій, методів та інструментів, які допомагають виявляти, аналізувати та запобігати поширенню діпфейків. Програми та компанії для захисту від діпфейків пропонують широкий спектр послуг, що включають автоматичне виявлення діпфейків, аналіз контенту на основі машинного навчання, автентифікацію цифрового контенту тощо.

5.1 Amber Video

Amber Video – компанія-розробник технологій боротьби з діпфейками. Вона спеціалізується на розробці інноваційних рішень, які допомагають виявляти, відстежувати та протидіяти діпфейкам.

Amber Video розробляє та використовує алгоритми машинного навчання для виявлення діпфейків у відео- та аудіоматеріалах. Компанія пропонує розширений аналіз відео та аудіо шляхом виявлення невідповідностей, аномалій та потенційних ознак діпфейків. Вони використовують сучасні алгоритми обробки даних та розпізнавання патернів, щоб виявити будь-які ознаки фальсифікації. Також компанія працює над створенням систем автоматичного виявлення та блокування діпфейків у режимі реального часу.

Крім того Amber Video пропонує консультації та навчальні програми з боротьби з діпфейками. Вони навчають користувачів розпізнавати та захищатися від діпфейків, а також надають рекомендації щодо кращих практик у цій сфері.

5.2 Truepic

Компанія Truepic пропонує різноманітні інструменти та технології, які допомагають перевіряти автентичність зображень і відео. Одним з ключових рішень є використання блокчейн-технології, що гарантує незмінність та відстежуваність зображень. Кожне зображення або відео, зроблене за допомогою

Truepic, отримує цифровий підпис, який може бути перевірений і підтверджений третьою стороною. Це робить неможливим будь-які зміни або редагування без виявлення.

Також Truepic використовує штучний інтелект для аналізу зображень та відео. Цей інструмент здатний виявляти незначні артефакти, які можуть свідчити про підроблення або інші типи маніпуляцій із зображеннями. Таким чином, вони можуть виявити навіть найдрібніші зміни, які були внесені в оригінальну копію.

5.3 Sensity

Компанія Sensity пропонує широкий спектр послуг та рішень, які допомагають виявляти, аналізувати та запобігати поширенню штучно створених відео, зображень та аудіоматеріалів.

Sensity використовує сучасні алгоритми та штучний інтелект для автоматичного виявлення діпфейків. Вони аналізують відео, зображення та аудіоматеріали, виявляють ознаки маніпуляцій та штучної обробки, що допомагає швидко і ефективно виявляти потенційні діпфейки. Крім того компанія має команду експертів, які детально аналізують підозрілі відео та зображення. Вони проводять дослідження, перевіряють джерела, проводять перехресну перевірку та експертну оцінку, щоб встановити правдивість чи фальшивість контенту.

Sensity розробляє та впроваджує передові технології для боротьби з діпфейками. Це включає в себе розробку алгоритмів, інструментів для виявлення фотомонтажу, впровадження блокчейн-технологій для підтвердження автентичності контенту тощо. Також Sensity активно працює над підвищенням обізнаності про діпфейки та небезпеки, які вони несуть. Компанія проводить тренінги, семінари та вебінари для громадськості, допомагаючи людям розпізнавати діпфейки та захищати себе від їх негативних наслідків.

5.4 Adobe Content Authenticity Initiative

Adobe Content Authenticity Initiative є програмою, розробленою компанією Adobe з метою боротьби з проблемою дезінформації, особливо в контексті поширення діпфейків. Основна мета ініціативи полягає в наданні засобів для підтвердження автентичності контенту, що розповсюджується в Інтернеті. Це досягається за допомогою використання цифрових підписів і блокчейн-технологій. Увесь контент, створений за допомогою підтримуваних програм Adobe, може бути позначений цифровим підписом, який дозволяє перевірити його автентичність та походження [16].

Переваги використання послуг Adobe Content Authenticity Initiative в боротьбі з діпфейками наступні:

- Завдяки цифровим підписам можна перевірити, автентичність контенту.
- Блокчейн-технологія дозволяє зберігати інформацію про походження контенту, що дає змогу відстежувати, де і коли він був створений.
- Ініціатива допомагає захищати права творців контенту, позначаючи його цифровим підписом і вказуючи на авторство.

Впровадження цифрових підписів у контент сприяє збереженню довіри споживачів до контенту та зменшує ймовірність поширення шкідливої інформації.

5.5 Microsoft Video Authenticator

Microsoft Video Authenticator - це інструмент, розроблений компанією Microsoft, який допомагає виявляти та боротися з діпфейками. Microsoft Video Authenticator використовує штучний інтелект та алгоритми комп'ютерного зору для аналізу відео та виявлення ознак діпфейків. Технологія базується на глибокому навчанні, що дозволяє автоматично розпізнавати підозрілі зміни в контенті [17].

Можливості Microsoft Video Authenticator:

- Microsoft Video Authenticator може швидко сканувати відео та виявляти невідповідності або аномалії, що свідчать про можливість діпфейку.

- Сервіс може класифікувати відео на основі ймовірності наявності дипфейку. Він надає оцінку достовірності відео, допомагаючи відрізнити оригінальний контент від підробленого.
- Коли Video Authenticator виявляє діпфейк, він може надіслати сповіщення адміністраторам або користувачам, щоб вони прийняли відповідні заходи для запобігання поширенню недостовірної інформації.

5.6 Serelay

Serelay - компанія, яка спеціалізується на боротьбі з дипфейками та наданні безпеки цифровому контенту. Компанія розробляє інноваційні рішення, які допомагають виявляти та відслідковувати діпфейки. Однією з ключових послуг компанії є технологія перевірки достовірності зображень та відео. Вона дозволяє автоматично аналізувати вміст і виявляти ознаки фальсифікації. Для виявлення діпфейків Serelay використовує машинне навчання, аналіз метаданих та глибини.

Крім того, Serelay надає можливість встановлення цифрових підписів, які гарантують цілісність та автентичність вмісту. Це дозволяє встановити, що зображення або відео не зазнали змін після створення та не піддалися втручанню. Такі підписи можуть бути використані як докази в судових процесах, а також довідками при перевірці достовірності медіа.

Одним з важливих аспектів послуг Serelay є їхня здатність працювати зі збереженням конфіденційності. Інформація про клієнтів та їхній вміст обробляється з дотриманням строгих стандартів безпеки та конфіденційності даних.

У цьому розділі були приведені різноманітні програми та компанії, які займаються захистом від діпфейків та забезпеченням інформаційної безпеки. Отже існують багато готових рішень автентифікації контенту. Поява та розвиток таких компаній допомагають у створенні безпечного інформаційного середовища.

6 ПРОПОЗИЦІЇ ЩОДО ЗАХИСТУ ВІД ПІДРОБКИ ВІДЕОЗОБРАЖЕНЬ ЗА ДОПОМОГОЮ ТЕХНОЛОГІЇ ДІПФЕЙК

У цьому розділі роздивимося можливі заходи щодо зменшення ризиків від діпфейків. Захист від діпфейків – задача, що потребує комплексного підходу, тому розіб’ємо пропозиції на три частини, відповідно до видів захисту від діпфейків, що були розглянуті раніше.

6.1 Виявлення

Інструменти, за допомогою яких можливо із впевненістю сказати чи є відео діпфейком, корисні у багатьох контекстах, будь то перевірка побаченого відео користувачами соціальних мереж, перевірка відео про інцидент новинними організаціями, встановлення автентичності відеодоказів при судовому розслідуванні або перед прийняттям стратегічних рішень державами.

Створення, поширення та інтегрування доступних та зрозумілих інструментів виявлення діпфейків може допомогти громадськості формувати об’єктивне уявлення про події та запобігти поширенню шахрайства, пропаганди та інформації, що направлена на дискредитацію осіб або державних установ.

Також можуть бути створені норми та протоколи, за якими повинна перевірятися інформація, що може бути підроблена перед її публікацією у ЗМІ. Це може допомогти зменшити кількість дезінформації, що потрапляє у ЗМІ та підвищити рівень довіри громадськості до новин. Подібні протоколи також необхідні при аналізі відеодоказів, що може запобігти помилковим вирокам у суді та невірній оцінці ситуацій при прийнятті рішень на державному рівні.

Ці протоколи перевірки можуть включати використання традиційних методів виявлення підробки відео з метою відсіювання діпфейків низької якості перед використанням більш технологічних методів з метою зменшення використання обчислювальних можливостей. Після відсіювання діпфейків, що легко

виявляються, відео може бути аналізовано за допомогою технологій глибокого навчання. Для виявлення діпфейків можуть використовуватися різні типи моделей глибокого навчання, які мають свої сильні та слабкі сторони, тому самим ефективним варіантом є використання методу ансамблів моделей, що дозволяє використовувати різні сильні сторони різних моделей та розширити варіативність діпфейків, які можуть бути виявлені.

6.2 Превентивні заходи

За допомогою використання методів автентифікації контенту можна ускладнити створення правдоподібних діпфейків та спростити їх виявлення. Впровадження стандартів захисту метаданих файлів ускладнить їх підробку, що підвищить довіру до них та створить перешкоду для тих, хто бажає видати підроблене відео за справжнє.

Особи та організації, що бажають захиститися від діпфейків, можуть використовувати цифрові підписи та блокчейн для забезпечення автентичності свого контенту та вставляти у свій контент водяні знаки та фізичні маркери, що ускладнить його використання для створення діпфейків.

Також можуть бути створені норми та протоколи, за якими відео та фото, що створюють державні установи та впливові компанії, мають бути створені із урахуванням заходів по боротьбі із діпфейками.

6.3 Запобігання поширенню діпфейків

Шкода від діпфейків наноситься лише у тому разі, коли їх бачить та їм вірить їх цільова аудиторія. Якщо встановити механізми обмеження поширення діпфейків та попередження користувачів про недостовірність інформації, шкоду від діпфейків можна мінімізувати.

Позначення надійних джерел на онлайн платформах може допомогти підвищити до них довіру та зменшити ризик появи у громадськості віри у те, що інформація, яка їм не подобається має бути підробкою. Також позначення або видалення підозрілого контенту може допомогти зупинити розповсюдження діпфейків.

Велику небезпеку становить розповсюдження зловмисниками діпфейків від імені достовірної або впливової особи або організації, якими можуть бути поважні члени онлайн спільнот, інфлюенсери, бренди, ЗМІ, політики тощо. Тому важливо, щоб онлайн-платформи мали адекватний захист облікових записів користувачів. Можуть бути створені норми захисту облікових записів, які мають реалізувати онлайн платформи, що мають реальний вплив на думку громадськості.

Організації та особи, що мають вплив на громадську думку, або можуть слугувати місцем поширення діпфейків можуть бути зобов'язані модерувати свій контент. Наприклад, соціальні мережі можуть бути примушені встановити засоби виявлення та позначення діпфейків. Офіційні ЗМІ та власники впливових онлайн каналів та облікових записів можуть бути покарані за випадки поширення ними дезінформації.

Також зменшити вплив діпфейків можуть програми цифрової грамотності. Поширення навичок та технік виявлення діпфейків та просування ідей критичного мислення при взаємодії з контентом в Інтернеті можуть значно посилити стійкість громадськості до введення в оману діпфейками та іншими типами дезінформації.

У цьому розділі були приведені рекомендації щодо заходів захисту від діпфейків, які можуть бути прийняті на рівнях особи, компанії та держави. Виконання цих рекомендацій може сприяти ефективному запобіганню поширенню дезінформації і підвищенню довіри до відео контенту.

ВИСНОВКИ

У роботі була досліджена технологія діпфейк та методи захисту від підробки відеозображень за допомогою неї. Було розглянуто принцип роботи технології, типи діпфейків, методи виявлення діпфейків та методи запобігання їх поширенню.

У результаті дослідження було встановлено, що підробка відео за допомогою технології діпфейк залишає сліди, у виявленні яких людське око та традиційні методи виявлення підробок можуть бути неефективними, але при виявлення діпфейків ефективна та ж сама технологія, на якій заснований діпфейк – глибоке навчання. Були розглянуті різні методи глибокого навчання та їх можливості, які можуть бути корисними при виявленні діпфейків.

Також були розглянуті превентивні методи захисту від діпфейків, засновані на автентифікації відеоконтенту, захисті метаданих та вставленні у відео маркерів, які ускладнюють створення діпфейків.

Крім того були розглянуті методи запобігання поширенню діпфейкам, спрямовані на надання користувачам онлайн платформ знань, навичок та інструментів для прийняття рішень щодо автентичності контенту. Також були розглянуті готові рішення у сфері автентифікації контенту та виявлення діпфейків.

Мета роботи була досягнена шляхом створення рекомендацій щодо забезпечення захисту від поширення дезінформації шляхом використання діпфейків. Результати дослідження вказують на те, що захист від діпфейків потребує комплексного рішення, що має включати розробку інструментів виявлення діпфейків, захисту медіа від підробки, автентифікації справжнього медіа, поширення знань про діпфейки та інстументів і навичок їх виявлення.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Bernaciak C., Ross D. How Easy Is It to Make and Detect a Deepfake? URL: https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/ (дата звернення: 29.04.2023).

2 Westerlund M. The Emergence of Deepfake Technology: A Review. URL: https://www.researchgate.net/publication/337644519_The_Emergence_of_Deepfake_Technology_A_Review (дата звернення: 28.04.2023).

3 A Deep Dive into Deepfake Types. URL: https://q5id.com/blog/a-deep-dive-into-deepfake-types (дата звернення: 28.04.2023).

4 Zhang W., Zhao C., Li Y. A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis. URL: https://www.researchgate.net/publication/339425119_A_Novel_Counterfeit_Feature_Extraction_Technique_for_Exposing_Face-Swap_Images_Based_on_Deep_Learning_and_Error_Level_Analysis (дата звернення: 28.04.2023).

5 Deep Learning for Deepfakes Creation and Detection: A Survey / Nguyena T.T. та інші. URL: https://arxiv.org/pdf/1909.11573.pdf (дата звернення: 28.04.2023).

6 Almars A.M. Deepfakes Detection Techniques Using Deep Learning: A Survey. URL: https://www.scirp.org/journal/paperinformation.aspx?paperid=109149 (дата звернення: 29.04.2023).

7 Weerawardana M., Fernando T. Deepfakes Detection Methods: A Literature Survey. URL: https://ieeexplore.ieee.org/document/9606067 (дата звернення: 02.05.2023).

8 DeepFake Detection using Convolutional Neural Networks. URL: https://techvidvan.com/tutorials/deepfake-detection-using-cnn/ (дата звернення: 02.05.2023).

9 Güera D., Delp E. J. Deepfake Video Detection Using Recurrent Neural Networks. URL: https://gangw.web.illinois.edu/class/cs598/papers/AVSS18-deepfake.pdf

10 Preeti, Kumar M., Sharma H.K. A GAN-Based Model of Deepfake Detection in Social Media. URL: https://www.sciencedirect.com/science/article/pii/S1877050923001916 (дата звернення: 02.05.2023).

11 Xu L. Face Manipulation with Generative Adversarial Network. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1848/1/012081 (дата звернення: 03.05.2023).

12 Malingan N. Attention Mechanism in Deep Learning. URL: https://www.scaler.com/topics/deep-learning/attention-mechanism-deep-learning/ (дата звернення: 03.05.2023).

13 Nguyen H.H., Yamagishi J., Echizen I. Use of a Capsule Network to Detect Fake Images and Videos. URL: https://arxiv.org/pdf/1910.12467.pdf (дата звернення: 03.05.2023).

14 Deep fake detection using cascaded deep sparse auto-encoder for effective feature selection / Balasubramanian S. B. та інші. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9299276/ (дата звернення: 03.05.2023).

15 Patil U., Chouragade P. M. Deepfake Video Authentication Based on Blockchain. URL: https://ieeexplore.ieee.org/document/9532725 (дата звернення: 03.05.2023).

16 Allen W. The Content Authenticity Initiative unveils content attribution tool within Photoshop and Behance. URL: https://blog.adobe.com/en/publish/2020/10/20/content-authenticity-initiative-unveils-content-attribution-tool-within-photoshop-behance (дата звернення: 05.05.2023).

17 Microsoft launches video authenticator to detect deepfake media. URL: https://tech.hindustantimes.com/tech/news/microsoft-launches-video-authenticator-to-detect-deepfake-media-71599114673370.html (дата звернення: 05.05.2023).