МЕТОДОЛОГІЯ ВИЯВЛЕННЯ СИНТЕЗОВАНОГО МЕДІАКОНТЕНТУ (DEEPFAKES) У СИСТЕМАХ КІБЕРЗАХИСТУ

Скрильник В.Ю., Кононенко І.В., Поддубний В.О.
Харківський національний університет радіоелектроніки, Харків, Україна

Стрімкий розвиток генеративного штучного інтелекту та глибинного навчання призвів до виникнення критичної загрози для цифрової безпеки – автоматизованого синтезу фальсифікованого аудіовізуального контенту (діпфейків) [1-3]. Створення надреалістичних підробок обличчя, міміки та голосу створює передумови для обходу дистанційної біометричної автентифікації, здійснення фінансового шахрайства та дезінформаційних операцій [1-3]. Документовані інциденти компрометації біометричних систем та мультимільйонних фінансових втрат у банківській сфері доводять, що синтетичний контент перетворився з медійного виклику на фундаментальну проблему кіберзахисту. Традиційні засоби пасивної цифрової криміналістики втрачають ефективність проти сучасних моделей на основі генеративно-змагальних мереж (GANs) та диффузійних архітектур (DDPMs), оскільки останні синтезують автентичний розподіл пікселів і частотно-спектральних характеристик без залишення явних слідів монтажу [1, 4].

Метою доповіді є розробка та обґрунтування комплексної методології виявлення синтезованого медіаконтенту на основі мультимодального аналізу та трансформерних нейромережевих архітектур для підвищення стійкості систем кіберзахисту проти наступного покоління генеративних атак.

Аналіз векторів загроз і обмежень існуючих засобів. Синтетичний медіаконтент підрозділяється на чотири ключові класи фальсифікації: заміни обличчя, зміну міміки, синхронізацію губ та анімацію статичного зображення у відеопотік. Найбільшу небезпеку для дистанційного Onboarding та KYC-процедур становлять підробки класу image-to-video, які створюють часово-узгоджене відео із єдиного фото без використання традиційного просторового стикування. Порівняльний бенчмаркінг класичних згорткових моделей (зокрема переможця DFDC Challenge – EfficientNet-B7) показав гостре падіння показника повноти виявлення до 24,0% на підробках типу image-to-video, оскільки згорткові шари перенавчаються на просторові шовні артефакти. Комерційні хмарні SaaS-платформи забезпечують високу специфічність (96,1%), проте випускають від 16% до 20% підробок типу image-to-video та створюють ризики витоку біометричних даних за межі корпоративного контуру [1]. Запропонована мультимодальна методологія детекції. Для подолання вказаних обмежень розроблено багаторівневу методологію детекції, що базується на інтеграції просторових, часових, акустичних та контекстних ознак в єдиний аналітичний пайплайн [2, 3]. Вхідний медіапотік $X = \{X_v, X_a, X_t\}$ піддається стандартизованій попередній обробці: для відео здійснюється виділення кадрів та кадрування обличчя (RetinaFace/MTCNN, розмірність 224x224); для аудіо – генерація спектрограм; для текстового супроводу – токенізація [2].

Базовим елементом методології є гібридна нейромережева архітектура, що поєднує локальне вилучення ознак за допомогою ConvNeXt та моделювання глобального контексту за допомогою Swin Transformer. Для моделювання міжмодальних зв'язків застосовано механізм спрямованого cross-attention ($Z_{v \leftarrow a}$} = CrossAttn(Z_v, Z_a)), який дозволяє моделі розпізнавати акустико-візуальну розсинхронізацію між фонемами та рухом губ [2–4]. Окремий модуль контекстної оцінки аналізує поведінкові індикатори (аномалії цифрового відбитка пристрою, IP/сесійні параметри, шаблон запиту), що дозволяє пріоритезувати перевірку під час транзакцій високого ризику [3].

Оцінка ефективності та результати експериментів. Інтеграція векторів ознак здійснюється у шарі каліброваної фузії ($Z_{fusion} = \Sigma \alpha_m Z_m$), який формує єдину оцінку ризику та показники невизначеності [2, 3]. Експериментальні дослідження запропонованої мультимодальної технології з cross-attention на репрезентативних датасетах (DFAVD, TTS-Deepfake) продемонстрували досягнення точності Accuracy = 0,950 (95,0%) та F1-score = 0,925 (92,5%), що суттєво перевищує результати одномодальних моделей (відео-CNN: Acc = 88,0%; аудіо-LSTM: Acc = 81,0%). Додавання механізму розпізнавання Cross-Attention забезпечує приріст F1-score на 4,5% за рахунок виявлення локальних міжмодальних неузгодженостей [3]. Застосування методології у реальних системах кіберзахисту дозволяє зменшити рівень хибно-позитивних спрацювань до 2,0% за умови залучення аналітика в спірних випадках. Практична імплементація даного підходу в операційні веб-контури та біометричні процеси запевнює зниження обсягів фінансового шахрайства на 45-60% у банках і фінтех-сервісах, зменшуючи середній час прийняття рішення до 2-4 секунд [3]. Методологія мультимодального аналізу із застосуванням Cross-Attention трансформерів є ефективним рішенням для локального захисту систем кібербезпеки від новітніх загрозових векторів deepfake. Вона забезпечує високу точність детекції, інтерпретованість рішень через карти уваги (Grad-CAM) та відповідність вимогам захисту персональних даних (GDPR).

Список літератури

1. Dzhaliuk N., Khoma V., Khoma Y. COMPARATIVE EVALUATION OF DEEPFAKE DETECTION MODELS FOR SECURE REMOTE BIOMETRIC VERIFICATION. Cybersecurity Education Science Technique. 2026. Vol. 2, no. 34. P. 597. URL: https://doi.org/10.28925/2663-4023.2026.34.1349.

2. Інформаційна технологія виявлення діпфейків на основі глибинного навчання та мультимодального аналізу для інтелектуальних систем інформаційної безпеки / V. Vysotska та ін. Системи управління, навігації та зв'язку. Збірник наукових праць. 2026. Т. 2, № 84. С. 52–61. URL: https://doi.org/10.26906/sunz.2026.2.052.

3. Amaresam R. K. Deepfake Detection and AI's Role in Preventing Digital Fraud. International Journal of Research and Applied Innovations. 2024. Vol. 7, no. 4. URL: https://doi.org/10.15662/ijrai.2024.0704011.

4. Dhope V. B., Murkute P. S., Nazare P. A. Deepfake Attacks and Detection Techniques in Cyber Security. International Journal of Research Publication and Reviews. 2026. Vol. 7, no. 5. P. 7557–7562. URL: https://doi.org/10.55248/gengpi.07.0526.c1309.