До проблеми методів детекції медіаконтенту, створеного за допомогою генеративного штучного інтелекту, під час проведення судової експертизи

Ігор Римар

кандидат історичних наук, завідувач відділу дослідження звуко- та відеозаписів, лабораторія досліджень у сфері інформаційних технологій, Київський науково-дослідний експертно-криміналістичний центр Міністерства внутрішніх справ України, м. Київ, Україна, ORCID: https://orcid.org/0000-0001-6153-0900, e-mail: ihor.rymar@outlook.com

Ганна Тулайдан

старший судовий експерт відділу фототехнічних та портретних досліджень, лабораторія досліджень у сфері інформаційних технологій, Київський науково-дослідний експертно-криміналістичний центр Міністерства внутрішніх справ України, м. Київ, Україна, ORCID: https://orcid.org/0009-0003-1192-5981, e-mail: annatulaydan@gmail.com

Досліджено виклики генеративного штучного інтелекту для судової експертизи. Розглянуто основні архітектурні підходи тренування генеративного штучного інтелекту. Систематизовано основні підходи до автентифікації медіа. Наголошено на потребі в нових методиках для верифікації синтетичного контенту.

Ключові слова: *судова експертиза; дипфейк; генеративний штучний інтелект; генеративно-змагальна мережа; дифузійна модель.*

On The Issue of Detection Methods for Media Content Created by Generative AI in Forensic Examination

Ihor Rymar, Anna Tulaydan

The challenges of generative AI for digital forensics are investigated. The core architectural approaches to training generative artificial intelligence are examined. The main methods for media authentication are systematized. The necessity for developing new techniques for synthetic content verification is emphasized.

Keywords*: forensic examination; deepfake; generative artificial intelligence; generative adversarial network; diffusion model.*

Епоха генеративного штучного інтелекту (далі — *ШІ*) кардинально змінює ландшафт цифрової криміналістики та судової експертизи. Сучасні технології дають змогу створювати синтетичні тексти, зображення, відео- та аудіо- настільки реалістичними, що їх неможливо відрізнити від автентичного контенту неозброєним оком [1, 16]. Ця технологічна революція породжує феномен «*упередження самозванця*» (англ. *impostor bias*) — когнітивного упередження, за якого люди починають сумніватися в автентичності будь-якого мультимедійного контенту через усвідомлення можливостей ШІ [1].

У контексті судової експертизи проблема ідентифікації ШІ-генерованого контенту набуває критичного значення. *Дипфейки* — синтетичні медіа, що імітують реальних людей, — можуть бути застосовані для фабрикації доказів у кримінальних справах, шантажі, дезінформаційних та дискредитаційних кампаніях тощо [6, 16]. Традиційні методи цифрової криміналістики, розроблені для виявлення маніпуляцій із мультимедійним контентом, виявляються недостатньо ефективними проти сучасних генеративних моделей, які не залишають типових артефактів стиснення або геометричних спотворень [17]. Відповіддю на ці виклики стало формування двох комплементарних підходів до детекції синтетичних медіа: пасивної та активної автентифікації [1, 3], покликаних забезпечити надійну доказову базу для верифікації та дослідження ШІ-генерованого контенту.

Звертаючись до проблеми, варто зазначити, що сучасні технології генеративного ШІ базуються на двох основних архітектурних підходах: генеративно-змагальних мережах *(Generative Adversarial Networks, GANs)* та дифузійних моделях *(Diffusion Models,*

DMs) [1]. *GANs*, запропоновані Яном Гудфеллоу в 2014 році, містять дві нейронні мережі — генератор та дискримінатор, які змагаються між собою, що призводить до створення все більш реалістичного синтетичного контенту. Дифузійні моделі, які набули популярності останніми роками, застосовують ітеративний процес *денойзингу* (англ. *denoising*) для генерації зображень високої якості з випадкового шуму [1, 4]. Обидві технології знайшли широке застосування в різних сферах — від розваг та мистецтва до медицини та освіти. Проте їхній потенціал для зловживань створює серйозні етичні та безпекові ризики [1].

У процесі технологічної еволюції генеративних моделей постійно підвищується якість синтезу та знижується бар'єр для входу. Якщо раніше створення переконливих дипфейків потребувало значних технічних знань та обчислювальних ресурсів, то сьогодні доступні численні онлайн-сервіси та додатки, які дають змогу генерувати синтетичний медіаконтент високої якості за лічені хвилини [8, 16]. Така демократизація технології значно розширює коло потенційних зловмисників та масштаб загрози, створює безпрецедентні виклики для системи судової експертизи. Традиційні методи цифрової криміналістики, розроблені для виявлення класичних маніпуляцій (копіювання-вставки, ретушування, монтажу тощо), базуються на пошуку специфічних артефактів: невідповідностей у метаданих *EXIF*, аномалій стиснення *JPEG*, порушень шумової структури або геометричних спотворень [17]. Проте сучасні генеративні моделі створюють контент «з нуля», не залишаючи цих традиційних слідів маніпуляції [1].

Навіть більше, генеративні моделі можуть бути навмисно натреновані для імітації специфічних характеристик реальних камер, включно з шумовими патернами, артефактами стиснення та навіть метаданими [17]. Дослідження демонструють, що навіть досвідчені експерти далеко не завжди можуть надійно відрізнити високоякісні дипфейки або ШІ-контент від автентичного контенту під час візуального огляду [6, 16], що робить детекцію синтетичного медіаконтенту надзвичайно складним завданням, потребує розроблення принципово нових підходів та методів проведення експертного дослідження подібних матеріалів, підкреслює критичну необхідність розроблення автоматизованих інструментів детекції, заснованих на глибокому навчанні та аналізуванні тонких статистичних аномалій, які не сприймаються людським зором [1, 5]. Проблема детекції україномовних ШІ-генерованих текстів є ще більш складною: наявні автоматизовані інструменти демонструють значно нижчу ефективність під час аналізу текстів українською мовою порівняно з англійською внаслідок недостатнього представлення в навчальних вибірках детекторів, постійному ускладненні ШІ-генерованих текстівтощо, що призводить до значних труднощів у визначенні належності тексту з високим ступенем ймовірності [14].

Додатковим викликом є швидкість еволюції генеративних технологій. Нові архітектури моделей з'являються щомісяця, кожна з них може мати унікальні характеристики та артефакти генерації [1, 9]. Це пришвидшує постійну «гонку технологій» між розробниками генеративних моделей та дослідниками методів детекції в парадигмі *«AI versus AI»* [14], де детектори швидко застарівають і потребують постійного оновлення [6, 13].

Отже, сучасні підходи до детекції ШІ-генерованого контенту можна класифікувати за кількома критеріями. За принципом роботи розрізняють пасивні та активні методи [1, 3]. Пасивні методи аналізують внутрішні характеристики контенту без попереднього втручання в процес його створення. Вони містять аналіз статистичних аномалій, виявлення артефактів генерації, дослідження фізіологічних невідповідностей (наприклад, аномалій моргання або дихання у відео) та застосування глибоких нейронних мереж для класифікації [1, 6, 15]. Активні методи, навпаки, передбачають проактивне впровадження механізмів верифікації на етапі створення контенту [3, 7]. До них належать цифрові водяні знаки (видимі та невидимі, як приклад — *SynthID, Digimarc Validate, Meta Stable Signature, Steganographic AI (Steg.AI)* тощо), метадані провенансу (специфікації *C2PA*), криптографічні підписи та технології блокчейн для забезпечення

незмінності записів про походження контенту [7, 11, 12]. За модальністю контенту методи детекції поділяються на спеціалізовані для зображень, відео-, аудіо- та тексту [2, 10]. Кожна модальність має специфічні характеристики та потребує адаптованих підходів. Наприклад, детекція дипфейк-відео може застосовувати темпоральну інформацію та аналіз послідовності кадрів, що недоступно для статичних зображень [1, 16]. За рівнем інтеграції у процесі створення контенту розрізняють *post-hoc* детекцію (аналіз уже створеного контенту) та вбудовану верифікацію (інтеграція механізмів автентифікації безпосередньо в генеративні моделі або пристрої захоплення) [3, 7]. Останній підхід вважається більш перспективним для довгострокового розв'язання проблеми, оскільки він забезпечує автентифікацію «за замовчуванням» та ускладнює створення неідентифікованого синтетичного контенту [5]. Варто зазначити, що дана класифікація не є вичерпною, це авторське бачення, намагання узагальнити й систематизувати наявні дослідження та підходи до проблеми.

Рано чи пізно ШІ-генерований контент стане повсякденним об'єктом експертного дослідження. Саме тому класифікація ШІ-контенту, ба більше, дослідження технічних принципів його ідентифікації, практик маркування, розроблення та впровадження валідованих методик є критично необхідними, якщо експертна спільнота прагне вижити в умовах цифрової революції першої половини XXI століття.

Перелік джерел посилання

1. Amerini I., Barni M., Battiato S., etc. Deepfake Media Forensics: State of the Art and Challenges Ahead. *arXiv preprint*. 2024. DOI: 10.48550/arXiv.2408.00388 (дата звернення: 18.01.2026).
2. Cao L. Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities. *ICLR 2025 Workshop on GenAI Watermarking*. 2025. DOI: 10.48550/arXiv.2504.03765 (дата звернення: 18.01.2026).
3. Deng J., Lin C., Zhao Z., Liu S., Peng Z., Wang Q., Shen C. A Survey of Defenses against AI-generated Visual Media: Detection, Disruption, and Authentication. *ACM Computing Surveys*. 2025. Vol. 58, Iss. 5. Art. No 123. P*p*. 1—35. DOI: 10.48550/arXiv.2407.10575 (дата звернення: 18.01.2026).
4. Fernandez P., Couairon G., Jégou H., Douze M., Furon T. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 2023. P*p*. 22466—22477. DOI: 10.1109/ICCV51070.2023.02053 (дата звернення: 18.01.2026).
5. Hilbert E., Greene G., Godwin M., Shirazyan S. Watermarking and Metadata for GenAI Transparency at Scale: Lessons Learned and Challenges Ahead. *ICLR 2025 Workshop on GenAI Watermarking*. 2025. URL: https://openreview.net/pdf?id=JdNfGpVH6r (дата звернення: 21.01.2026).
6. Hydara E., Kikuchi M., Ozono T. Empirical assessment of deepfake detection: Advancing judicial evidence verification through artificial intelligence. *IEEE Access*. 2024. Vol. 12. P*p*. 151188—151203. DOI: 10.1109/ACCESS.2024.3480320 (дата звернення: 18.01.2026).
7. Jiang Z., Guo M., Hu Y., Gong N.Z. Watermark-based Detection and Attribution of AI-Generated Content. *arXiv preprint*. 2024. DOI: 10.48550/arxiv.2404.04254 (дата звернення: 18.01.2026).
8. Kumar M. AI in Digital Forensics: Detecting Deepfakes and Synthetic Media Attacks. *International Journal of Science, Engineering and Technology*. 2025. Vol.13. Iss. 2. URL: https://surl.li/spxzbk (дата звернення: 23.01.2026)
9. Kumar N., Sharma C. Deepfake Detection Methods and Trends: A Comprehensive Analysis. *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*. 2025. DOI: 10.1109/ISAC364032.2025.11156395 (дата звернення: 18.01.2026).
10. Kumarage T., Agrawal G., Sheth P., Moraffah R., Chadha A., Garland J., Liu H.A Survey of AI-generated Text Forensic Systems: Detection, Attribution, and Characterization. *arXiv preprint*. 2024. DOI: 10.13140/RG.2.2.23893.60645 (дата звернення: 18.01.2026).
11. Mastoi Q.-u.-A., Memon M. F., Jan S., Jamil A., Faique M., Ali Z., Lakhan A., Syed T.A. Enhancing Deepfake Content Detection Through Blockchain Technology. *International Journal of Advanced Computer Science and Applications*. 2025. Vol. 16. No. 6. P*p*. 48—56. DOI: 10.14569/IJACSA.2025.0160607 (дата звернення: 18.01.2026).

12. Nandwani S., Ostrowski D.A. Utilizing NFT Technology and Generative AI for Deep Fake Detection and Media Authentication. *2024 Conference on AI, Science, Engineering, and Technology (AIxSET).* 2024. DOI: 10.1109/AIxSET62544.2024.00059 (дата звернення: 18.01.2026).
13. Ölvecký M., Huraj L., Brlej I. Evaluating the Effectiveness of Deepfake Video Detection Tools: A Comparative Study. *TEM Journal*. 2025. Vol. 14. Iss. 1. P*p.* 64—77. DOI: 10.18421/TEM141-07 (дата звернення: 18.01.2026).
14. Prykhodchenko S.D., Prykhodchenko O.Yu., Shevtsova O.S., Olishevskiy I.H. Software detection of Ukrainian-language texts generated by AI: methods, estimations, challenges. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu.* 2025. No 4. P*p.* 150—159. DOI: 10.33271/nvngu/2025-4/150 (дата звернення: 18.01.2026).
15. Sohail S., Sajjad S.M., Zafar A., Iqbal Z., Muhammad Z., Kazim M. Deepfake Detection Using Deep Learning: A Unified Forensic Approach to Detect AI-Generated Images and Videos with Fusion of Eye, Nose, and Mouth Landmarks. *Preprints.* 2025. DOI: 10.20944/preprints202502.0552.v1 (дата звернення: 18.01.2026).
16. Soni N. Deepfake Detection and Multimedia Forensics: Investigating Synthetic Media, Image Forgery, and Video Manipulation in Cybercrime Cases. *ARC Journal of Forensic Science*. 2025. Vol. 9. Iss. 2. P*p.* 36—39. DOI: 10.20431/2456-0049.0902005 (дата звернення: 18.01.2026).
17. Verdoliva L. Media Forensics and DeepFakes: an overview. *IEEE Journal of Selected Topics in Signal Processing. 2020. Vol. 14, Iss. 5. P. 910-932.* DOI: 10.1109/JSTSP.2020.3002101