https://doi.org/10.31649/1997-9266-2025-180-3-126-132

УДК 004.8

М. Б. Марчук[1]
В. В. Лукічов[1]

АНАЛІЗ МЕТОДІВ ТА ЗАСОБІВ АКТИВНОГО ЗАХИСТУ ВІД DEEPFAKE

[1]Вінницький національний технічний університет

Розвиток штучного інтелекту є одним з головних трендів останніх років. Хоч така технологія багато в чому служить на благо людства, але вона може використовуватись і для зловмисних дій, як-от поширення дезінформації чи шантаж.

Для реалізації цих злочинних дій часто застосовуються технології для створення так званого Deepfake-контенту. Deepfake — це загальна назва для зображень, відео- та аудіофайлів, створених за допомогою штучних нейронних мереж, які зображають, як певна людина говорить чи вчиняє щось, чого в дійсності ніколи не було. Завдяки використанню штучного інтелекту такі матеріали здаються справжніми особам, що не знають про їхнє походження. Без використання спеціальних технічних засобів людям важко відрізнити справжнє зображення від підробленого.

Технології генерації контенту за допомогою штучного інтелекту, зокрема Deepfake, все частіше застосовуються для створення матеріалів, які використовуються для дезінформаційних кампаній, шантажу та інших зловмисних дій. В зв'язку з цим існує необхідність в розробці засобів для захисту від зловмисного використання Deepfake. Пасивні методи захисту, які ґрунтуються на можливості моделей машинного навчання відрізнити автентичний контент від згенерованого, залежать від архітектур моделей для генерації Deepfake-контенту, тому швидко застарівають з темпом розвитку останніх. Тому існує необхідність в розвитку активних методів захисту, які ґрунтуються на використанні водяних знаків з метою їх відслідковування або перешкоджають роботі моделей для генерації Deepfake.

У роботі подано опис та аналіз наявних активних методів захисту від Deepfake, а також проаналізовано напрямки їх цільових застосувань, їхні переваги та недоліки, технічні характеристики, і запропоновано напрямки для подальших досліджень у цій галузі. Активна увага приділяється засобам захисту, що ґрунтуються на методах стеганографії та із застосуванням водяних знаків. Розглянуто переваги активних методів захисту над пасивними.

Ключові слова: Deepfake, стеганографія, водяні знаки, штучний інтелект, машинне навчання, інформаційна атака, захист інформації.

Вступ

Розвиток штучного інтелекту є одним з головних незмінних трендів останніх років. Розвиток цих технологій є настільки стрімким і неконтрольованим, що навіть спробували закликати компанії-лідери, які працюють в цьому напрямку, тимчасово зупинити подальші розробки [1]. Але, незважаючи на це, розробки рішень штучного інтелекту лише прискорились [2].

Іншим трендом, що розвивався в останні роки, є поширення дезінформації в мережі Інтернет. Цьому сприяв розвиток соціальних мереж, які дозволяють поширювати інформацію швидко та без валідації неупереджених осіб. У контексті використання штучного інтелекту в створені дезінформації зазвичай згадується термін "Deepfake". Під цим терміном розуміється низка методів для створення та/або маніпуляції аудіовізуальним контентом, в результаті чого утворюються відео з зображенням осіб, на яких вони виконують або говорять щось, чого ніколи не траплялось в дійсності. Такі маніпуляції можливі завдяки використанню засобів штучного інтелекту. Незважаючи на те, що відео, створені цим методом, ніколи не знімались на камеру, для пересічного спостерігача складається враження, що вони автентичні.

Методи генерації Deepfake-контенту можуть використовуватись для різноманітних випадків незаконної діяльності як-от шахрайство, поширення дезінформації та імітації особистості без її

© М. Б. Марчук, В. В. Лукічов, 2025

згоди. Через це існує потреба в методах, що можуть допомогти в виявленні Deepfake-контенту.

Більшість методів розпізнавання Deepfake, що існують на момент написання, є бінарними методами класифікації для розподілення справжніх та підроблених зображень. На момент написання розвиток інструментів для створення Deepfake не досяг рівня, за якого неможливо розрізнити згенеровані зображення від справжніх. В середньому люди можуть розпізнати Deepfake в 60…70 % випадків [5]. Але люди зазвичай переоцінюють свої шанси в розпізнавані згенерованих зображень [6]. Хоч шанси людей розпізнавати Deepfake не є ідеальними, це не стає перепоною в аспектах життя, пов'язаних зі споживанням медіаконтенту, як цифрова гігієна, де інтуїтивне розуміння того, що саме може бути Deepfake-контентом, є достатнім.

Але наразі не існує достатньо інструментів у цій області для таких випадків як, наприклад, кримінальні розслідування [7] або публікація зображень медіаресурсами, де бінарна класифікація реальних та згенерованих медіа є недостатньою без історії походження медіафайлу.

Іншою проблемою наявних методів розпізнавання згенерованого контенту є те, що інструменти для створення Deepfake постійно розвиваються. Оскільки нові інструменти генерації не мають недоліків попередніх поколінь, інструменти розпізнавання з часом втрачають свою актуальність. Ця проблема в основному стосується пасивних методів пошуку, які розпізнають Deepfake-контент за неточностями, що залишаються після роботи моделей генерації. Але активні методи розпізнавання є стійкішими в цьому плані, оскільки вони залежать не від недоліків моделей генерації контенту, а від завчасно встановлених механізмів захисту.

Одним з видів активного захисту від Deepfake є використання методів стеганографії та водяних знаків. В більшості методах захист відбувається за рахунок маркування зображення стеганографічним прихованим повідомленням або водяним знаком. У випадку, якщо таке зображення використано для створення контенту без погодження власника оригінального зображення, це можна довести через наявність маркування.

Метою статті є аналіз та порівняння наявних методів активного захисту від зловмисного використання Deepfake, що використовують методи стеганографії та водяних знаків; також розглянути, як ці методи дозволяють відслідковувати походження цифрових файлів, які їхні переваги та недоліки та які подальші роботи можуть бути виконанні в цьому напрямку.

Постановка проблеми

Deepfake — це загальна назва для способів маніпуляції аудіовізуальним контентом, щоб зобразити дії певних осіб, які ті ніколи не вчиняли. Саме слово Deepfake утворене комбінацією словосполучення "Deep Learning", що позначає один з популярних методів використання штучних нейронних мереж, та "Fake", що в перекладі означає «підробка» [8]. Приклад створення Deepfake за допомогою двох зображень показано на рис. 1.

Рис. 1. Приклад створення Deepfake за допомогою двох зображень.

Незважаючи на негативну конотацію у назві, технології для Deepfake мають і позитивні кейси використання, загалом для креативних індустрій. Наприклад, при створенні фільмів вони можуть використовуватись для омолодження акторів в кадрі або ж змінення сцен без їх перезнімання. Також Deepfake можуть використовуються для розважальних цілей, наприклад, створення пародійних відео на YouTube.

Використання цих методів для зловмисних дій є поширеним явищем. Наприклад, Deepfake-контент може використовуватись для поширення дезінформації шляхом створення підроблених фото та відео, що нібито зображують певні події, які ніколи не траплялись, чи що нібито певні публічні особи казали слова, яких ті ніколи не промовляли.

Розповсюджене використання Deepfake підриває довіру споживачів [12].

Згідно зі звітом компанії Sumsub в 2024 році кількість зафіксованих випадків використання

Deepfake в світі зросла на 245 %. Україна є однією з країн, де використання Deepfake зустрічалось найчастіше [10].

Іншим аспектом використання Deepfake, що може бути спрямований проти моралі, є дискредитація певної особистості або ж взяття контролю за її публічним представленням. Через Deepfake-медіа можна дискредитувати певну особу, створивши відео антисоціального характеру нібито як за участю певної особи, але без її згоди та без її реальної участі у створенні відео. Це можуть бути відео порнографічного характеру, зображення злочинних дій, або інший контент, що суперечить загальноприйнятій моралі.

Для того, щоб боротися зі зловживаннями використання Deepfake, необхідно застосовувати спеціальні інструменти, які можуть допомогти їх розпізнати. Усі такі засоби можна поділити на дві категорії: пасивні і активні [9].

Пасивні методи розпізнавання працюють за рахунок того, що методи генерації Deepfake створюють неідеальні зображення і залишають по собі артефакти, відслідкувавши які можна визначити з певною ймовірністю, чи було зображення створене генеративною мережею. Головним недоліком таких методів розпізнавання є те, що методи генерації постійно оновлюються і поліпшуються, через що створювані ними зображення стають щораз більше наближеними до реальності і з меншою кількістю артефактів. Через це наявні пасивні методи розпізнавання з часом стають неактуальними.

В зв'язку з цим існує потреба в розробці активних методів розпізнавання Deepfake.

Метрики непомітності прихованих повідомлень

Непомітність визначає сприймання схожості оригінального зображення та зображення, захищеного прихованим повідомленням. Стеганографічно приховане повідомлення або водяний знак мають бути важкопомітними, щоб кінцевий користувач фізично не міг сприймати будь які візуальні відмінності. Хоч приховане повідомлення не повинно змінювати якісь зображення, незначні об'єми деградації допустимі [11].

Основними метриками непомітності прихованих повідомлень є PSNR та SSIM.

Пікове співвідношення сигналу до шуму або *PSNR* (від англ. *Peak Signal-to-Noise Ratio*) — є мірою визначення різниці у відхилені за значеннями пікселів між оригінальним та захищеним зображенням [9].

Для обчислення *PSNR* використовується така формула:

$$PSNR(x, x') = -10 \cdot \log_{10}\left(MSE(x, x')\right), \tag{1}$$

де x — оригінальне зображення, x' — захищене зображення, а *MSE* — функція середньої квадратичної похибки.

Індекс структурної подібності або *SSIM (*від англ. *Structural Similarity Index Measure*) — це міра структурної невідповідності між оригінальним та захищеним зображенням, що визначається втратою кореляції, викривленням яскравості та викривленням контрасту [3].

Для обчислення *SSIM* використовується така формула:

$$SSIM(x, y) = \frac{\left(2\mu_x 2\mu_y + c_1\right)\left(2\delta_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\delta_x^2 + \delta_y^2 + c_2\right)}, \tag{2}$$

де x — оригінальне зображення; μ_x — середнє значення пікселів x; μ_y — середнє значення пікселів y; δ_{xy} — коваріантність x і y; δ_x^2 — дисперсія x; δ_y^2 — дисперсія y; c_1 — $\left(k_1 L\right)^2$; c_2 — $\left(k_2 L\right)^2$; L — динамічний діапазон значень пікселів.

Огляд засобів активного захисту

FakeTagger — це рішення, що працює на базі алгоритмів глибинного навчання, для позначення зображень з метою відслідковування походження Deepfake-контенту. Походження відслідковується за рахунок вбудовування прихованого повідомлення в файл. Якщо файл буде використано для створення Deepfake-контенту, приховане повідомлення можна буде верифікувати і визначити оригінальне зображення [4].

Таке рішення для чотирьох поширених видів Deepfake: заміна обличчя, перенесення жестикуляції, редагування атрибутів та повний синтез.

Хоч за словами авторів це рішення є доволі стійким, а приховане повідомлення в більшості випадків є непомітним для розпізнавання, воно має декілька недоліків. Приховане повідомлення вразливе до атак, за яких на захищене зображення накладається шум, через що повідомлення неможливо верифікувати. Також існує необхідність компромісу між непомітністю повідомлення та можливістю його верифікувати. Для того, щоб повідомлення в захищеному зображенні збереглося після маніпуляцій моделями GAN для створення Deepfake-зображення, в саме повідомлення додається надлишкова інформація, через що збільшується розмір повідомлення і, відповідно, шанс того, що повідомлення буде помітним.

FaceSigns — це система для захисту зображень напівкрихкими водяними знаками на базі глибинного машинного навчання. Цей метод використовується для аутентифікації візуальних медіа, що вони не були згенеровані моделями машинного навчання. Те, що водяний знак є напівкрихким в цьому контексті означає,, що водяний знак може бути зруйнований у разі маніпуляції моделями машинного навчання, наприклад для заміни обличчя, але не може бути зруйнований у разі звичніших і поширеніших маніпуляцій зображенням, як-от зміна розміру чи компресія.

Іншим методом є *PDDvIW* (від англ. *Proactive Deepfake Defence via Identity Watermarking)*. Автори пропонують метод використання водяних знаків, створених з використанням архітектури Encoder-Decoder і основаних на особливостях ідентифікації обличчя. Запропонований метод досягає середньої точності виявлення на рівні понад 80 %.

Введення водяного знаку складається з трьох етапів: виділення ознак, генерація водяного знаку та реконструкція зображення. На етапі виділення ознак дві мережі виділяють представлення ідентичності та представлення атрибутів. Модель Encoder для визначення ознак ідентичності отримує високорівневі біометричні ознаки людини, які дозволяють ідентифікувати конкретну особу. Encoder атрибутів виділяє просторові характеристики, такі як поза, вираз обличчя, фон тощо. Водяний знак генерується шляхом додавання побітової двійкової послідовності, яка може слугувати приватним ключем, до представлення ідентичності. Водяний знак накладається на оригінальне зображення на етапі реконструкції зображення.

Основним обмеженням цього методу є те, що він значною мірою покладається на ідентифікаційні ознаки обличчя і не може виявити маніпуляції з іншими частинами зображення, такими як фон або нелицьові атрибути тіла.

Цей метод не є вразливим до нешкідливих операцій обробки зображень, оскільки він не впливає безпосередньо на риси обличчя. Однак є незначні погіршення продуктивності, якщо вхідне зображення обробляється шляхом стиснення або розмиття [14].

FaceSigns — це проактивний метод виявлення Deepfake, який використовує водяні знаки, згенеровані на основі рис обличчя, замість того, щоб покладатися на виявлення артефактів або фіксовані схеми нанесення водяних знаків. Метод, який отримав назву FaceProtect, забезпечує точність виявлення, що перевищує 90 %, за різних маніпуляцій з ідентифікаційними даними та редагуванні атрибутів обличчя.

Метод складається з трьох основних компонентів: механізму односторонньої динамічної генерації водяних знаків на основі GAN (GODWGM), стратегії верифікації на основі водяних знаків (WVS) та модуля порівняння водяних знаків. GODWGM використовує 128-вимірні вектори рис обличчя як вхідні дані для генерації водяних знаків на основі зображень шляхом одностороннього незворотного відображення. Це не дозволяє зловмисникам реконструювати оригінальні риси обличчя і гарантує, що водяний знак динамічно змінюється разом зі вмістом зображення. Водяний знак вбудовується в оригінальне зображення за допомогою методів стеганографії, створюючи візуально ідентичне змішане зображення.

WVS складається з мережі приховування, що базується на U-Net і SENet, і мережі відновлення з використанням згорткової архітектури. Відновлення та перевірка водяних знаків виконується шляхом порівняння відновленого водяного знаку та водяного знаку на основі рис обличчя, витягнутого з отриманого зображення. Для порівняння використовується косинусоїдальна подібність з порогом 0,8, що вказує на автентичність.

Основним обмеженням методу є його залежність від рис обличчя; маніпуляції за межами області обличчя (наприклад, тіло або фон) залишаються невиявленими. До того ж стійкість до типових операцій обробки зображень, таких як стиснення та розмиття, безпосередньо не розглядається, хоча автори припускають, що включення таких перетворень у навчання може підвищити продуктивність.

Проте метод демонструє високі можливості узагальнення і чудову продуктивність порівняно як з пасивними методами виявлення, так і з наявними проактивними підходами. Він також перевершує метод бінарних послідовностей за стійкістю та можливістю відновлення завдяки використанню водяних знаків у відтінках сірого.

Порівняння наявних методів захисту, їхніх типів, архітектур моделей машинного навчання та вхідного ключа для генератора повідомлень подано в табл. 1.

Таблиця 1

Загальне порівняння методів

Метод	Тип захисту	Архітектура моделі	Вхід генератора секретних повідомлень
FakeTagger [21]	Позначення зображень	Encoder-Decoder	—
FaceSigns [13]	Напівкрихкий водяний знак	Encoder-Decode	128 біт
PDDvIW [14]	Водяний знак на основі ознак ідентичності	Encoder-Decode	512-вимірний вектор
FaceProtect [15]	Водяний знак на основі ознак ідентичності	Згорткова архітектура	128-вимірний вектор

В табл. 2 подано методи та їхні значення метрик непомітності PSNR та SSIM. За обома показниками найкраще себе демонструє FaceProtect, але в процесі навчання моделі вказаного методу у функції втрат не враховувались пертурбації зображень, як-от Гаусове розмивання чи компресія [15]. Функції втрат інших методів враховували це, хоч їхні характеристики не досягають показників FaceProtect.

Таблиця 2

Порівняння методів за метриками непомітності

Метод	SSIM	PSNR
FakeTagger (заміна обличчя) [21]	0,931	32,45
FakeTagger (перенесення жестикуляції) [21]	0,939	33,78
FakeTagger (повний синтез) [21]	0,948	35,21
FakeTagger (редагування атрибутів) [21]	0,942	34,70
FaceSigns (без трансформацій) [13]	0,973	36,38
FaceSigns (стійкий) [13]	0,971	35,91
FaceSigns (напівкрихкий) [13]	0,975	36,08
PDDvIW [14]	0,94	33,32
FaceProtect [15]	0,986	42,23

Подальші дослідження

Наявні методи проактивного захисту від Deepfake мають низку недоліків та вразливостей. Насамперед захищені повідомлення різних методів різною мірою можуть бути атаковані через додавання шуму, що генерується моделями машинного навчання таким чином, щоб модифікація зображення була візуально непомітною, але при цьому приховане повідомлення неможливо було верифікувати.

Іншими наявними проблемами є невелика кількість інформації, що може бути прихована, або ж необхідність додавати надлишок задля того, щоб повідомлення могло бути збережено попри різноманітні маніпуляції над зображенням.

Для подальших досліджень пропонується знаходження рішень для цих проблем.

Іншою проблемою є створення рішення, яке дозволить відслідковувати походження початкових медіафайлів, використаних для створення згенерованого контенту.

Висновки

В роботі розглянуто проактивні методи захисту від зловмисного використання технології Deepfake. Проведено порівняльний аналіз наявних засобів, що ґрунтуються на використанні при-

хованих повідомлень у вигляді водяних знаків, а також їхні недоліки та вразливості. Запропоновано напрямки подальших досліджень для вирішення зазначених проблем.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] *Pause Giant AI Experiments: An Open Letter*. [Electronic resource]. Available: https://futureoflife.org/open-letter/pause-giant-ai-experiments . Accessed: 26.03.2025.

[2] *Six Months Ago Elon Musk Called for a Pause on AI. Instead Development Sped Up*. [Electronic resource]. Available: https://www.wired.com/story/fast-forward-elon-musk-letter-pause-ai-development/ . Accessed: 26.03.2025.

[3] Z. Wang, E. P. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, 1398 p.

[4] R. Wang, F. Juefei-Xu., M. Luo., Y. Liu., and L. Wang, "FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking," *Proceedings of the 29th ACM International Conference on Multimedia*, 2020.

[5] E. L. Josephs, C. L. Fosco, and A. Oliva, "Artifact magnification on deepfake videos increases human detection and subjective confidence," *ArXiv*, 2023.

[6] S. D. Bray, S. D. Johnson, and B. Kleinberg, "Testing Human Ability To Detect Deepfake Images of Human Faces," *Journal of Cybersecurity*, 2022.

[7] T. Wang, X. Liao, K. Chow, X. Lin, and Y.Wang, "Deepfake Detection: A Comprehensive Survey from the Reliability Perspective," *ACM Computing Surveys*, 2022, 58 p.

[8] Y. Mirsky, and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys (CSUR)*, 2021. v. 54, 1 p.

[9] H. Nguyen-Le, V. Tran, D. Nguyen, and N. Le-Khac, "Deepfake Generation and Proactive Deepfake Defense," *A Comprehensive Survey*, 2024, 1 p. (Preprint)

[10] *Deepfake Cases Surge in Countries Holding 2024 Elections, Sumsub Research Shows.* [Electronic resource]. Available: https://sumsub.com/newsroom/deepfake-cases-surge-in-countries-holding-2024-elections-sumsub-research-shows/ . Accessed: 26.03.2025.

[11] A. Malanowska, W. Mazurczyk, T. K. Araghi, and D. Megías, "Digital Watermarking — A Meta-Survey and Techniques for Fake News Detection," *IEEE Access*, 2024, vol. 12, 36311 p.

[12] М. В. Рябокінь, Є. В. Котух, і О. І. «Папилев, Вплив технології Deepfake на Fintech сектор: поточний стан в Україні та стратегії запобігання кіберзлочинності,» *Проблеми і перспективи економіки та управління*, № 1 (37), с. 310-328, 2024.

[13] P. Neekhara, S. Hussain, X. Zhang, K. Huang, "FaceSigns: Semi-fragile Watermarks for Media Authentication," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, pp. 1-21, 2024.

[14] Y. Zhao, B. Liu, M. Ding, B. Liu, T. Zhu, and X. Yu, "Proactive Deepfake Defence via Identity Watermarking," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 4602-4611.

[15] S. Lan, K. Liu, Y. Zhao, C. Yang, "Facial Features Matter: a Dynamic Watermark based Proactive Deepfake Detection Approach," 2024, Preprint *arXiv*.

Рекомендована кафедрою захисту інформації ВНТУ

Стаття надійшла до редакції 10.04.2025

Марчук Михайло Борисович — аспірант кафедри захисту інформації, e-mail: dzgamech@gmail.com ;

Лукічов Віталій Володимирович — канд. техн. наук, доцент, доцент кафедри захисту інформації, e-mail: lukichov.vitalyi@vntu.edu.ua .

Вінницький національний технічний університет

M. B. Marchuk[1]
V. V. Lukichov[1]

Analysis of Methods and Tools of Proactive Defense against Deepfake

[1]Vinnytsia National Technical University

The development of artificial intelligence has been one of the main trends in recent years. Although this technology serves for the benefit of humanity, it can also be used for malicious purposes, such as spreading disinformation or blackmail. For the realization of the above-mentioned purposes, technologies are often used to create so-called Deepfake content.

The article describes the results of a study of methods and means for active protection against the malicious use of Deepfake. Deepfake is a general name for images, video or audio files, created by means of artificial neural networks and show how a person speaks or performs something that he did not perform in reality. As a result of using artificial intelligence such materials seem to be real for persons who do not know their origin. Without usage of special tools it will be difficult to distinguish the real image from faked one.

Technologies of the content generation by means of artificial intelligence, in particular Deepfake, are often used for the creation of the materials, used for disinformation campaigns, blackmail or other malicious purposes. In this connection, there is a need to develop means for protection against malicious use of Deepfake.

Passive protection methods based on the ability of machine learning models to distinguish authentic content from generated content depend on the architectures of models for generating Deepfake content, so they quickly become outdated as the latter develop more rapidly. Therefore, there is a need to develop active protection methods based on the use of watermarks to track them or to interfere with the operation of Deepfake generation models.

The paper describes and analyzes the existing active methods of protection against Deepfake, as well as analyzes the areas of their target applications, their advantages and disadvantages, technical characteristics, and suggests directions for further research in this area. Special attention is paid to the protection methods based on steganography and watermarking. The advantages of active protection methods over passive ones are considered.

Keywords: deepfake, steganography, watermarking, artificial intelligence, machine learning, information attack, information security.

Marchuk Mykhailo M. — Post-Graduate Student of the Chair of Information Security, e-mail: dzgamech@gmail.com ;

Lukichov Vitalii V. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Information Security, e-mail: lukichov.vitalyi@vntu.edu.ua