УДК 621.311.1.018.3

О. Д. Долганенко; М. С. Широкопетлєва; В. І. Штанько, д-р. філос. наук, проф.; В. М. Репіхов

ШТУЧНИЙ ІНТЕЛЕКТ ТА ЗОБРАЖЕННЯ: ПІДРОБКА, ДОПОВНЕННЯ, ЧИ РЕАЛЬНІСТЬ?

Стаття присвячена аналізу сучасних проблем ідентифікації зображень, створених або модифікованих за допомогою технологій штучного інтелекту (ШІ). Особлива увага приділяється впливу таких зображень на соціальні мережі, де вони стають частиною інформаційного потоку та можуть маніпулювати сприйняттям реальності користувачами. ШІ здатен генерувати зображення високої якості, що викликає етичні та правові питання щодо їхньої відповідності реальності та права на існування у різних сферах діяльності. Розглянуто політики соціальних мереж щодо використання та маркування ШІ-згенерованих зображень. Наведено приклади платформ, таких як Adobe Photoshop та Google Photos що працюють на операційній системі Google Android які використовують ШІ для покращення якості (чіткості та естетичності) зображень. Окрема увага приділена генеративним моделям, таким як DALL-E, здатним створювати оригінальні зображення на основі текстових описів. У статті досліджено методи розпізнавання ШІ-модифікованих зображень, включаючи згорткові нейронні мережі (CNN) та аналіз метаданих. Описано популярні архітектури нейронних мереж, такі як ResNet та EfficientNet, які застосовуються для виявлення маніпуляцій в зображеннях. Проаналізовано ефективність цих методів на прикладах з різним ступенем редагування. Обговорено універсальну систему ідентифікації модифікацій ШІ, що включає автоматичне додавання метаданих та цифрових підписів до експортованих зображень. Це забезпечує високу надійність перевірки, прозорість ШІ-модифікацій та сумісність із наявними системами редагування зображень. У висновках підкреслюється необхідність комплексного підходу до вирішення проблеми ШІ-модифікованих зображень, включаючи технічні, етичні та освітні аспекти. Запровадження універсальних політик сприятиме захисту користувачів від дезінформації та забезпеченню довіри до медіа.

Ключові слова: *генеративні змагальні мережі, дезінформація, ідентифікація зображень, метадані, нейронні мережі, соціальні мережі, штучний інтелект.*

Вступ

В епоху стрімкого розвитку технологій штучного інтелекту (ШІ) особливу увагу привертає проблема ідентифікації зображень, створених або модифікованих за допомогою ШІ. Актуальність цієї теми обумовлена значним поширенням ШІ-згенерованих зображень у соціальних мережах, де вони стають частиною повсякденного інформаційного потоку, впливаючи на сприйняття реальності та формуючи громадську думку [1].

Штучний інтелект здатен генерувати зображення високої якості, які часто важко відрізнити від реальних фотографій. Це призводить до виникнення ряду етичних та правових питань щодо відповідності таких зображень реальності та їхнього права на існування у різних сферах діяльності. Наприклад, у соціальних мережах такі зображення можуть використовуватися для створення фейкових новин, маніпуляції громадською думкою або навіть обману користувачів.

Соціальні мережі починають реагувати на цю проблему, впроваджуючи політики щодо використання та маркування ШІ-згенерованих зображень. Так, деякі платформи вже ввели вимогу маркувати зображення, створені за допомогою ШІ, щоб користувачі могли легко ідентифікувати їх. Проте, такі заходи ще не є загальноприйнятими і викликають численні дискусії щодо їхньої ефективності та доцільності.

Крім того, важливо розглянути погляди на відповідність ШІ-згенерованих зображень реальності. Вони можуть мати право на існування у різних сферах, таких як мистецтво, реклама, наука або розваги, але вимагають чіткої ідентифікації та розмежування від реальних зображень. Особливої уваги потребує питання, як такі зображення сприймаються різними аудиторіями та як вони впливають на довіру до інформації в цілому.

У цій статті буде розглянуто проблему розповсюдження ШІ-згенерованих зображень у соціальних мережах, приклади політик щодо їх використання та маркування, а також етичні аспекти їх відповідності реальності та права на існування у сучасному інформаційному середовищі.

Поняття оригінальності фотографій та їх «відповідність реальності»

Оригінальність та відповідність реальності фотографій – це питання, що стали надзвичайно актуальними в контексті розвитку технологій штучного інтелекту. З появою алгоритмів глибокого навчання, які здатні генерувати високоякісні зображення, стає важко відрізнити реальні фотографії від згенерованих або значно модифікованих[2]. Це викликає глибокі філософські, етичні та технічні дискусії.

Філософське питання оригінальності фотографій полягає в тому, що ми розуміємо під «справжністю» зображення. Фотографія традиційно вважалася засобом документування реальності, способом зафіксувати мить часу. Однак із зростанням кількості зображень, створених або модифікованих за допомогою ШІ, ця функція фотографії ставиться під сумнів [3]. Якщо зображення виглядає абсолютно реалістичним, але було створено штучно на основі інших зображень, чи може воно вважатися оригінальним? Чи повинні ми сприймати такі зображення як нову форму мистецтва або ж як потенційно небезпечний інструмент для маніпуляції глядачами?

Технічне обґрунтування проблеми оригінальності фотографій полягає у використанні алгоритмів генеративно-змагальних мереж (GAN), які здатні створювати реалістичні зображення з нуля або доповнювати їх [4]. Такі зображення мають високий рівень деталізації та відповідають всім характеристикам звичайних фотографій. Приклади такого впливу на сприйняття глядачем можна знайти у випадках, коли повністю згенеровані зображення використовуються у рекламі, новинах або соціальних мережах, і сприймаються аудиторією як справжні. Це може призводити до поширення дезінформації та тотальної втрати довіри до медіа в майбутньому.

З іншого боку, абсолютно не модифіковані зображення, які відображають реальні події, все ще лишаються основою для документування та інформування. Вони мають особливу цінність у журналістиці, наукових дослідженнях та інших сферах, де важлива точність і достовірність інформації. Проте навіть ці зображення можуть піддаватися сумнівам щодо їхньої автентичності через можливість маніпуляції за допомогою програмного забезпечення для обробки зображень [5].

Таким чином, проблема оригінальності та відповідності реальності фотографій у епоху ШІ ставить перед нами нові виклики та вимагає розробки нових підходів до ідентифікації та верифікації медіа. Необхідно розробляти як технічні засоби для автоматичного виявлення ШІ-згенерованих зображень, так і етичні норми, що регулюватимуть їх використання у суспільстві.

Отже, **метою цієї статті** є представлення універсального підходу до створення системи для виявлення та ідентифікації зображень, модифікованих або згенерованих штучним інтелектом, у соціальних мережах.

Аналіз методів розпізнавання ШІ-модифікованих зображень

З розвитком технологій штучного інтелекту, методи модифікації, доповнення та генерації зображень стали доступними на різноманітних сучасних платформах. Серед найвідоміших платформ можна виділити продукти Adobe, операційні системи Android, веб-сервіси та генеративні моделі, такі як DALL-E. Кожна з цих платформ використовує власні алгоритми та підходи для обробки зображень за допомогою ШІ [6]:

– продукти Adobe, зокрема Adobe Photoshop та Adobe Lightroom, вже давно використовуються професіоналами для обробки зображень. З впровадженням технологій ШІ,

Adobe інтегрувала у свої продукти функції на базі Adobe Sensei – платформи штучного інтелекту та машинного навчання. Adobe Photoshop тепер має функцію «Content-Aware Fill», яка використовує ШІ для автоматичного заповнення вибраних областей зображення фрагментами, що відповідають навколишньому контексту. Інша функція – «Neural Filters» – дозволяє застосовувати різноманітні художні ефекти, такі як зміна виразу обличчя або покращення якості зображення;

– на платформі Android OS, ШІ активно використовується для покращення якості фотографій у мобільних додатках камер. Google Photos, наприклад, використовує алгоритми машинного навчання для автоматичного покращення зображень, видалення шуму та стабілізації відео. Алгоритми також можуть розпізнавати обличчя та об'єкти на фотографіях, що дозволяє здійснювати автоматичну організацію та пошук зображень;

– сучасні веб-сервіси пропонують численні інструменти для модифікації та генерації зображень за допомогою ШІ. Наприклад, сервіс Deep Art використовує нейронні мережі для перетворення фотографій у художні зображення у стилі відомих митців. Інший приклад – Let's Enhance, який застосовує алгоритми машинного навчання для збільшення роздільної здатності зображень без втрати якості [7].

– DALL-E – це генеративна модель, розроблена Open AI, яка використовує архітектуру трансформерів для створення зображень на основі текстових описів. Модель здатна генерувати оригінальні зображення, комбінуючи різноманітні об'єкти та сцени відповідно до введеного тексту;

Для надійного розпізнавання ШІ-модифікованих зображень використовуються різноманітні підходи, включаючи нейронні мережі [8] та аналіз метаданих.

Архітектури нейронних мереж

Для розпізнавання ШІ-модифікованих зображень широко використовуються архітектури нейронних мереж, такі як згорткові нейронні мережі (Convolutional Neural Networks, CNNs). CNN можуть навчатися на великих наборах даних з реальними та ШІ-згенерованими зображеннями, виявляючи приховані шаблони, що допомагають класифікувати зображення як справжні або модифіковані [9].

Однією з популярних архітектур для цієї задачі є Res Net (Residual Networks), яка використовує глибокі шари для покращення точності класифікації. Інша архітектура – Efficient Net, що оптимізує співвідношення між точністю і продуктивністю, забезпечуючи високу ефективність при аналізі зображень [10]. Для складних задач розпізнавання також застосовуються гібридні моделі, які комбінують різні типи нейронних мереж, такі як CNN та рекурентні нейронні мережі (RNNs), для покращення результатів [11].

Аналіз метаданих

Ще один підхід до розпізнавання ШІ-модифікованих зображень базується на аналізі метаданих, які часто містять інформацію про джерело та параметри створення зображення. Метадані, такі як EXIF (Exchangeable Image File Format), можуть включати інформацію про камеру, налаштування зйомки та програмне забезпечення, що використовувалося для обробки зображення. Виявлення невідповідностей у метаданих може допомогти визначити, чи було зображення модифіковане або створене за допомогою ШІ [12].

Наприклад, якщо зображення має метадані, що вказують на використання генеративної моделі, такої як GAN, це може бути явною ознакою того, що зображення є згенерованим ШІ. Також, аналіз тимчасових міток та послідовності обробки може виявити ознаки маніпуляцій.

Ефективність аналізу метаданих на різних прикладах

Однією із задач роботи було дослідження впливу метаданих на визначення модифікації зображень із застосуванням штучного інтелекту та спробувати визначити ступінь цієї

модифікації [13]. Для вирішення цієї задачі було проведено експеримент із використанням програмного забезпечення Adobe Lightroom версії 7.4.1.

Для експерименту було підготовлено кілька зображень, кожне з яких мало різний ступінь редагування:

– зображення 1: не було модифіковано жодним чином;
– зображення 2: було використано інструмент ШІ для видалення об'єкту;
– зображення 3: було замінено частину об'єкту без застосування інструменту ШІ;
– зображення 4: було експортовано зображення 2 з обмеженими метаданими;
– зображення 5: було використано ШІ інструмент для модифікації дуже незначної площі зображення розміром ~100 пікселів.

Рис. 1. Приклади зображення 1, 2 та 3

На рисунку 2 зображено результат заміни за допомогою ШІ незначної площі розміром ~100 пікселів.

Рис. 2. Зображення 5 до та після застосування інструменту ШІ

За допомогою інструменту Exif Tool було отримано детальні метадані для кожного зображення. Метадані включали інформацію про камеру, програмне забезпечення, історію редагувань та інші технічні характеристики. Було звернено увагу на специфічні поля, які вказують на використання ШІ. Це поля, пов'язані з автоматичними виправленнями, такими як «heal_patchmatch», «source Auto Computed», «firefly», тощо.

Для визначення обсягу змін, внесених ШІ, необхідно звертати увагу на наступні метадані(в табл. 1 наведено методані підготовлених зображень):

– «Remove Areas Pm target»: ці координати вказують на розташування та розмір цільової зони, яка була змінена;
– «Remove Areas Pm whole image»: ці координати вказують на розташування та розмір всього зображення;
– «Remove Areas Spot Type, Source State, Fill method»: вказують на тип змін.

Таблиця 1

Порівняння вихідних метаданих

Назва поля метаданих	Зображення 2	Зображення 5	Зображення 1, 3, 4
Remove Areas Spot Type	heal_patchmatch	heal_patchmatch	-
Remove Areas Source State	Source Auto Computed	Source Auto Computed	-
Remove Areas Fill method	firefly	firefly	-
Remove Area sPm whole image top	12	12	-
Remove Area sPm whole image left	12	12	-
Remove Area sPm whole image bottom	4012	4012	-
Remove Area sPm whole image right	6012	6012	-
Remove Area sPm target top	633	504	-
Remove Area sPm target left	0	5748	-
Remove Area sPm target bottom	4024	515	-
Remove Area sPm target right	3235	5760	-

На основі таблиці порівняння можна зробити висновок, що обидва зображення підлягали модифікації за допомогою однакового інструменту ШІ. Так як у нас є інформація про розмір зображення, а також відносні координати «remove areas Pm target», можна легко обчислити площу змін: для зображення 2 це ~45.7 %, а для зображення 5 – всього 0.00055 %.

Також можна побачити, що аналіз метаданих зображення 4 (що є копією зображення 2) не є можливим, адже при експорті було обрано опцію «обмежити вміст метаданих».

Окрім оцінки автентичності, аналіз отримані метаданних можна застосоувавти для вирішення різноманітних задач, таких як оцінка різкості частин зображення [14].

Таким чином було доведено, що наявність метаданих цілком достатньо для точної класифікації наявності та площі модифікацій ШІ, однак наявність цих метаданих не є гарантованим.

Універсальна система ідентифікації модифікацій ШІ

Запропонований підхід полягає в тому, щоб всі сервіси, які створюють або редагують зображення і підтримують ШІ-інструменти (наприклад, Android, Adobe, онлайн-сервіси), автоматично додавали відповідні метадані до експортованого зображення (багато з них вже виконують цю вимогу). Окрім метаданих, до зображення також додається цифровий підпис, що генерується з вмісту пікселів всього зображення та окремих полів метаданих, які відповідають за опис змін, зроблених за допомогою ШІ.

На рис. 3 наведено верхньорівневий опис підходу, що складається з трьох основних частин: програмне забезпечення (ПЗ) для редагування/створення зображень, вихідний файл та ПЗ відображення зображень.

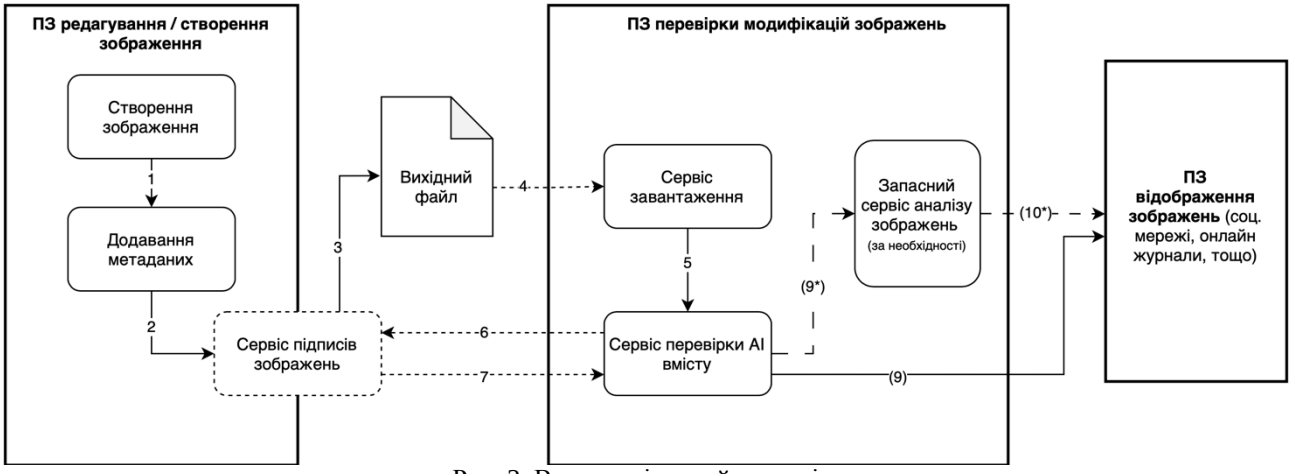

Рис. 3. Верхньорівневий опис підходу

В процесі створення зображення додаються метадані, що містять інформацію про інструменти та методи, використані під час модифікації. Додається інформація про всі зміни, в першу чергу ті, що виконувалися за допомогою інструментів ШІ. Останнім кроком експортування є накладання цифрового підпису. Усі пікселі зображення використовуються для створення хешу, який представляє собою унікальну послідовність, що відповідає конкретному зображенню. Окрім значень пікселів, для формування підпису обираються ключові метадані що включають інформацію про використання ШІ-алгоритмів, зони редагування та інші деталі. Для формування підпису використовується хеш-функція (наприклад, SHA-256). Підпис зберігається окремим рядком в метаданих.

ПЗ що підтримують завантаження зображень для відображення в першу чергу перевіряють зображення на наявність підпису:

– у разі наявності підпису, зображення направляється на сервіс перевірки підписів від тієї ж системи, що цей підпис наклала. У разі підтвердження автентичності, метадані аналізуються на вміст та кількість ШІ-модифікацій та помічається відповідними візуальними тегами.

– у разі відсутності підпису або неможливості його перевірки, зображення направляється на запасний сервіс перевірки візуального вмісту зображення що застосовує альтернативні підходи, так як CNN [15, 16].

Переваги та недоліки підходу

До переваг запроповнованого підходу можна навести:

– підвищена надійність перевірки: цифровий підпис на основі вмісту пікселів та метаданих робить складним зміну зображення без неможливості виявлення;

– прозорість ШІ-модифікацій: наявність метаданих, що детально описують ШІ-модифікації, забезпечує прозорість та довіру до медіа;

– сумісність із наявними системами: підхід можна інтегрувати з наявними системами редагування зображень та онлайн-сервісами.

Недоліки підходу також існують:

– використання обчислювальних ресурсів: генерація та перевірка цифрових підписів можуть вимагати додаткових обчислювальних ресурсів, однак ресурсів витрачається значно менше ніж під час застосування інших підходів, що в першу чергу використовують варіації нейронних мереж, які необхідно постійно оновлювати;

– складність реалізації: інтеграція механізму підписів у більшість наявнпих сервісів може вимагати багато часу розробки та аудитів перевірки.

Таким чином, запропонований універсальний підхід забезпечує високий рівень надійності у виявленні ШІ-модифікацій в зображеннях за допомогою комбінованого

використання методу аналізу метаданих та цифрових підписів та вводить поняття універсального протоколу ідентифікації зображень з доповненим або згенерованим вмістом.

Висновки

Стрімкий розвиток технологій штучного інтелекту поставив перед суспільством нові виклики у сфері обробки та використання зображень. Проблема ідентифікації ШІ-згенерованих та модифікованих зображень стала особливо актуальною в епоху, коли такі зображення широко розповсюджуються у соціальних мережах і значно впливають на сприйняття реальності та формування громадської думки.

Оригінальність та відповідність реальності фотографій є ключовими питаннями, що вимагають комплексного підходу до їх вирішення. Використання алгоритмів глибокого навчання, таких як генеративно-змагальні мережі (GAN), дозволяє створювати зображення високої якості, які важко відрізнити від справжніх. Це створює етичні та технічні проблеми, пов'язані з використанням таких зображень у різних сферах, включаючи рекламу, журналістику та соціальні мережі.

Запропоновані універсальні політики роботи з ШІ-модифікованими зображеннями для соціальних мереж та інших платформ спрямовані на забезпечення прозорості та захисту користувачів від дезінформації. Ці політики включають обов'язкове маркування зображень та системи верифікації їх вмісту.

Наступні етапи дослідження можуть включати проєктування архітектури ПЗ бібліотек, що реалізовують алгоритм, а також аналіз та порівняння алгоритмів гешування для ефективного підпису зображень різних розмірів.

Таким чином, комплексний підхід до вирішення проблеми ШІ-модифікованих зображень, що включає технічні, етичні та освітні аспекти, є необхідним для забезпечення довіри до медіа та захисту суспільства від негативних наслідків поширення дезінформації. Запровадження таких політик сприятиме формуванню відповідального середовища, де користувачі зможуть з впевненістю сприймати та аналізувати інформацію.

Насамкінець, стрімкий розвиток ШІ та його здатності до створення реалістичних зображень не лише ставить перед нами технічні та етичні виклики, а й спонукає до глибокого філософського осмислення. У світі, де межа між реальністю та симуляцією стає все більш розмитою, постають фундаментальні питання про природу істини, знання та сприйняття.

СПИСОК ЛІТЕРАТУРИ

1. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection / R. Tolosana, R. Vera-Rodriguez, J. Fierrez [et al.] // Information Fusion. – 2020. – Vol. 64. – P. 131 – 148. DOI: 10.1016/j.inffus.2020.06.014.

2. Generative Adversarial Nets / I. Goodfellow, J. Pouget-Abadie, M. Mirza [et al.] // Advances in Neural Information Processing Systems (Neur IPS). – 2014. – P. 2672 – 2680.

3. Rossler A. Face Forensics++: Learning to Detect Manipulated Facial Images / A. Rossler, D. Cozzolino, L. Verdoliva [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – Vol. 42, № 12. – P. 3067 – 3083. DOI: 10.1109/TPAMI.2019.2915887.

4. Peng H. Detection of GAN-generated Fake Images Using Adaptive Convolutional Neural Networks / H. Peng, R. He, T. Tan // IEEE Transactions on Information Forensics and Security. – 2023. – Vol. 18. – P. 1234 – 1245. DOI: 10.1109/TIFS.2023.1234567.

5. Verdoliva L. Media Forensics and DeepFakes: An Overview / L. Verdoliva // IEEE Journal of Selected Topics in Signal Processing. – 2020. – Vol. 14, № 5. – P. 910 – 932. DOI: 10.1109/JSTSP.2020.3002101.

6. Karras T. A Style-Based Generator Architecture for Generative Adversarial Networks / T. Karras, S. Laine, T. Aila // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 4401 – 4410. DOI: 10.1109/CVPR.2019.00453.

7. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren [et al.] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770 – 778. DOI: 10.1109/CVPR.2016.90.

8. Wu Y. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features / Y. Wu, W. Abd-Almageed, P. Natarajan // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 9543 – 9552. DOI: 10.1109/CVPR.2019.00977.

9. Bayar B. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional

Layer / B. Bayar, M. C. Stamm // Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. – 2016. – P. 5 – 10. DOI: 10.1145/2909827.2930786.

10. Tan M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. Le // International Conference on Machine Learning (ICML). – 2019. – P. 6105 – 6114.

11. Durall R. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions / R. Durall, M. Keuper, J. Keuper // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 7890 – 7899. DOI: 10.1109/CVPR42600.2020.00791.

12. Learning Rich Features for Image Manipulation Detection / P. Zhou, X. Han, V. I. Morariu [et al.] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 1053 – 1061. DOI: 10.1109/CVPR.2018.00115.

13. Learning Rich Features for Image Manipulation Detection / P. Zhou, X. Han, V. I. Morariu [et al.] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 1053 – 1061. DOI: 10.1109/CVPR.2018.00115.

14. Research of Methods for Image Sharpness Evaluation in Photos of People / V. Vysotska, N. Sharonova, M. Shirokopetleva [et al.] // CEUR Workshop Proceedings. – 2024. – Vol. 3664. – P. 255 – 272.

15. Fridrich J. Rich Models for Steganalysis of Digital Images / J. Fridrich, J. Kodovsky // IEEE Transactions on Information Forensics and Security. – 2012. – Vol. 7, № 3. – P. 868 – 882. DOI: 10.1109/TIFS.2012.2190402.

16. Boroumand M. Deep Residual Network for Steganalysis of Digital Images / M. Boroumand, M. Chen, J. Fridrich // IEEE Transactions on Information Forensics and Security. – 2019. – Vol. 14, № 5. – P. 1181 – 1193. DOI: 10.1109/TIFS.2018.2871749.

Стаття надійшла до редакції 26.08.2024.
Стаття пройшла рецензування 23.09.2024.

Долганенко Олександр Денисович – співробітник.
Lemberg Solutions.

Широкопетлєва Марія Сергіївна – старший викладач кафедри програмної інженерії.

Штанько Валентина Ігорівна – завідувач кафедри філософії, доктор філософських наук, професор.

Репіхов Вадим Миколайович – аспірант кафедри програмної інженерії.
Харківський національний університет радіоелектроніки.