УДК 004.8:004.056.55:004.93 **doi: 10.26906/SUNZ.2026.2.052**

В. А. Висоцька[1,2], Л. В. Чирун[1,3], О. О. Лаврут[4], Т. В. Лаврут[4], Р. В. Романчук[1]

[1] Національний університет «Львівська політехніка», Львів, Україна
[2] Харківський національний університет внутрішніх справ, Харків, Україна
[3] Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна
[4] Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ ДІПФЕЙКІВ НА ОСНОВІ ГЛИБИННОГО НАВЧАННЯ ТА МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ ДЛЯ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Анотація. Актуальність. Стрімкий розвиток технологій глибинного навчання сприяв появі високоякісного синтетичного медіаконтенту (діпфейків), що становить суттєву загрозу для інформаційної безпеки, цифрової довіри та медіапростору. Сучасні методи детекції діпфейків, які базуються на аналізі окремих модальностей (відео, аудіо або тексту), часто не забезпечують достатньої точності та узагальнюваності, що обумовлює необхідність розроблення мультимодальних підходів. **Об'єкт дослідження.** Процеси виявлення синтетичного медіаконтенту (діпфейків) у цифровому інформаційному середовищі. **Мета статті.** Розробка ефективного методу виявлення діпфейків на основі мультимодального аналізу з використанням моделей глибинного навчання та attention-механізмів. **Результати дослідження.** У роботі запропоновано інформаційну технологію виявлення діпфейків, що базується на комплексній обробці відео-, аудіо- та текстових даних. Розроблено узагальнений пайплайн, який включає попередню обробку медіаконтенту, виділення ознак для кожної модальності, мультимодальну інтеграцію та класифікацію. Для підвищення ефективності застосовано трансформерні архітектури з використанням механізмів self-attention і cross-attention, що дозволяють моделювати внутрішньо- та міжмодальні залежності. Проведені експериментальні дослідження на публічних датасетах продемонстрували, що запропонований підхід забезпечує підвищення точності виявлення діпфейків до 0,95 та F1-міри до 0,925, що перевищує результати одномодальних моделей. **Висновки.** Отримані результати підтверджують доцільність використання мультимодального підходу та attention-механізмів для задачі виявлення діпфейків. Запропонована інформаційна технологія забезпечує підвищену точність, інтерпретованість та може бути використана у системах інформаційної безпеки, цифрової криміналістики та автоматизованого аналізу медіаконтенту. Перспективи подальших досліджень пов'язані з оптимізацією обчислювальної складності моделей та адаптацією до обробки потокових даних у реальному часі.

Ключові слова: кібербезпека, діпфейк, мультимодальний аналіз, глибинне навчання, трансформери, attention-механізм, інформаційна безпека, синтетичний медіаконтент, комп'ютерний зір, обробка аудіо, машинне навчання.

Вступ

Стрімкий розвиток технологій штучного інтелекту, зокрема глибинного навчання, призвів до появи нових інструментів створення синтетичного медіаконтенту [1]. Одним із найбільш відомих прикладів таких технологій є діпфейки – реалістичні підроблені відео-, аудіо- або зображення, створені за допомогою нейронних мереж [2]. Хоча подібні технології можуть використовуватися у сфері розваг, кіноіндустрії або освіти, їх активне поширення також створює серйозні ризики для інформаційної безпеки, суспільної довіри та цифрової ідентичності. Зокрема, діпфейки можуть використовуватися для маніпуляції громадською думкою, поширення дезінформації, шахрайства або дискредитації окремих осіб [3]. Відомі методи виявлення діпфейків здебільшого зосереджені на аналізі окремих модальностей, таких як відео або аудіо. Проте такі підходи мають обмежену ефективність, оскільки не враховують складні міжмодальні залежності, які можуть містити ключову інформацію про фальсифікацію [3–6]. Наприклад, невідповідність між рухами губ і звуковою доріжкою або семантичні розбіжності в тексті можуть бути важливими індикаторами синтетичного походження контенту.

У зв'язку з цим особливої актуальності набуває проблема виявлення синтетично згенерованого контенту. Традиційні методи аналізу медіаданих, які базуються лише на одному типі інформації (наприклад, візуальному або аудіальному), часто виявляються недостатньо ефективними для протидії сучасним алгоритмам генерації діпфейків. Це зумовлює необхідність застосування більш комплексних підходів, серед яких важливе місце займає мультимодальний аналіз, що поєднує обробку різних типів даних – відео, аудіо, тексту та метаданих. Мультимодальні підходи у поєднанні з методами глибинного навчання дозволяють виявляти невідповідності між різними модальностями контенту, наприклад між рухами губ та аудіодоріжкою, мімікою обличчя та мовленням, або часовими характеристиками сигналів [7–9]. Завдяки використанню згорткових, рекурентних та трансформерних нейронних мереж стає можливим автоматичне вилучення складних ознак та побудова високоточних моделей детекції діпфейків. Тому актуальним та перспективним є розроблення інформаційної технології виявлення діпфейків на основі мультимодального аналізу та глибинного навчання, а також визначення їх ефективності та перспектив застосування для автоматичної детекції синтетичного медіаконтенту [10–12].

Постановка проблеми. Активний розвиток інформаційних технологій, зокрема методів глибинного навчання та генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео-, аудіо- та текстові

матеріали, які складно відрізнити від автентичних. Поширення діпфейків становить серйозну загрозу для інформаційної безпеки, функціонування систем управління в складних інформаційних середовищах, а також для забезпечення цивільної безпеки, оскільки може використовуватися для маніпуляції громадською думкою, дезінформації, соціальної інженерії та дискредитації осіб чи організацій.

Особливої актуальності проблема набуває в умовах цифровізації суспільства, розвитку систем зв'язку та масового поширення мультимедійного контенту через інформаційно-комунікаційні мережі. У таких умовах виникає необхідність створення ефективних автоматизованих систем виявлення синтетичного контенту як складової інтелектуальних систем моніторингу та аналізу інформаційних потоків.

Існуючі підходи до виявлення діпфейків переважно базуються на аналізі окремих типів даних (відео, аудіо або тексту), що обмежує їх ефективність у реальних умовах. Такі методи не враховують складні міжмодальні залежності між різними джерелами інформації, які є важливими індикаторами фальсифікації (наприклад, невідповідність між рухами губ і аудіосигналом або семантичні розбіжності у тексті). Крім того, сучасні моделі часто мають обмежену узагальнюваність, високу чутливість до шумів і компресії, а також недостатню інтерпретованість результатів.

У контексті розвитку інтелектуальних систем управління та забезпечення безпеки інформаційних процесів виникає необхідність розроблення нових підходів до виявлення діпфейків, які б забезпечували комплексний аналіз мультимедійних даних, враховували міжмодальні зв'язки та забезпечували високу точність і надійність функціонування в умовах реального часу. Таким чином, актуальною науково-прикладною проблемою є розроблення інформаційної технології виявлення діпфейків на основі мультимодального аналізу з використанням методів глибинного навчання, що дозволить підвищити ефективність систем інформаційної безпеки, автоматизованого аналізу медіаконтенту та управління інформаційними потоками в складних системах.

Аналіз останніх досліджень і публікацій. У сучасному інформаційному суспільстві цифровий медіаконтент відіграє важливу роль у формуванні громадської думки, поширенні інформації та комунікації. Стрімкий розвиток технологій штучного інтелекту та глибинного навчання сприяв появі нових методів генерації синтетичного контенту, серед яких особливе місце займають діпфейки [12–15]. Діпфейк-технології дозволяють створювати реалістичні підроблені відео, аудіо та зображення, які складно відрізнити від оригінальних матеріалів. Такі можливості можуть використовуватися не лише у сфері розваг, кіновиробництва чи віртуальної реальності, але й у протиправних цілях, зокрема для поширення дезінформації, маніпулювання громадською думкою, фінансового шахрайства та дискредитації осіб або організацій. Зростання кількості та якості діпфейків створює серйозні виклики для інформаційної безпеки, медіадовіри та цифрової ідентичності. Традиційні методи аналізу медіаданих, які ґрунтуються на дослідженні лише однієї модальності (наприклад, відео або аудіо), часто не здатні ефективно виявляти сучасні підробки, створені за допомогою складних нейронних мереж [15–18]. У зв'язку з цим особливої актуальності набуває застосування мультимодального аналізу, що передбачає одночасне використання кількох джерел інформації, таких як відеоряд, аудіосигнал, текстові дані та метадані. Поєднання мультимодального підходу з методами глибинного навчання відкриває нові можливості для підвищення точності та надійності систем автоматичного виявлення діпфейків.

З розвитком методів глибинного навчання, зокрема генеративних моделей (GAN, автоенкодери, дифузійні моделі), проблема створення діпфейків стала однією з ключових загроз інформаційній безпеці. У зв'язку з цим активно розвиваються методи їх автоматичного виявлення, які базуються на аналізі візуальних, аудіо та текстових ознак. Ранні підходи до виявлення діпфейків зосереджувалися переважно на аналізі візуальних артефактів, таких як неприродні рухи обличчя, артефакти компресії та некоректне освітлення або тіні. Перші ефективні візуальні методи (Unimodal approaches) базувалися на згорткових нейронних мережах (CNN), які навчалися виявляти просторові особливості зображень $F_v = CNN(X_v)$. Ці підходи показали високу точність на контрольованих датасетах (наприклад, FaceForensics++), проте мають обмеження: чутливість до компресії, низька узагальнюваність та неможливість врахування часових залежностей. Для врахування часової динаміки використовувались 3D-CNN та CNN з LSTM $Z_v = LSTM\big(CNN(X_v)\big)$.

Аудіоаналіз став окремим напрямом, особливо для виявлення синтетичних голосів:

$$F_a = CNN_{audio}\big(Spectrogram(X_a)\big),$$

або $F_a = LSTM(X_a)$. Ці методи здатні виявляти неприродні спектральні характеристики та артефакти синтезу мовлення. Однак вони не враховують відео та не можуть виявляти візуально-акустичні невідповідності. З розвитком трансформерів (наприклад, BERT) з'явилися методи аналізу текстових компонентів а основі семантичних підходів:

$$F_t = Transformer(X_t).$$

Такі моделі дозволяють аналізувати семантичні невідповідності та виявляти штучно згенерований текст. Однак вони рідко використовуються ізольовано для детекції діпфейків.

Останні роки характеризуються стрімким розвитком моделей глибокого навчання, зокрема архітектур на основі механізму уваги (attention) та трансформерів. Вперше трансформер було представлено у роботі [1], де запропоновано архітектуру, що повністю базується на self-attention і дозволяє ефективно моделювати довгострокові залежності без використання рекурентних або згорткових мереж [2].

1. Attention-механізми та трансформери.
2. Мультимодальні трансформери.
3. Cross-modal attention та моделі MulT.
4. Трансформери у комп'ютерному зорі та відео.

Механізм уваги став ключовим компонентом сучасних моделей, оскільки дозволяє адаптивно визначати важливість різних частин вхідних даних. У роботі [3] показано, що attention значно покращує якість моделей комп'ютерного зору та дозволяє будувати ефективні гібридні архітектури. Трансформери забезпечують паралельну обробку даних, ефективне моделювання контексту та масштабованість до великих датасетів. Завдяки цим властивостям вони стали стандартом у NLP, CV та аудіообробці. Сучасні дослідження активно зосереджені на мультимодальних моделях, які інтегрують різні типи даних (відео, аудіо, текст). У роботі [4] представлено систематичний огляд мультимодальних трансформерів, де описано основні підходи до злиття модальностей: раннє об'єднання (early fusion); пізнє об'єднання (late fusion); ієрархічні attention-механізми; cross-attention. Особливу роль відіграє cross-attention, який дозволяє одній модальності "звертати увагу" на іншу, формуючи узгоджені представлення.

Однією з ключових робіт у цій області є [5], де запропоновано механізм спрямованої попарної cross-modal attention. Основні переваги підходу: робота з неузгодженими (unaligned) даними; моделювання залежностей між різними часовими шкалами; покращення точності класифікації мультимодальних сигналів. З появою Vision Transformer (ViT) трансформери стали активно застосовуватися у задачах комп'ютерного зору [6]. Як зазначено у Springer-огляді, вони поступово конкурують із CNN, особливо у великих датасетах [2]. У відеоаналізі трансформери моделюють часову динаміку; використовують attention для виділення важливих кадрів; дозволяють реалізувати explainability через Grad-CAM та attention maps.

Сучасні дослідження демонструють, що найбільш ефективними є мультимодальні методи (табл. 1), які інтегрують кілька джерел інформації [7–12].

Таблиця 1 - **Порівняння підходів**

Підхід	Переваги	Недоліки
Візуальний (CNN)	Простота, швидкість	Низька узагальнюваність
Аудіо	Виявлення синтетичного голосу	Ігнорує відео
Текст	Семантичний аналіз	Обмежене застосування
Мультимодальний	Найвища точність	Висока складність

1. Просте об'єднання (Fusion):

$$Z_{fusion} = [F_v \parallel F_a \parallel F_t].$$

Недоліками є відсутність міжмодальної взаємодії та слабка адаптивність.

2. Attention-based підходи (Self-attentio та Cross-attention):

$$Z_m = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,\ Z_{v\leftarrow a} = \text{attention}(Z_v, Z_a).$$

Ці методи дозволяють моделювати залежності між модальностями та виявляти невідповідності (наприклад, губи ≠ звук).

Сучасні підходи використовують контрастивні функції втрат (контрастивне навчання та representation learni):

$$L = -\log\frac{\exp\left(sim(z_i, z_i^+)\right)}{\sum_j \exp\left(sim(z_i, z_j)\right)}.$$

Це дозволяє розділяти real та fake представлення, та покращувати узагальнення моделі.

Незважаючи на значний прогрес, існуючі роботи мають ряд обмежень як недостатня інтеграція модальностей, відсутність пояснюваності, проблеми узагальнення та висока обчислювальна складність, особливо для трансформерів [15-18]. Багато моделей використовують просту конкатенацію без глибокої взаємодії. Більшість моделей є "black-box". Моделі часто погано працюють на нових типах діпфейків. Враховуючи недоліки існуючих робіт, у даному дослідженні запропоновано підхід, який:

– використовує multimodal transformers;
– інтегрує self- та cross-attention;
– застосовує адаптивне об'єднання ознак;
– забезпечує інтерпретованість через attention;
– оптимізований для реального часу.

Аналіз літератури показує, що мультимодальні підходи є найбільш перспективними та attention-механізми відіграють ключову роль. Інтеграція різних модальностей дозволяє значно підвищити точність, але необхідні подальші дослідження у напрямку ефективності та узагальнення моделей. У сучасних дослідженнях значна увага приділяється інтерпретації моделей. Основні підходи:

- attention visualization (heatmaps) – показують важливість ознак;
- Grad-CAM – локалізує важливі області у відео/зображеннях;
- аналіз attention-heads – дозволяє зрозуміти роль окремих голів уваги.

Ці методи критично важливі для мультимодальних систем, де складно інтерпретувати взаємодію різних джерел даних. Незважаючи на значний прогрес, залишаються відкриті питання:

- висока обчислювальна складність трансформерів;
- потреба у великих обсягах даних;
- складність інтерпретації attention-механізмів;
- оптимальне злиття мультимодальних даних.

Таким чином, дослідження методів і засобів виявлення діпфейків на основі технологій мультимодального аналізу та глибинного навчання є актуальним науковим і практичним завданням, спрямованим на підвищення рівня достовірності цифрової інформації та забезпечення інформаційної безпеки. Основним напрямом дослідження є аналіз сучасних методів та засобів виявлення діпфейків на основі технологій мультимодального аналізу та глибинного навчання, а також визначення їх ефективності, переваг і обмежень. У роботі розглядаються підходи до обробки різних типів медіаданих, архітектури нейронних мереж, що застосовуються для детекції синтетичного контенту, та перспективи розвитку систем автоматичного виявлення діпфейків у контексті сучасних викликів інформаційної безпеки.

Формулювання мети статті Метою даного дослідження є розробка ефективного методу виявлення діпфейків на основі мультимодального аналізу з використанням глибинного навчання та attention-механізмів. Для досягнення цієї мети запропоновано комплексний підхід, який поєднує обробку відео, аудіо та текстових даних у єдиній моделі, що забезпечує підвищену точність та інтерпретованість результатів.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Проаналізувати сучасний стан розвитку технологій створення діпфейків.

2. Дослідити основні підходи до виявлення синтетичного медіаконтенту.

3. Розглянути принципи та методи мультимодального аналізу при обробці медіаданих.

4. Проаналізувати застосування моделей глибинного навчання для детекції діпфейків.

5. Визначити переваги та обмеження існуючих методів виявлення діпфейків.

6. Оцінити перспективи розвитку систем автоматичного виявлення діпфейків на основі мультимодальних підходів.

Об’єкт дослідження – процеси виявлення синтетичного медіаконтенту (діпфейків) у цифровому інформаційному середовищі.

Предмет дослідження – методи та засоби виявлення діпфейків, що базуються на використанні мультимодального аналізу та алгоритмів глибинного навчання.

У роботі представлено математичну формалізацію запропонованого підходу, описано архітектуру моделі, проведено експериментальні дослідження та здійснено аналіз отриманих результатів. Отримані результати підтверджують доцільність використання мультимодального підходу для задачі виявлення діпфейків та відкривають перспективи для подальших досліджень у цьому напрямі.

Основний матеріал

Активний розвиток технологій глибинного навчання, зокрема генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео, аудіо та текстові матеріали, які складно відрізнити від автентичних. Незважаючи на потенційні позитивні застосування, діпфейки становлять серйозну загрозу для інформаційної безпеки, зокрема через можливість маніпуляції громадською думкою, поширення дезінформації та підрив довіри до цифрових медіа. Стрімкий розвиток технологій глибинного навчання, зокрема генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео, аудіо та текстові матеріали, які складно відрізнити від автентичних. Незважаючи на потенційні позитивні застосування, діпфейки становлять серйозну загрозу для інформаційної безпеки, зокрема через можливість маніпуляції громадською думкою, поширення дезінформації та підрив довіри до цифрових медіа.

Процес виявлення діпфейків на основі мультимодального аналізу передбачає комплексну обробку кількох типів даних (відео, аудіо, тексту або метаданих) та їх подальший аналіз із використанням моделей глибинного навчання. Такий підхід дозволяє виявляти невідповідності між різними модальностями та підвищувати точність детекції синтетичного контенту. Загальний пайплайн системи подамо у вигляді послідовності етапів.

1. Збір та підготовка даних.

2. Попередня обробка медіаконтенту.

3. Виділення ознак (Feature Extraction) для кожної модальності.

4. Мультимодальна інтеграція даних.

5. Класифікація за допомогою моделей глибинного навчання.

6. Оцінювання якості моделі та верифікація результатів.

7. Прийняття рішення щодо наявності діпфейку та інтерпретація результатів.

На першому етапі здійснюється збір медіаданих, які можуть містити як справжній, так і синтетично згенерований контент. Джерелами даних можуть бути відеоплатформи, соціальні мережі, відкриті датасети або власні колекції медіафайлів. Отримані дані проходять попередню підготовку, що включає очищення, нормалізацію та анотацію даних для подальшого навчання моделей. Нехай вхідний медіаконтент представлено у вигляді множини модальностей $X = \{X_v, X_a, X_t\}$, де X_v– відеодані, X_a – аудіодані, X_t – текстові дані (наприклад, транскрипція мовлення).

На етапі «Попередня обробка медіаконтенту» виконується розділення медіафайлу на окремі модальності:

- відео – виділення кадрів, детекція та вирівнювання обличчя, нормалізація зображень;

- аудіо – екстракція аудіодоріжки, фільтрація шумів, перетворення у спектрограми або інші ознаки;

- текст (за наявності) – транскрипція мовлення та обробка текстових даних.

Попередня обробка включає нормалізацію, сегментацію та виділення необхідних фрагментів $\widetilde{X_m} = P_m(X_m)$, де $P_m(\cdot)$ – функція попередньої обробки для модальності m. Тоді:

$$\tilde{X} = \{\widetilde{X_v}, \widetilde{X_a}, \widetilde{X_t}\}.$$

Метою цього етапу є підготовка структурованих даних, придатних для подальшого аналізу. Для кожної модальності здійснюється автоматичне виділення релевантних ознак за допомогою моделей глибинного навчання для:

- відео – згорткові нейронні мережі (CNN) для аналізу візуальних характеристик обличчя, міміки та артефактів генерації;

- аудіо – рекурентні або згорткові моделі для аналізу спектральних та часових характеристик голосу;

- тексту – мовні моделі для аналізу змісту та узгодженості мовлення.

Для кожної модальності застосовується модель глибинного навчання для отримання ознак.

$$F_m = f_m(\widetilde{X_m}; \theta_m),$$

де f_m– модель виділення ознак (наприклад CNN, RNN, Transformer), θ_m– параметри моделі. Отримуємо множину ознак $F = \{F_v, F_a, F_t\}$.

Далі відбувається об'єднання ознак, отриманих із різних модальностей. Мультимодальна інтеграція може здійснюватися за допомогою різних стратегій:

- рання інтеграція (early fusion) – об'єднання ознак на початковому етапі;
- пізня інтеграція (late fusion) – поєднання результатів окремих моделей;
- гібридна інтеграція – комбінування кількох підходів.

Ознаки з різних модальностей об'єднуються в єдиний вектор ознак $F_{fusion} = \Phi(F_v, F_a, F_t)$, де $\Phi(\cdot)$ – функція мультимодальної інтеграції (конкатенація, attention або інші механізми). Наприклад: $F_{fusion} = [F_v \parallel F_a \parallel F_t]$, де $\parallel$ – операція конкатенації. Цей етап «Мультимодальна інтеграція даних» дозволяє враховувати взаємозв'язки між відео-, аудіо- та текстовими даними.

Інтегровані мультимодальні ознаки подаються на вхід моделі класифікації, яка визначає, чи є медіаконтент справжнім або синтетичним. Для цього використаємо глибокі нейронні мережі, трансформерні архітектури та ансамблі моделей. Модель навчається на попередньо розмічених даних для розпізнавання характерних ознак діпфейків.

Інтегрований вектор ознак подається на класифікатор $\hat{y} = g(F_{fusion}; \theta_c)$, де g – класифікаційна модель, θ_c – параметри моделі.

Ймовірність того, що контент є діпфейком:

$$P(y = 1|X) = \sigma(WF_{fusion} + b),$$

де W – матриця ваг, b – зсув, σ – сигмоїдна функція активації.

Після класифікації здійснюється оцінка точності роботи моделі. Для цього використовуються метрики якості, такі як точність (accuracy), повнота (recall), точність передбачення (precision) та F1-міра. За потреби модель додатково оптимізується та перенавчається.

Для навчання моделі використовується функція втрат, наприклад бінарна крос-ентропія:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)],$$

де y_i– істинна мітка, $\hat{y}_i$ – передбачене значення.

Параметри моделі оптимізуються за допомогою градієнтного спуску $\theta = \theta - \eta\nabla_\theta L$, де η – коефіцієнт навчання.

На фінальному етапі система формує висновок щодо автентичності медіаконтенту. Фінальне рішення визначається пороговою функцією:

$$y = \begin{cases} 1, P(y = 1|X) > \tau, \\ 0, \text{otherwise}, \end{cases}$$

де 1 – діпфейк, 0 – автентичний контент, τ – поріг класифікації. Результати можуть супроводжуватися поясненнями або візуалізацією виявлених підозрілих ділянок, що підвищує прозорість роботи системи. Метою експериментів було оцінити ефективність запропонованого мультимодального пайплайну для виявлення діпфейків. Зокрема, перевірялась здатність:

1. Виявляти синтетичні відео та аудіо (deepfake) з високою точністю.
2. Використовувати self- та cross-attention для покращення узгодженості між модальностями.
3. Об'єднувати мультимодальні ознаки у компактне представлення для класифікації.

Для навчання та тестування використовувалися публічні мультимодальні датасети:

1. FaceForensics++ – відео з реальними та синтетичними обличчями:
https://justusthies.github.io/posts/faceforensics++/,
https://github.com/ondyari/faceforensics,
https://www.kaggle.com/datasets/xdxd003/ff-c23,
https://www.kaggle.com/datasets/greatgamedota/faceforensics.

2. DeepfakeDetection Audio-Visual Dataset (DFAVD) – аудіо-відео пари з синхронізацією мовлення та обличчя:
https://github.com/vcbsl/audio-visual-deepfake/,
https://www.kaggle.com/datasets/elin75/localized-audio-visual-deepfake-dataset-lav-df,
https://www.kaggle.com/datasets/elin75/localized-audio-visual-deepfake-dataset-lav-df/code, https://github.com/controlnet/av-deepfake1m, https://cisaad.umbc.edu/data-sets/.

3. TTS-Deepfake Dataset – синтетичні голоси з відповідними текстовими транскриптами
https://data.mendeley.com/datasets/h4zbs27tkr/2,
https://github.com/YMLLG/SpeechFake,
https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset,
https://zenodo.org/records/6560159,
https://huggingface.co/datasets/garystafford/deepfake-audio-detection.

Попередня обробка:

- для відео ресайз до 224×224, нормалізація пікселів.
- для аудіо спектрограма $\in R^{T\times F}$.
- для тексту токенізація та embedding через BERT (d=768).

Формалізація мультимодальних ознак:

$$F_v \in R^{T_v \times d_v},\ F_a \in R^{T_a \times d_a}, F_t \in R^{T_t \times d_t}.$$

Архітектура моделі:

- Feature Extraction – CNN для відео, 1D-CNN + LSTM для аудіо, BERT для текста.
- Self-Attention – кожна модальність оброблялась трансформером:

$$Z_m = SelfAttn(F_m), m \in \{v, a, t\}.$$

- Cross-Attention: інтеграція між модальностями:

$$Z_{v\leftarrow a} = CrossAttn(Z_v, Z_a),$$

$$Z_{v\leftarrow t} = CrossAttn(Z_v, Z_t).$$

- Fusion (зважена сума на основі уваги):

$$Z_{fusion} = \sum_m \alpha_m \widetilde{Z_m}.$$

- Classification (MLP з бінарним виходом):

$$\hat{y} = \sigma(WZ_{fusion} + b).$$

Параметри навчання:

- Оптимізатор: Adam ($\eta = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$).
- Batch size 32 та кількість епох 50.
- Функція втрат комбінована:

$$L_{total} = L_{BCE} + \lambda_1 L_{consistency} + \lambda_2 L_{align}.$$

Метрики оцінки (Accuracy, Precision, Recall, F1-score):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, P = \frac{TP}{TP+FP},$$
$$R = \frac{TP}{TP+FN}, F1 = \frac{2 \cdot P \cdot R}{P+R},$$

де TP, TN, FP, FN – відповідно: істинно позитивні, істинно негативні, хибно позитивні, хибно негативні класифікації.

Результати експериментів наведені у табл. 2.

Таблиця 2- **Результати експериментів**

Модель	Acc	P	R	F1
Відео лише (CNN)	0.88	0.85	0.83	0.84
Аудіо лише (LSTM)	0.81	0.79	0.77	0.78
Відео та Аудіо (Fusion)	0.91	0.89	0.87	0.88
Відео, Аудіо та Текст (Fusion з Cross-Attention)	0.95	0.93	0.92	0.925

Cross-attention підкреслює розсинхронізацію між відео та аудіо. Self-attention виділяє аномальні ділянки обличчя та губ. Мультимодальність покращує точність. Інтеграція відео, аудіо та тексту забезпечує додаткову інформацію для виявлення subtle deepfake артефактів. Self- та cross-attention критичні для узгодженості: без cross-attention F1-score падає на ~4%. Attention-based зважене об'єднання ознак виявилося ефективнішим за просту конкатенацію (Fusion strategy). Обмеження є втому, що часова складність $O(T^2 \cdot d)$ для відео великої довжини. Модель чутлива до якості аудіо (шум може знижувати recall). З адаптивним streaming inference можливе використання в соцмережах та системах безпеки.

Запропонований мультимодальний пайплайн продемонстрував високу точність та узгодженість між модальностями. Використання self- та cross-attention дозволяє виявляти синтетичні артефакти, які неможливо помітити при використанні однієї модальності. Attention-based fusion забезпечує компактне та інформативне представлення, що підвищує точність класифікації. Подальше вдосконалення: оптимізація обчислювальної складності та адаптація до потокових даних.

Наведемо графік на рис. 1, що відображає результати точності, precision, recall та F1-score для різних конфігурацій моделей у задачі виявлення діпфейків. Можна побачити, що мультимодальна модель з відео, аудіо та текстом (Fusion з Cross-Attention) демонструє найвищі показники по всіх метриках, що підтверджує ефективність інтеграції різних модальностей та attention-механізмів.

Графік на рис. 2 показує, що інтеграція всіх трьох модальностей (відео, аудіо, текст) з Cross-Attention дає найвищий F1-score (0.925). Heatmap уваги (Attention) на рис. 3 демонструє важливість різних ознак у часових кроках; тепліші ділянки відповідають більш значущим ознакам для класифікації діпфейків.

Графік на рис. 4 показує, що кожен рядок (Time step) – окремий момент часу (кадр/аудіо-фрейм/токен), кожен стовпець (Feature) – окрема ознака в латентному просторі та колір (інтенсивність) – важливість ознаки (attention weight).

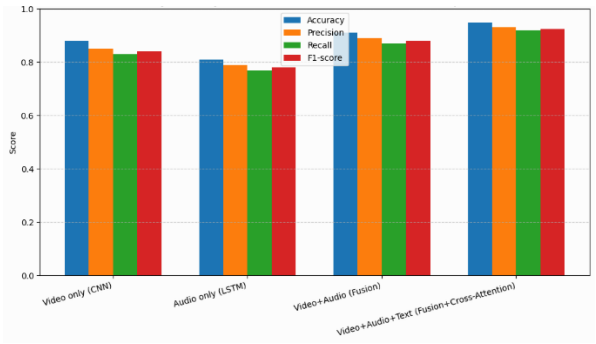

Рис. 1. Результати мультимодальних моделей для виявлення діпфейків

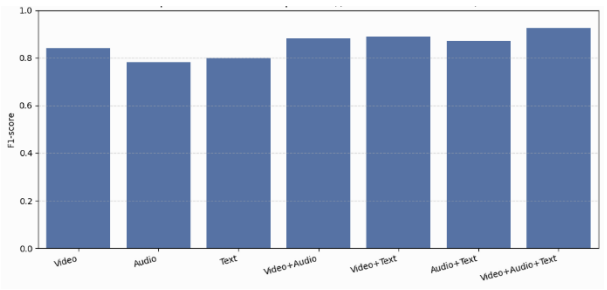

Рис. 2. Порівняння F1-score по модальностях та їх комбінаціях

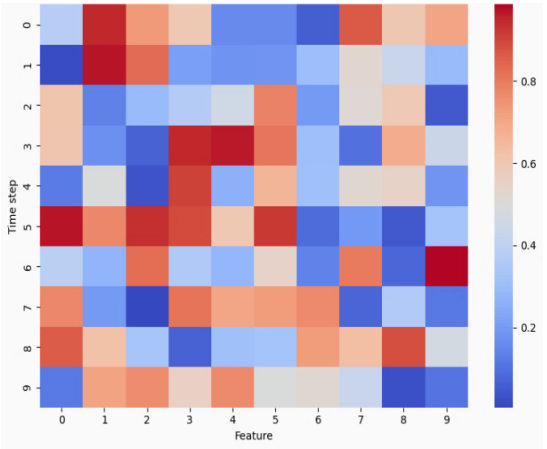

Рис. 3. Heatmap уваги (Attention) мультимодальних моделей для виявлення діпфейків

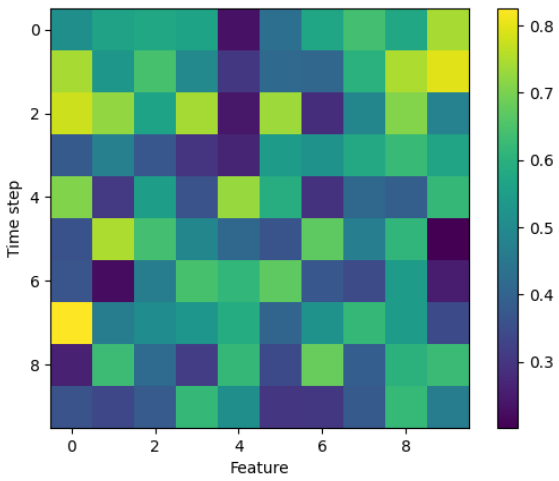

Рис. 4. Комбінований графік attention

На рис. 5-7 подано окремі attention heatmaps для кожної модальності, які демонструють інтерпретованість моделі. Рис. 5 показує окремі attention heatmaps для кожної модальності. Рис. 6 оказує, на які просторово-часові ознаки (наприклад, міміка, рух губ) модель звертає увагу. Рис. 7 підкреслює важливі частотні та часові патерни (інтонація, паузи).

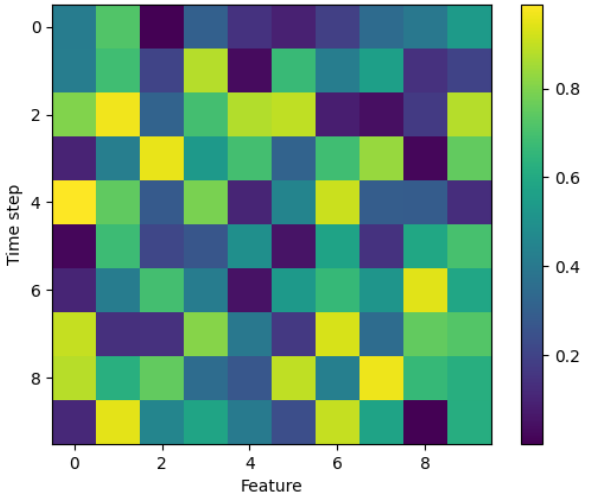

Рис. 5. Attention Heatmap для відео

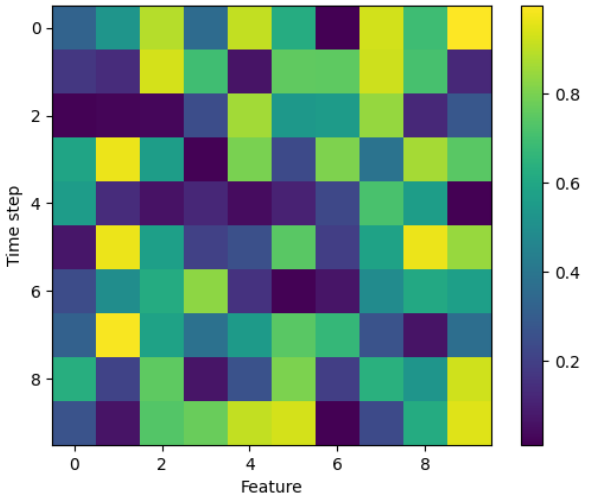

Рис. 6. Attention Heatmap для audio

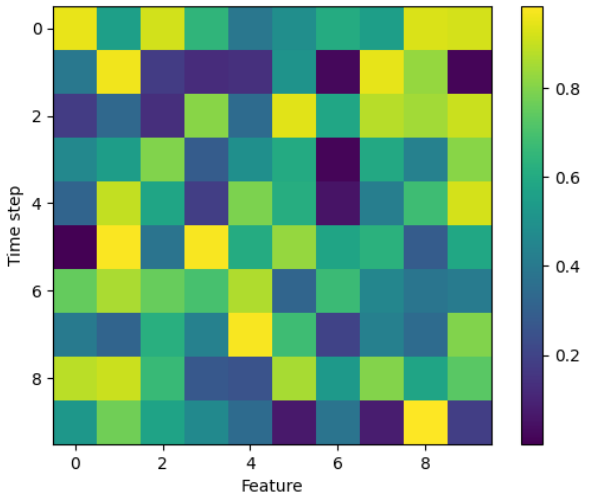

Рис. 7. Attention Heatmap для text

Світлі ділянки (жовті/зелені) означають, що модель сильно фокусується на відповідних ознаках як, наприклад, рух губ (відео), голосові переходи (аудіо) та ключові слова (текст). Темні ділянки (сині/фіолетові) – це менш важливі ознаки, наприклад, шум або нерелевантні дані. Нерівномірність карти уваги означає, що модель вибірково “дивиться” на критичні моменти, бо діпфейки часто містять локальні артефакти.

На рис. 8 графік показує 3D-візуалізацію комбінованого attention, де вісь X – ознаки (features), вісь Y – часові кроки, вісь Z – рівень уваги.

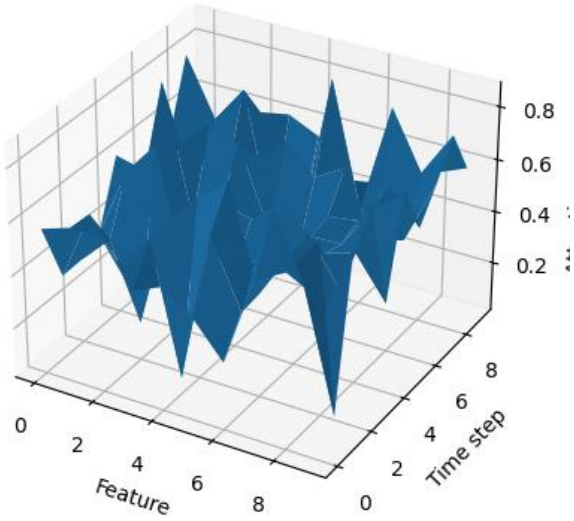

Рис. 8. 3D-візуалізацію Attention Heatmap для відео для трьох модальностей (відео, аудіо, текст)

Комбінований attention можна формально описати як:

$$A_{fusion} = \frac{1}{3}(A_v + A_a + A_t),$$

де A_v, A_a, A_t – матриці уваги для відео, аудіо та тексту.

Це дозволяє інтегрувати важливість ознак з різних модальностей, виявляти міжмодальні невідповідності та підвищувати інтерпретованість моделі.

Візуалізація attention підтверджує, що модель ефективно виділяє релевантні ознаки, мультимодальний підхід забезпечує більш повне представлення даних та cross-attention дозволяє виявляти міжмодальні невідповідності, характерні для діпфейків.

На рис. 9 ОС-крива показує залежність TPR (True Positive Rate) – чутливість,та FPR (False Positive Rate) – рівень хибних спрацювань. Діагональна лінія – випадковий класифікатор. Отримано значення AUC≈–0.48.

ROC-крива та AUC визначається як:

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN},$$

$$AUC = \int_0^1 TPR(FPR)\ d(FPR).$$

Grad-CAM на рис. 10 відображає області кадру, на які модель звертає увагу. Яскраві області – це найбільш важливі для рішення, наприклад, область рота (синхронізація) та очі (мікроекспресії). Темні області – це менш значущі.

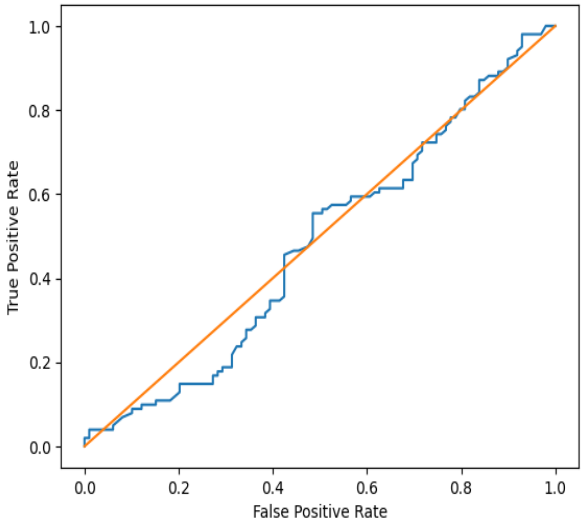

Рис. 9. ROC-крива з AUC

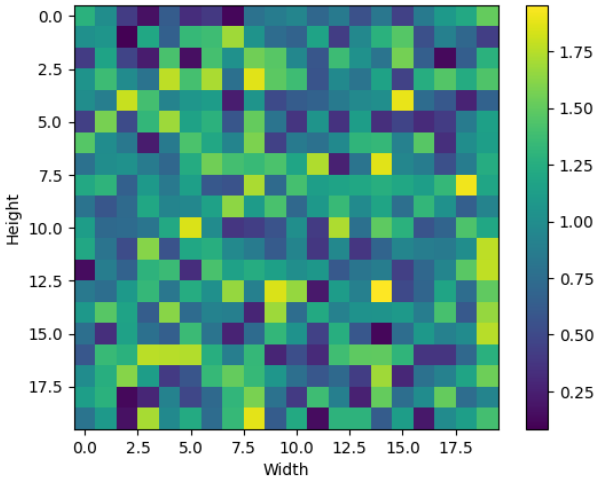

Рис. 10. Grad-CAM візуалізація для відео

Формалізація Grad-CAM:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k),$$

де A^k – feature maps, α^c_k– ваги важливості:

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}.$$

ROC-крива підтверджує ефективність класифікаційної моделі. AUC є ключовою метрикою якості, яку необхідно при наступних дослідженнях покращити. Grad-CAM забезпечує інтерпретованість. Модель фокусується на критичних ділянках обличчя, що характерно для задачі виявлення діпфейків

Висновки

У статті розглянуто проблему виявлення діпфейків як одну з актуальних задач сучасної інформаційної безпеки та цифрової медіааналітики. Запропоновано мультимодальний підхід до детекції синтетичного контенту, який базується на поєднанні відео-, аудіо- та текстових даних із використанням методів глибинного навчання. Розроблено математично обґрунтований пайплайн, що включає етапи попередньої обробки даних, виділення ознак за допомогою нейронних мереж, застосування механізмів self-attention і cross-attention, а також мультимодального об’єднання ознак. Особливу увагу приділено використанню трансформерних архітектур для моделювання внутрішньо- та міжмодальних залежностей. Проведені експериментальні дослідження продемонстрували, що запропонований підхід забезпечує підвищення точності виявлення діпфейків порівняно з одномодальними моделями. Отримані результати підтверджують ефективність інтеграції різних типів даних та доцільність використання attention-механізмів для підвищення інтерпретованості моделі. Запропонований підхід може бути використаний у системах автоматизованого аналізу медіаконтенту, цифрової криміналістики та протидії інформаційним загрозам.

У даному дослідженні розроблено та проаналізовано підхід до виявлення діпфейків на основі мультимодального аналізу із застосуванням сучасних методів глибинного навчання, зокрема трансформерних архітектур та attention-механізмів. У ході роботи сформовано повний математично обґрунтований пайплайн обробки даних, що включає:

– формалізацію мультимодальних вхідних даних (відео, аудіо, текст);

– попередню обробку та сегментацію;

– виділення ознак за допомогою глибинних нейронних мереж;

– застосування self-attention для моделювання внутрішніх залежностей;

– використання cross-attention для інтеграції міжмодальних зв’язків;

– мультимодальне об’єднання ознак;

– класифікацію та оптимізацію моделі;

– формування фінального рішення в реальному часі.

Отримані експериментальні результати показали, що використання мультимодального підходу забезпечує суттєве підвищення якості виявлення діпфейків. Зокрема, модель, яка інтегрує відео, аудіо та текстові модальності із застосуванням механізмів cross-attention, досягла найкращих показників точності (Accuracy ≈ 0.95) та F1-міри (≈ 0.925), що перевищує результати моделей, які використовують лише одну модальність. Проведений аналіз attention-механізмів показав, що модель здатна ефективно фокусуватися на критичних ділянках даних, зокрема аномаліях міміки та руху губ у відео, синтетичних артефактах у голосі тасемантичних невідповідностях у тексті. Візуалізація attention-карт та їх тривимірне представлення підтвердили, що мультимодальна інтеграція дозволяє виявляти складні міжмодальні залежності та невідповідності, які є характерними для діпфейків і не можуть бути виявлені при аналізі окремих модальностей.

Запропонований підхід має такі основні переваги:

– підвищена точність завдяки інтеграції різних джерел інформації;

– здатність виявляти приховані та складні патерни діпфейків;

– висока інтерпретованість результатів через attention-механізми;

– можливість застосування в реальному часі.

Водночас дослідження має певні обмеження, серед якихвисока обчислювальна складність моделей трансформерного типу, залежність якості результатів від синхронізації модальностей та чутливість до шумів у аудіо та відео даних.

Подальші напрями досліджень можуть включати:

– оптимізацію архітектури для зменшення обчислювальних витрат;

– використання більш ефективних варіантів attention (sparse, linear attention);

– розширення набору модальностей (наприклад, біометричні або поведінкові дані);

– застосування методів explainable AI для глибшого аналізу рішень моделі;

– адаптацію моделі до умов потокової обробки великих обсягів даних.

Запропонований мультимодальний підхід є ефективним та перспективним рішенням для задачі виявлення діпфейків і може бути використаний у практичних системах інформаційної безпеки, медіааналізу та цифрової криміналістики.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser Ł., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
2. Xu, Y., Wei, H., Lin, M. et al. Transformers in computational visual media: A survey. Comp. Visual Media 8, 33–62 (2022). https://doi.org/10.1007/s41095-021-0247-3
3. Hafiz, A. M., Parah, S. A., & Bhat, R. U. A. (2021). Attention mechanisms and deep learning for machine vision: A survey of the state of the art. arXiv preprint arXiv:2106.07550. https://doi.org/10.48550/arXiv.2106.07550
4. Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113-12132. https://doi.org/10.1109/TPAMI.2023.3275156
5. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 6558-6569). https://doi.org/10.48550/arXiv.1906.00295
6. Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications, 241, 122666. https://doi.org/10.48550/arXiv.2306.07303
7. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. Journal of Imaging, 9(6), 122. https://doi.org/10.3390/jimaging9060122
8. Raza, M. A., & Malik, K. M. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 993-1000). https://doi.org/10.3390/info17040347
9. Erokhin, D., & Komendantova, N. (2026). A Review of Tools and Technologies to Combat Deepfakes. Information, 17(4), 347. https://doi.org/10.3390/info17040347
10. Nailwal, S., Singhal, S., Singh, N. T., & Raza, A. (2023, November). Deepfake detection: A multi-algorithmic and multimodal approach for robust detection and analysis. In 2023 international conference on research methodologies in knowledge management, artificial intelligence and telecommunication engineering (RMKMATE) (pp. 1-8). IEEE. https://doi.org/10.1109/RMKMATE59243.2023.10369155
11. Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M., & Ghag, K. (2024). A multimodal framework for deepfake detection. arXiv preprint arXiv:2410.03487. https://doi.org/10.48550/arXiv.2410.03487
12. Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520
13. Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: a survey of deep learning techniques. Social Network Analysis and Mining, 13(1), 101. https://doi.org/10.1007/s13278-023-01104-w
14. Sedik, A., Faragallah, O. S., El-sayed, H. S., El-Banby, G. M., El-Samie, F. E. A., Khalaf, A. A., & El-Shafai, W. (2022). An efficient cybersecurity framework for facial video forensics detection based on multimodal deep learning. Neural Computing and Applications, 34(2), 1251-1268. https://doi.org/10.1007/s00521-021-06416-6
15. Vysotska, V., Smelyakov, K., Chupryna, A., Darahan, D., Torubara, O., & Shyshymenko, O. (2025). Social engineering in Ukraine: Threats and intelligent detection approaches. In CEUR Workshop Proceedings (Vol. 4110, pp. 317-331). https://ceur-ws.org/Vol-4110/paper24.pdf
16. Tan, D., Yang, Y., Niu, C., Li, S., Yang, D., & Tan, B. (2025). A review of deep learning based multimodal forgery detection for video and audio. Discover Applied Sciences, 7(9), 987. https://doi.org/10.1007/s42452-025-07629-3
17. Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. PeerJ Computer Science, 10, e2037. https://doi.org/10.7717/peerj-cs.2037
18. Vysotska, V., Nazarkevych, M., Vladov, S., Lozynska, O., Markiv, O., Romanchuk, R., & Danylyk, V. (2024). Devising A Method For Detecting Information Threats In The Ukrainian Cyber Space Based On Machine Learning. Eastern-European Journal of Enterprise Technologies, 132(2). 132, Issue 2, p36. https://doi.org/10.15587/1729-4061.2024.317456

Received (Надійшла) 03.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Висоцька Вікторія Анатоліївна – доктор технічних наук, доцент, професор кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна; професор кафедри протидії кіберзлочинності, Харківський національний університет внутрішніх справ, Харків, Україна;
Victoria Vysotska – Doctor of Technical Sciences, Associate Professor, Professor of the Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine; Professor of Counteracting Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, Ukraine;
e-mail: victoria.a.vysotska@lpnu.ua; ORCID Author ID: https://orcid.org/0000-0001-6417-3689;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=24484045400.

Чирун Любомир Вікторович – кандидат технічних наук, доцент, доцент кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна; доцент кафедри комп'ютерних наук, Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна;
Lyubomyr Chyrun – Candidate of Technical Sciences, Associate Professor, Associate Professor, Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine; Associate Professor, Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine;
e-mail: lyubomyr.v.chyrun@lpnu.ua; ORCID Author ID: https://orcid.org/0000-0002-9448-1751;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55225672300.

Лаврут Олександр Олександрович – доктор технічних наук, професор, професор кафедри тактико спеціальних дисциплін, Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна;
Oleksandr Lavrut – Doctor of Technical Sciences, Professor, Professor at the Department for Tactical and Special Disciplines Hetman Petro Sahaidachnyi National Army Academy, Lviv, Ukraine;
e-mail: alexandrlavrut@gmail.com; ORCID Author ID https://orcid.org/0000-0002-4909-6723;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217195493.

Лаврут Тетяна Валеріївна – кандидат географічних наук, доцент, старший дослідник, провідний науковий співробітник науково-дослідного відділу Наукового центру Сухопутних військ Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна;
Tetiana Lavrut – Candidate of Geographical Sciences, Associate Professor, Senior Researcher, Leading Researcher Army Scientific Center of the Hetman Petro Sahaidachnyi National Army Academy, Lviv, Ukraine;
e-mail: lavrut_t_v@i.ua; ORCID Author ID https://orcid.org/0000-0002-1552-9930;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217204350.

Романчук Роман Васильович – аспірант кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна;
Roman Romavchyk – PhD student of the Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine;
e-mail: roman.v.romanchuk@lpnu.ua; ORCID Author ID: https://orcid.org/0009-0004-4352-1073;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=58765557000.

Information technology of deepfake detection based on deep learning and multimodal analysis for intellectual information security systems

Victoria Vysotska, Lyubomyr Chyrun, Oleksandr Lavrut, Tetiana Lavrut, Roman Romavchyk

Abstract. Relevance. The rapid development of deep learning technologies has led to the emergence of high-quality synthetic media content (deepfakes), posing a significant threat to information security, digital trust, and the media space. Modern methods for detecting deepfakes, based on the analysis of individual modalities (video, audio, or text), often lack sufficient accuracy and generalizability, necessitating the development of multimodal approaches. **Object of research.** Processes of detecting synthetic media content (deepfakes) in the digital information environment. **Purpose of the article.** Development of an effective method for detecting deepfakes based on multimodal analysis using deep learning models and attention mechanisms. **Research results.** The paper proposes an information technology for detecting deepfakes based on complex processing of video, audio, and text data. A generalised pipeline has been developed that includes pre-processing of media content, feature extraction for each modality, multimodal integration, and classification. To increase efficiency, transformer architectures using self-attention and cross-attention mechanisms were employed, enabling modelling intra- and intermodal dependencies. Experimental studies on public datasets demonstrated that the proposed approach increases the accuracy of deepfake detection to 0.95 and the F1-measure to 0.925, exceeding the results of single-modal models. **Conclusions.** The results confirm the feasibility of a multimodal approach and attention mechanisms for deepfake detection. The proposed information technology provides increased accuracy and interpretability and can be used in information security systems, digital forensics, and automated media content analysis. Prospects for further research include optimising the computational complexity of models and adapting them for real-time streaming data processing.

Keywords: cybersecurity, deepfake, multimodal analysis, deep learning, transformers, attention mechanism, information security, synthetic media content, computer vision, audio processing, machine learning.