УДК 004.056.55:004.932.4:004.9
DOI https://doi.org/10.32689/maup.it.2025.1.4

Денис БУКАТОВ
викладач кафедри математичних дисциплін та інноваційного проектування,
Приватний вищий навчальний заклад «Європейський університет», denys.bukatov@e-u.edu.ua
***ORCID:** 0009-0009-9320-2181*

Яніна КОЛОДІНСЬКА
старший викладач кафедри математичних дисциплін та інноваційного проектування,
Приватний вищий навчальний заклад «Європейський університет», yanina.kolodinska@e-u.edu.ua
***ORCID:** 0000-0002-3330-7565*
***Scopus Author ID:** 57466561800*

Ігор ФРОЛОВ
викладач кафедри математичних дисциплін та інноваційного проектування
Приватний вищий навчальний заклад «Європейський університет», ihor.frolov@e-u.edu.ua
***ORCID:** 0009-0004-1163-8787*

Олег РОМАНЕНКО
викладач кафедри математичних дисциплін та інноваційного проектування,
Приватний вищий навчальний заклад «Європейський університет», oleg.romanenko@e-u.edu.ua
***ORCID:** 0009-0008-4703-217X*

АКТУАЛЬНІ ЗАГРОЗИ ТА ЗАХИСНІ СТРАТЕГІЇ ДЛЯ ВІЗУАЛЬНОГО КОНТЕНТУ В ІНТЕРАКТИВНИХ СИСТЕМАХ ІТ ПРОЄКТІВ

Анотація. *У сучасних умовах цифровізації та стрімкого розвитку штучного інтелекту захист візуального контенту набуває особливої актуальності. Зростання загроз підробки, викрадення та несанкціонованого використання цифрових зображень обумовлює потребу у вдосконаленні методів безпеки та розробці інноваційних стратегій протидії порушенням авторських прав.*

Мета роботи. *Мета даної роботи полягає в аналізі сучасних методів захисту візуального контенту в інтерактивних ІТ-проєктах, з урахуванням стрімкого розвитку цифрових технологій, зростання ризиків підробки цифрового контенту, недосконалості правових норм у сфері інтелектуальної власності та впливу штучного інтелекту.*

Методологія. *У дослідженні застосовано метод порівняльного аналізу існуючих технік захисту цифрового контенту, включаючи шифрування, водяні знаки, обмеження доступу, застосування метаданих, нейромережеві підходи, а також вивчено методи адаверсаріальних атак, зокрема технологію Nightshade. Особливу увагу приділено інтеграції заходів захисту на рівні інтерфейсів користувача та технічних обмежень у вебсередовищах, що дозволяє забезпечити комплексний підхід до безпеки мультимедійних даних.*

Наукова новизна. *Наукова новизна полягає у комплексному розгляді новітніх технологій захисту візуального контенту від автоматизованого збору і використання штучними інтелектами, виявленні ролі метаданих для підтвердження автентичності цифрових зображень, а також запропонуванні сучасних практик застосування нейромереж для виявлення маніпуляцій із візуальним контентом з урахуванням актуальних викликів кібербезпеки.*

Висновки. *Стаття підкреслює важливість багаторівневого захисту цифрового контенту з використанням поєднання різних підходів – від шифрування і водяних знаків до нейромережевих технологій та спеціальних протиадаверсаріальних методів. Виявлено, що ефективна стратегія має охоплювати як технічні, так і організаційні заходи безпеки, забезпечуючи гнучкість систем захисту в умовах постійно зростаючих цифрових загроз. Перспективи подальших досліджень визначені у напрямку розробки нових алгоритмів захисту, інтеграції їх у ІТ-продукти, та оптимізації існуючих методів для підвищення стійкості до сучасних викликів.*

Ключові слова: *методи захисту, нейромережі, адаверсаріальні атаки, захист цифрового контенту, цифровий контент.*

Denys BUKATOV, Yanina KOLODINSKA, Ihor FROLOV, Oleh ROMANENKO. CURRENT THREATS AND PROTECTION STRATEGIES FOR VISUAL CONTENT IN INTERACTIVE IT PROJECT SYSTEMS

Abstract. *In the context of rapid digitalization and the accelerated development of artificial intelligence technologies, protecting visual content has become increasingly important. The growing threats of forgery, theft, and unauthorized use of digital images necessitate the improvement of security methods and the development of innovative strategies to counter violations of intellectual property rights.*

Purpose of the study. *The aim of this work is to analyze modern methods of protecting visual content in interactive IT projects, taking into account the rapid development of digital technologies, the increasing risks of digital content forgery, the imperfections of intellectual property regulations, and the influence of artificial intelligence.*

Methodology. *The study employs a comparative analysis of existing digital content protection techniques, including encryption, watermarking, access restrictions, metadata application, and neural network–based approaches. It also examines*

adversarial attack methods, particularly the Nightshade technology. Special attention is paid to the integration of user interface–level protections and technical restrictions in web environments, ensuring a comprehensive approach to the security of multimedia data.

Scientific novelty. *The scientific novelty lies in the comprehensive consideration of advanced technologies for protecting visual content from automated collection and use by artificial intelligence systems. The study highlights the critical role of metadata in verifying the authenticity of digital images and suggests modern practices for applying neural networks to detect visual content manipulations, taking into account current cybersecurity challenges.*

Conclusions. *The article emphasizes the importance of multi-level protection of digital content by combining various approaches – from encryption and watermarking to neural network technologies and specialized adversarial methods. It has been found that an effective strategy must encompass both technical and organizational security measures, ensuring flexibility of security systems amid the ever-growing digital threats. Future research directions are defined in the development of new protection algorithms, their integration into IT products, and the optimization of existing methods to enhance resilience to modern challenges.*

Key words: *protection methods, neural networks, adversarial attacks, digital content protection, digital content.*

Постановка проблеми та її актуальність. Захист цифрової інформації є надзвичайно важливою умовою у сучасних ІТ-проєктах, де головну роль відіграє візуальна взаємодія з користувачем.

Особливо це важливо у сферах розробки та створення візуально-цифрового контенту. Зображення, 3D-моделі, візуальні елементи інтерфейсу, анімації та відео часто є не лише інструментами комунікації, а й частиною критично важливої логіки застосунку. При цьому саме ці компоненти найчастіше стають об'єктами підробки, викрадення або несанкціонованого використання – як з боку користувачів, так і автоматизованих систем штучного інтелекту.

З розвитком інструментів реверс-інжинірингу, автоматичного видалення водяних знаків, діпфейків, генеративних моделей та хмарних сервісів для зберігання візуального контенту зросли і ризики. Проблема полягає не лише у витоку даних або несанкціонованому доступі, а й у можливості маніпулювання зображеннями в інтерактивному середовищі, підміні вмісту, або використанні оригінальних візуальних елементів для тренування сторонніх ШІ-моделей без згоди правовласника.

Для захисту візуального контенту є різноманітні методи захисту, такі як шифрування, обмеження доступу, водяні знаки, навмисне зменшення роздільної здатності зображень та інші.

Додаткову складність становить потреба збалансувати захист із безперебійною інтерактивністю: надмірні обмеження або погіршення якості відображення може негативно вплинути на користувацький досвід.

Аналіз останніх досліджень і публікацій. Питання захисту візуального контенту в сучасних ІТ-проєктах стає дедалі актуальнішим у зв'язку з широким використанням цифрових зображень, відео, UI-компонентів та іншого конткенту. Розвиток хмарних технологій, зростання популярності гібридних застосунків, поширення генеративного ШІ та нових форматів зберігання й передачі мультимедійних даних обумовлюють необхідність формування нових стратегій цифрового захисту.

У роботах останніх років особлива увага приділяється технологіям водяного маркування [10, с. 110; 4, с. 2224–2225]. Так, M. Begum і M. Uddin підкреслюють значення видимих і прихованих watermark для підтвердження авторства. У дослідженні A. F. Eldaoushy та ін. запропоновано гібридний підхід до вставки водяних знаків із використанням спектральних перетворень.

Мета статті полягає у проведенні більш детального аналізу сучасних методів захисту візуального контенту в інтерактивних ІТ-системах, таких як вебзастосунки, мобільні платформи, ігрові проєкти та. Основна увага приділяється методам, що реалізуються у кінцевих продуктах: шифрування графічних ресурсів, застосування видимих і невидимих водяних знаків (Begum & Uddin, 2020; Eldaoushy et al., 2023), захист на основі метаданих і цифрових підписів (El-Hajj et al., 2019), а також технологіям протидії ШІ-моделям, як-от Nightshade (Zhao et al., 2023).

Виклад основного матеріалу дослідження. Для захисту цифрових зображень є дуже багато різноманітних методів захисту, такі як шифрування, водяні знаки, обмеження доступу, навмисне зменшення роздільної здатності та інші методи.

Одним із базових підходів до захисту візуального контенту є шифрування, яке передбачає перетворення вмісту у вигляд, незрозумілий для стороннього користувача. На практиці застосовують як симетричне AES шифрування, так і асиметричні алгоритми на кшталт RSA, а для мультимедійних даних використовуються спеціалізовані протоколи безпечної передачі [3, с. 186958–186973].

Нейромережі та машинне навчання можуть бути застосовані до перевірки цифрових зображень. Наприклад, застосування нейронних мереж для виявлення змін колірних діапазонів та інших форм маніпулювання зображеннями дозволяє виявляти спроби підробки цих зображень.

Захист цифрових зображень вимагає використання різноманітних та всебічних технік та підходів.

В художній індустрії фахівці можуть використовувати різноманітні методи та інструменти для захисту свого контенту. Одним з таких методів є водяний знак, який використовується для захисту авторських прав та належності зображення до конкретного автора.

Варто зазначити, що в сучасних ІТ-проєктах, де розробка візуального контенту інтегрується у бізнес-моделі ціифрових проєктів, питання безпеки зображень набуває практичного значення вже на етапі формування інноваційної ідеї. Зокрема, дослідження Колодінської Я.О., Скляренко О.В. та Ніколаєвського О.Ю. показує, що цифрові сервіси активно використовуються як платформи для реалізації стартапів, де візуальний контент є ключовим елементом презентації, маркетингу та взаємодії з цільовою аудиторією [11, с. 53–60]. У такому контексті захист графічних матеріалів має не лише технічний, а й стратегічний характер.

Іншим методом захисту є техніка блокування копіювання, що дозволяє обмежити можливість копіювання зображень без дозволу автора. Ця техніка може бути реалізована за допомогою вбудованих обмежень в програмах для перегляду та завантаження зображень або шляхом застосування спеціальних налаштувань функціоналу сайту.

Ще одним методом захисту а точніше перевірки та отримання додаткової інформації про зображення є використання метаданих, які дозволяють дізнатись додаткові дані про зображення.

Метадані дозволяють зберігати додаткову інформацію: дані про автора, дату створення чи дані програмного забезпечення. У сучасних підходах нерідко застосовують приховане маркування через контрольні суми або цифрові підписи, що дає змогу простежити автентичність візуального матеріалу навіть після його редагування [5].

Також, дуже важливо проводити регулярну перевірку мережі на наявність незаконного використання зображень та вживати заходів щодо їх вилучення. З цим дуже добре допомагають сервіси пошуку за зображенням, зокрема Google Reverse Image Search або TinEye, а також спеціалізовані платформи захисту авторських прав, наприклад Pixsy чи Digimarc, що можуть відслідкувати копії контенту в мережі.

Серед інших методів є такі варіанти як використання промптів, шуму, обмеження функціональності на сайтах і тд.

Найпоширеніші варіанти:

Обмеження доступу (Access restrictions)

У практиці розробки вебзастосунків обмеження доступу до зображень може бути реалізоване через заборону функцій контекстного меню за допомогою JavaScript-функцій, наприклад, event. preventDefault() або return false; у відповідь на події типу contextmenu. Хоча цей метод не гарантує абсолютного захисту, він ускладнить базове копіювання зображення. Встановлення невидимих перекриттів на елементи графіки або запровадження політик безпеки на рівні заголовків HTTP, таких як Content Security Policy (CSP), що обмежує джерела, з яких може бути завантажений контент, а також забороняє вбудовування зображень на сторонніх ресурсах (anti-hotlinking) [9].

Інтеграція промптів (Prompts integration)

Використання промптів у зображеннях може стати ефективним способом ускладнення або запобігання використанню зображень для навчання ШІ без дозволу.

Промпти, які вбудовують текстові або бінарні дані в зображення, дозволяють передавати приховану інформацію, що може бути використана для верифікації автентичності зображення або його власника, або як протилежний варіант може використовуватись для приховування шкідливого «коду» або «функцій» всередині здається невинних зображень, або самих моделей які потім можуть бути поширені через електронну пошту, соціальні мережі або звичайні сайти. Як приклад це небезпечно у випадках, якщо використовується модель ШІ яка може не тільки аналізувати а й виконувати «функції» записані у зображення.

Шум (Noise)

Техніка додавання шуму вимагає ретельного підходу до вибору типу та інтенсивності шуму. Важливо збалансувати між маскуванням інформації та збереженням якості зображення. Один із способів – використання чорно-білого шуму, який випадково змінює значення пікселів на чорне або біле. Цей метод ефективний для зображень з високим контрастом, де такі точки будуть менш помітні. Для більш плавних зображень застосовується спектральний шум, що вносить зміни на рівні частотних компонентів, роблячи зміни менш очевидними для людського ока, але ускладнюючи автоматизовану обробку зображень.

Колірні канали (Color channels)

Ще як варіант іноді використовують модифікацію каналів зображення. Варіювання кольорових каналів у спосіб, який робить дані непридатними для навчання ШІ [8] і загалом робить зображення менш чітким та візуально перенасиченим, наприклад, через алгоритми, що зміщують колірні гістограми

зображення [1]. Найбільша проблема методу у тому що у grayscale зображені не зсунеш канал кольору так як залишається всього один варіант градації каналу кольору, від чорного до білого. Цей grayscale можна призначити будь якому колірному каналу але лише для оного єдиного і відображається як під ефектом одноколірного фільтру, інакше зображення втрачає колірну глибину.

Адаверсаріальні атаки (Adversarial attacks)

Також для захисту можна примінити «адаверсальні атаки».

Адаверсаріальні атаки – це методи маніпуляції даними для створення ситуацій в яких алгоритми машинного навчання або нейронні мережі приймають неправильні рішення при аналізі даних. Цей процес включає зміну або створення вхідних даних, які на перший погляд здаються нормальними, але спричиняють помилки в роботі алгоритмів. Загалом це спеціально модифіковані вхідні дані, які призводять до помилок у класифікації зображень або прогнозуванні моделі ШІ [6].

Наприклад, до зображення додаються незначні зміни (шум), які майже непомітні для людського ока, але значно змінюють результат класифікації.

Такі типи атак використовують для:

– маніпуляції вхідними даними так, щоб вони уникли виявлення або неправильного класифікування

– модифікації (отруєння) навчальних даних для введення помилок у модель

– відтворення або отримання доступу до внутрішньої логіки або параметрів моделі

Водяні знаки (Watermark)

Водяні знаки, як правило, використовуються для додавання невидимих або малопомітних для ока штампів на зображення, що дозволяє захистити авторські права та запобігти незаконному розповсюдженню. Такі водяні знаки можуть бути вбудовані у формі символів, логотипів або інших графічних елементів, які стають помітними тільки після спеціальної обробки зображення та його ретельного обстеження, забезпечуючи ефективний захист без значного зниження якості. Або є ще варіанти з використанням навпаки сильно помітних або напівпрозорих водяних знаків або їх модифікованою версією з використанням «Blur» методів.

Nightshade

Nightshade – це технологія яка дозволяє захищати цифрові зображення від несанкціонованого використання, включаючи захист від тренування штучного інтелекту на основі цих зображень шляхом їх модифікації. Це теж є методом який відносять до типу адаверсаріальних атак [2].

Основна ідея Nightshade полягає в тому, щоб змінювати пікселі зображення таким чином, щоб вони виглядали для людини нормально, але створювали значні труднощі для алгоритмів ШІ. (Мікрорівневі зміни (Pixel-Level Alterations))

Зміни на рівні пікселів викликають помилки у моделях машинного навчання, які використовуються для розпізнавання або класифікації зображень.

Зображення, оброблені Nightshade, можуть негативно вплинути на модель ШІ. Це означає, що алгоритми не зможуть правильно інтерпретувати ці зображення або використовувати їх для навчання моделей. Часто результат такого методу залежить від кількості «отруєних» зображень які використовуються у моделях ШІ, чим їх більше тим довше модель буде аналізувати дані.

Крім Nightshade, є ще технологія Glaze, яка змінює стилістичні характеристики зображення, ускладнюючи для генеративних моделей точне копіювання художнього стилю автора без помітного спотворення самого зображення візуально, хоча технологія не ідеальна але це принаймні уповільнює час навчання ШІ-моделі специфічним характеристикам стилю автора.

Висновки. У цій статті було проведено всебічний аналіз сучасних загроз і методів захисту візуального контенту, що використовується в інтерактивних системах ІТ-проєктів. Було встановлено, що сучасні методи захисту контенту, такі як промпти, шум, модифікація колірних каналів, водяні знаки, обмеження доступу та нейромережеві підходи, є ефективними для забезпечення безпеки цифрового контенту. Використання цих методів показало свою доцільність у поточному стані індустрії.

Проаналізовані техніки – включаючи шифрування, обмеження доступу, використання метаданих, watermarking, модифікацію колірних каналів, а також додавання шуму та застосування спеціальних «промптів» – демонструють свою ефективність у захисті зображень у цифровому середовищі. Значну увагу приділено нейромережевим та машинно-навчальним підходам, які застосовуються як для виявлення маніпуляцій, так і для формування нових засобів цифрової ідентифікації.

Перспективи подальшого розвитку у даному напрямі включають:

– інтеграцію захисних технологій у сучасні платформи та інструменти;

– розробку нових алгоритмів та методів, які покращать стійкість додатків та програм до різних типів атак вцілому;

– оптимізацію та адаптацію існуючих методів захисту для підвищення та пролонгованості підтримки продуктів старого та нового типу.

Список використаних джерел:

1. Колодінська Я. О., Скляренко О. В., Ніколаєвський О. Ю. Практичні аспекти розробки інноваційних бізнес-ідей з використанням цифрових сервісів. *Економіка і управління*. 2022. № 4. С. 53–60. DOI: https://doi.org/10.36919/2312-7812.4.2022.53.

2. Kumar S. A review on attacks and defense strategies in generative AI models. Artificial Intelligence Review. 2023. DOI: 10.1007/s10462-023-10450-2.

3. Liu S., Pan Z., Song H. An improved image watermarking method based on DCT and fractal coding. *Signal Processing: Image Communication*. 2021. Vol. 94. P. 116205.

4. Begum M., Uddin M.S. Digital image watermarking techniques: A review. *Information*. 2020. Vol. 11. No. 2. P. 110.

5. Eldaoushy A. F., Desouky M. I., El-Dolil S. A., El-Fishawy A. S., Abd El-Samie F. E. Efficient hybrid digital image watermarking. *Journal of Optics*. 2023. Vol. 52. No. 11. P. 2224–2238.

6. Epic Games. Unreal Engine 5 Documentation: Pak File Encryption. 2022. URL: https://docs.unrealengine.com/5.0/en-US/encrypting-pak-files-in-unreal-engine (дата звернення: 25.04.2025).

7. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. Proceedings from ICLR 2015: International Conference on Learning Representations. 2015. URL: https://arxiv.org/abs/1412.6572 (дата звернення: 25.04.2025).

8. Figma Inc. Security at Figma. 2023. URL: https://www.figma.com/security (дата звернення: 25.04.2025).

9. Zhao B., Yang J., Li Y., et al. Nightshade: Poisoning data to prevent unauthorized training on generative models. Proc. of the ACM Conf. on Computer and Communications Security (CCS '23). 2023. URL: https://nightshade.cs.uchicago.edu (дата звернення: 25.04.2025).

10. Robust watermarking of medical images using DICOM standards. *Scientific Reports*. 2021. Vol. 11. Article 11820. DOI: https://doi.org/10.1038/s41598-021-91035-z.

11. El-Hajj M., Fadlallah A., Chamoun M., Serhrouchni A., et al. Secure and lightweight image encryption technique for IoT applications. *IEEE Access*. 2020. Vol. 8. P. 186958–186973.