https://doi.org/10.31891/2307-5732-2025-351-74
УДК 004.773.2

ГУНЯ ВАСИЛЬ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0003-9796-9909
e-mail: vasyl.hunia.kn.2021@lpnu.ua
ШАХОВСЬКА НАТАЛІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-6875-8534
Nataliya.b.shakhovska@lpnu.ua

ІНФОРМАЦІЙНА СИСТЕМА ВИЯВЛЕННЯ ВІДЕО ПІДРОБОК ОБЛИЧ, СТВОРЕНИХ ЗА ДОПОМОГОЮ ШТУЧНОГО ІНТЕЛЕКТУ

У даній роботі розглянуто проблему виявлення підроблених відеозаписів облич людей, створених за допомогою технологій deepfake. Запропоновано інформаційну систему, що використовує гібридну нейронну мережу, яка поєднує згорткові нейронні мережі (CNN) для вилучення візуальних ознак та LSTM для аналізу часових залежностей. Завдяки інтеграції методу Grad-CAM, система надає можливість пояснювати прийняті рішення через візуалізацію критичних областей кадрів, що впливають на класифікацію. Проведені експерименти на наборах даних DFDC та FakeAVCeleb демонструють хорошу ефективність запропонованого підходу, що свідчить про його потенціал у боротьбі з дезінформацією та кіберзлочинністю.

Ключові слова: нейронні мережі, CNN, LSTM, Grad Cam, Deepfake.

HUNIA VASYL, SHAKHOVSKA NATALIYA
Lviv Polytechnic National University

INFORMATION SYSTEM FOR DETECTING AI-GENERATED FACE FORGERY VIDEOS

This study addresses the challenge of detecting manipulated videos featuring forged faces generated by deepfake technologies. We propose an information system based on a hybrid neural network architecture that integrates Convolutional Neural Networks (CNN) for extracting spatial features with Long Short-Term Memory (LSTM) networks for analyzing temporal dependencies. By incorporating the Grad-CAM technique, the system offers explainability by visualizing the key regions influencing the classification decision. Experiments conducted on the DFDC and FakeAVCeleb datasets demonstrate valuable effectiveness of the proposed approach, underlining its potential in the fight against disinformation and cybercrime.

The results on the DFDC set are slightly better than on FakeAVCeleb, because the model was trained on data from this set and is better adapted to its features. The results on the FakeAVCeleb set are slightly lower. The Precision, Recall, and F1-score metrics on the DFDC set for both classes are quite close, which indicates a balance in recognizing both real and fake videos. Compared to the results on DFDC, on FakeAVCeleb there is a decrease in the metrics (Accuracy, Precision, Recall, F1-score, AUC) and an increase in the loss value (BCE). For DFDC, the matrix shows an almost symmetric distribution of errors, as for FakeAVCeleb. The high AUC indicator (0.94 for DFDC and 0.83 for FakeAVCeleb) demonstrates that the model separates classes well regardless of the chosen classification threshold.

Explanatory AI (XAI) makes the model's work clearer by providing explanations for its decisions. As mentioned above, the system implements the model's explanations using Grad Cam. Grad-CAM visualizes the regions of the image that are important for classification using a heat map (red is high activity, blue is low). In real videos, we can see more "active" areas on the face, while in the case of a fake, the model focuses on various artifacts that occur in deepfakes. Grad Cam shows high activity in these areas (red).

Keywords: neural networks, CNN, LSTM, Grad Cam, Deepfake.

Стаття надійшла до редакції / Received 11.04.2025
Прийнята до друку / Accepted 26.04.2025

Вступ

Бурхливий розвиток технологій штучного інтелекту призвів до створення моделей, які здатні генерувати різноманітний "штучний" контент, який стало важко відрізнити від реального. Нова технологія, що дозволяє створювати високо реалістичний медіаконтенту, проводити різні маніпуляцій із зображеннями, відео- та аудіо записами отримала назву deepfake. Ця технологія переходить від традиційних графічних методів до підходів, заснованих на глибокому навчанні (DL). Deepfake технології стали дуже популярними після того, як у 2017 році з'явилося перше відео, на якому було замінене обличчя на обличчя знаменитості [1]. В 2023 році кількість діпфейків різко зросла. Згідно з даними, приблизно 500,000 таких було завантажено на ресурси соціальних медіа до кінця року [2].

Технологія deepfake має значний вплив на реальний та віртуальний світ. Вона створює такі ризики, як крадіжка персональних даних, комп'ютерне шахрайство, шантаж, маніпуляції з голосом або зображенням під час аутентифікації та створення фальшивих доказів. Наприклад, зображення та відео, що використовуються як докази в судових процесах та розслідуваннях, колись вважалися надійними доказами. На ранніх етапах розвитку технологій зображення зазвичай піддавалися більш очевидним змінам, як-от маніпуляції з якістю або вставкою несумісних елементів. Такі підробки було простіше визначити через "криве" редагування та відсутність кращих технологій для створення реалістичних зображень. Однак з появою технології deepfake довіряти таким доказам стає дедалі важче. Перевірка таких доказів перетворюється на складний процес, що сповільнює розгляд справ. Діпфейки відіграють все більш значущу роль в інформаційних війнах, виступаючи потужним інструментом дезінформації, маніпуляції громадською думкою та підриву довіри до медіа та політиків.

Мета цієї роботи полягає в тому, щоб розробити інформаційну систему, за допомогою якої користувачі могли б перевіряти відео на вміст “діпфейків” облич. У роботі було реалізовано гібридну нейронну мережу CNN-LSTM для перевірки відео вмісту. Пояснюваність моделі була забезпечена за допомогою підходу Grad CAM, який візуалізує теплову карту для кожного кадру з відео і дає розуміння, які саме області на кадрі найбільше вплинули на рішення моделі.

Аналіз літературних джерел

У галузі виявлення відео діпфейків дослідники пропонують широкий спектр підходів, які часто базуються на адаптації методів аналізу зображень до відео вмісту. Одним із поширених методів є усереднення прогнозів, отриманих для окремих кадрів, із фокусом на аналізі обличчя [3, 4]. Такі підходи зазвичай доповнюються спеціалізованими техніками, що враховують як просторові, так і часові характеристики відео.

У роботі Wubet, W.M. під назвою “The Deepfake Challenges and Deepfake Video Detection” запропоновано підхід, що базується на комбінації згорткових нейронних мереж (CNN) і рекурентних мереж LSTM. Автором використано архітектури ResNet і VGG-16 без попереднього навчання для класифікації станів очей (відкритих чи закритих), а LSTM-модель застосовувалася для аналізу послідовності кадрів. Дослідження проводилося на наборі даних UADFV, де аналізувались моргання очей протягом певного інтервалу часу. Зокрема, визначалися висота та ширина очей у різних станах, що дозволило виявляти аномалії, характерні для діпфейків. [5]. Інший підхід описано у статті Nguyen та співавторів “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos”. Тут використано концепцію капсульних мереж, розширену для виявлення різних типів діпфейків як на зображеннях, так і на відео. Метод включає кілька етапів: попередню обробку, під час якої відео розбивається на кадри з подальшим виявленням і масштабуванням облич; витяг прихованих ознак за допомогою мережі VGG-19; класифікацію за допомогою капсульної мережі; а також постобробку, де обчислюються середні ймовірності для визначення автентичності. Автори порівняли ефективність цього методу з еталонними підходами, продемонструвавши його конкурентоспроможність [6].

Розвиток гібридних архітектур для аналізу просторово-часових характеристик відео представлено у роботі Kaddar та співавторів під назвою “Deepfake Detection Using Spatiotemporal Transformer”. У цьому дослідженні поєднано згорткові нейронні мережі (CNN) із трансформерною моделлю Vision Transformer (ViT). CNN на базі Xception вилучає локальні просторові ознаки з кадрів, тоді як ViT завдяки механізму самоуваги забезпечує глобальне контекстуальне навчання та виявлення довготривалих залежностей. Попередня обробка включає розбиття відео на кадри, обрізання та вирівнювання облич для мінімізації впливу фонового шуму. Експерименти на наборах даних FaceForensics++, DFDC-p і Celeb-DF показали високу здатність моделі до узагальнення [7].

Методи та моделі

Опис наборів даних. У роботі було обрано набір даних Deepfake Detection Challenge. Цей набір містить більше ніж 124,000 коротких відео (близько 10 секунд тривалістю) записаних з допомогою 3426 оплачуваних акторів та поділений на 50 підвибірок. Для тренування були використанні відео з 0-6 та 9-11 підвибірок (3978 відео). Для валідації і тестування було використано 8 (468 відео) та 7 (573 відео) підвибірки відповідно, що приблизно відповідає поділу 80/10/10.

Також для тестування було використано набір даних FakeAVCeleb, а саме підвибірки, де підроблені відео були згенеровані за допомогою наступних методів: Faceswap, and FSGAN. FakeAVCeleb поділено на чотири групи; справжні відео зі справжнім аудіо, справжні відео з підробленим аудіо, підроблені відео з справжніми аудіозаписами, і підроблені відео з підробленими аудіозаписами. FakeAVCeleb містить короткі відео відомих людей різних рас та статті, які дозволяють краще оцінити роботу моделі. Для тестування було використано 1000 відео з двох групи, а саме справжні відео зі справжнім аудіо, підроблені відео з справжнім аудіо. Кількість кадрів за секунду в обох наборах знаходиться в межах 25-30 fps.

Обробка відео даних. Було використано відео у яких лише одне обличчя, адже модель на вхід приймає одну послідовність обличчя і класифікує її як підроблену чи реальну, що дає більш детальний аналіз відео. Процес підготовки даних складався з кількох етапів. На рисунку 1 наведено візуалізацію цього процесу.

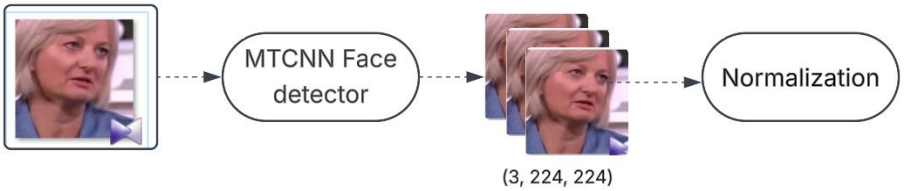

Рис 1. Процес попередньої обробки відео

Спочатку з кожного відео рівномірно витягувалося 20 кадрів, щоб забезпечити репрезентативну вибірку протягом усього відеоряду. Для виявлення обличчя на кожному кадрі використовувалася нейронна мережа MTCNN (Multi-task Cascaded Convolutional Neural Network), яка точно локалізує ключові точки обличчя, такі як очі, ніс і рот, та визначає його межі. Щоб захопити

більше контекстної інформації, наприклад, частину шиї чи волосся, рамка навколо виявленого обличчя розширювалася на 30% від початкового розміру, встановленого за допомогою MTCNN. Після цього всі отримані кадри облич приводилися до однакового розміру 224×224 пікселів. На завершальному етапі проводиться нормалізація: значення пікселів трансформувалися до діапазону [0, 1], що покращує стабільність і точність класифікації.

Ідея методу та алгоритм проведеного експерименту. Для початку відбувається попередня обробка відео, подана у розділі вище. Для класифікації відео послідовностей облич була розроблена гібридна нейронна мережа, яка поєднує згорткову нейронну мережу (CNN) для вилучення просторових ознак із кадрів і рекурентну мережу (LSTM) для аналізу часових залежностей. У роботі було використано попередньо натреновану модель EfficientNet-B3, яка була навчена на наборі даних ImageNet. Вихідний розмір ознак мережі становить 1536. У нашій реалізації використовується лише частина EfficientNet-B3 — до останнього шару перед класифікатором, що дозволяє отримувати компактні та інформативні вектори ознак для кожного кадру. Основний блок мережі EfficientNet – MBConv використовує SE-блоки (Squeeze-and-Excitation), які динамічно зважують канали, посилюючи важливі ознаки (так звана канальна увага). Під час тренування усі шари моделі були розморожені. Після вилучення ознак із кадрів вони передаються в рекурентну мережу LSTM, яка аналізує послідовність кадрів у часі. Двонаправлена оброка у LSTM була відключена. Було використано два шари LSTM та між ними було додано dropout зі значенням 0.1 для уникнення перенавчання. Останній вихід LSTM, що містить інформацію про всю послідовність, передається у повнозв'язний класифікатор, що складається з двох шарів: перший нелінійно трансформує вхід, а другий видає підсумкову оцінку (ймовірність фальсифікації). Між шарами додано dropout для уникнення перенавчання. Поріг для фейку встановлений за замовчування 0.5.

Для пояснення роботи моделі інтегровано метод Grad-CAM (Gradient-weighted Class Activation Mapping), який візуалізує області кадрів, що найбільше вплинули на ухвалення рішення нейромережею. Grad-CAM використовує градієнти виходу класифікатора відносно активацій останнього згорткового шару для створення теплової карти, яка показує важливі області на зображенні. У моделі використовується hook для збереження активацій та градієнтів із останнього згорткового шару CNN, що дозволяє в подальшому генерувати теплові карти важливих областей для ухвалення рішення. Таким чином, Grad-CAM допомагає пояснити, які частини обличчя або інших областей відеокадру містять характерні артефакти. Під час навчання моделі була використана L2 регуляризація. Для моніторингу функції втрат та уникнення перенавчання було використано Early stopping callback. Для тренування мережі обрані гіперпараметри, подані в Таблиці 1.

Таблиця 1

Гіперпараметри при тренуванні моделі

Назва гіперпараметру	**CNN-LSTM**
Оптимізатор	Adam
Коефіцієнтом регуляризації L2	5e-4
Коефіцієнт швидкості навчання	0.00001
Dropout у LSTM	0.1
Dropout у класифікаторі	0.4
Розмір батчу	4
Кількість епох	25
Кількість шарів LSTM	2
Кількість нейронів у прихованому стані LSTM	32

Функція втрат – BCEWithLogitsLoss. Для кожного кадру використовувались однакову аугментації за допомогою бібліотеки torchvision, щоб уся послідовність була оброблена аналогічно. Аугментацію включають RandomHorizontalFlip, RandomAffine, RandomRotation, ColorJitter, RandomGrayscale та Normalize (значення mean та std взяті з набору ImageNet). Для навчання було використано платформу Kaggle, а саме P 100 GPU.

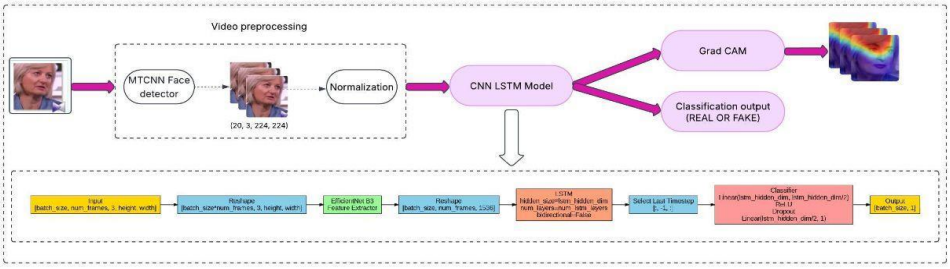

Рис 2. Візуалізація запропонованого алгоритму роботи системи

Експерименти

Для оцінки роботи моделі були використані наступні метрики: Accuracy, Loss (значення функції втрат), Recall, Precision, f1-score та AUC. Нижче наведено графіки функцій втрат та точності під час тренування на тренувальних та валідаційних даних.

Таблиця 2

Значення метрик для CNN LSTM моделі

Метрики на тестовому наборі DFDC						
Клас	Precision	Recall	F1-score	Accuracy	Loss (BCE)	AUC
0	0.88	0.85	0.87	0.86	0.3699	0.94
1	0.84	0.87	0.85			
Метрики на тестовому наборі FakeAVCeleb						
Клас	Precision	Recall	F1-score	Accuracy	Loss (BCE)	AUC
0	0.76	0.75	0.75	0.75	0.5738	0.83
1	0.75	0.76	0.75			

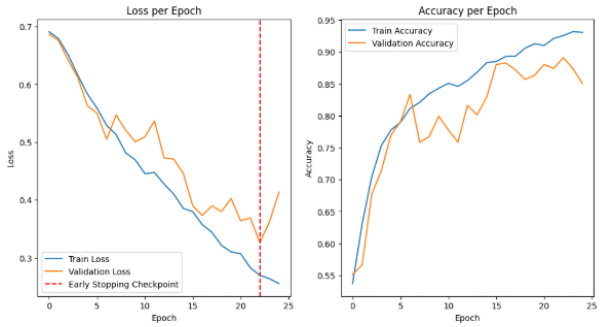

Рис. 3 Графік функції втрат та точність для CNN LSTM моделі

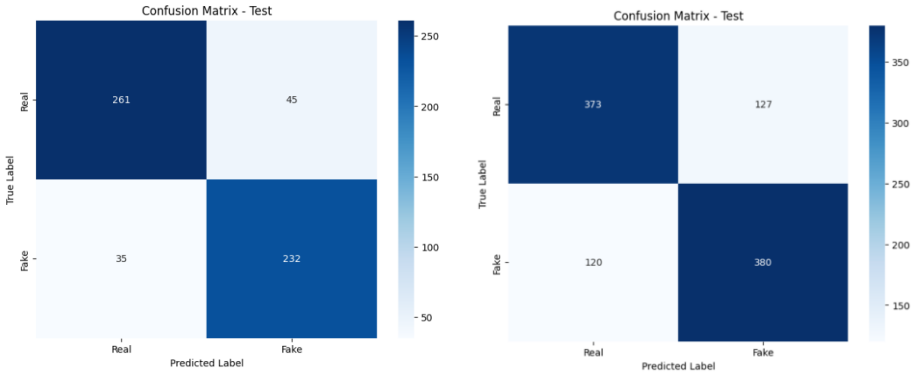

Рис. 4 Матриця плутанини для тестового набору DFDC (ліворуч) та для FakeAVCeleb (праворуч)

Результати на наборі DFDC дещо кращі, адже модель навчалась на даних з цього набору та краще адаптована до його особливостей. На наборі FakeAVCeleb результати трохи нижчі. Метрики Precision, Recall та F1-score на наборі DFDC для обох класів досить близькі, що свідчить про збалансованість у розпізнаванні як реальних, так і фейкових відео. Порівняно з результатами на DFDC, на FakeAVCeleb спостерігається зниження метрик (Accuracy, Precision, Recall, F1-score, AUC) та підвищення значення втрат (BCE). Для DFDC матриця демонструє майже симетричний розподіл помилок, як і для FakeAVCeleb. Високий показник AUC (0.94 для DFDC та 0.83 для FakeAVCeleb) демонструє, що модель добре розділяє класи незалежно від обраного порогу класифікації.

Grad-CAM візуалізує важливі для класифікації регіони зображення за допомогою теплової карти (червоний – висока активність, синій – низька). У реальних відео ми можемо бачити більше “активних” областей на обличчі (див. рис. 5), тоді як у випадку підробки модель фокусується на різних артефактах, які виникають при діпфейках. Так на рис. 6 бачимо на обличчі артефакти на носі та щоках. Grad Cam у цих областях показує високу активність (червоний колір).

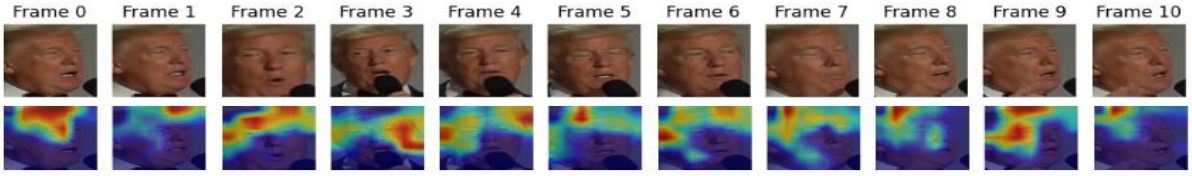

Рис 5. Візуалізація Grad Cam для реального обличчя

Рис 6. Візуалізація Grad Cam для підробленого обличчя

Висновок

У цій роботі було запропоновано та експериментально обґрунтовано інформаційну систему виявлення діпфейків, що ґрунтується на гібридній архітектурі CNN-LSTM з інтегрованим методом Grad-CAM для пояснення рішень моделі. Експериментальні результати, отримані на наборах даних DFDC та FakeAVCeleb, свідчать про хорошу здатність моделі до узагальнення. На DFDC показники Precision, Recall, F1-score та AUC свідчать про високий рівень точності, тоді як на FakeAVCeleb спостерігається невелике зниження метрик, що пояснюється відмінностями в характеристиках даних. Впровадження Grad-CAM забезпечило прозорість прийнятих рішень, сприяючи підвищенню довіри до системи як інструменту для боротьби з кіберзлочинністю та дезінформацією. Загалом, запропонований підхід демонструє значний потенціал у контексті сучасних викликів інформаційної безпеки та може стати важливим інструментом у протидії діпфейкам.

Літератури

1. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., & Tao, D. (2024). *Deepfake generation and detection: A benchmark and survey*. arXiv. https://arxiv.org/abs/2403.17881v3

2. Ulmer A, Tong A (2023) Deepfaking it: America's 2024 election collides with AI boom. https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-3

3. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.pdf

4. Wang, Z., Bao, J., Zhou, W., Wang, W., & Houqiang, L. (2023). AltFreezing for more general video face forgery detection. *arXiv preprint* arXiv:2307.08317. https://arxiv.org/abs/2307.08317

5. Wubet, W.M. (2020) The deepfake challenges and deepfake video detection. *Int. J. Innov. Technol. Explor. Eng.*, *9*, https://www.ijitee.org/wp-content/uploads/papers/v9i6/E2779039520.pdf

6. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) https://ieeexplore.ieee.org/document/8682602

7. Kaddar, B., Fezza, S. A., Akhtar, Z., Hamidouche, W., Hadid, A., & Serra-Sagristà, J. (2024). Deepfake detection using spatiotemporal transformer. *ACM Transactions on Multimedia Computing, Communications, and Applications*. https://doi.org/10.1145/3643030

References

1. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., & Tao, D. (2024). *Deepfake generation and detection: A benchmark and survey*. arXiv. https://arxiv.org/abs/2403.17881v3

2. Ulmer A, Tong A (2023) Deepfaking it: America's 2024 election collides with AI boom. https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30

3. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.pdf

4. Wang, Z., Bao, J., Zhou, W., Wang, W., & Houqiang, L. (2023). AltFreezing for more general video face forgery detection. *arXiv preprint* arXiv:2307.08317. https://arxiv.org/abs/2307.08317

5. Wubet, W.M. (2020) The deepfake challenges and deepfake video detection. *Int. J. Innov. Technol. Explor. Eng.*, *9*, https://www.ijitee.org/wp-content/uploads/papers/v9i6/E2779039520.pdf

6. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) https://ieeexplore.ieee.org/document/8682602

7. Kaddar, B., Fezza, S. A., Akhtar, Z., Hamidouche, W., Hadid, A., & Serra-Sagristà, J. (2024). Deepfake detection using spatiotemporal transformer. *ACM Transactions on Multimedia Computing, Communications, and Applications*. https://doi.org/10.1145/3643030