Multimodal Audio Visual Fake Detection with Explainable Artificial Intelligence

Dr.D.R.Jiji Mol[1,2], Dr. Deepak Gupta[3]
[1] Postdoctoral Researcher, Lincoln University College, Malaysia. Email: pdf.jiji@lincoln.edu.my
[2]Assistant Professor, SRM Arts and Science College, Tamilnadu, India. Email: dr.jiji@gmail.com
[3]Professor, Maharaja Agrasen Institute of Technology, India. Email:drdeepakgupta.cse@gmail.com

Abstract:

The development of deepfake generation methods has brought a number of challenges in the field of digital forensics, especially realism and multimodality of manipulated content. Most of the current detection methods are unimodal, which are not interpretable and not effective in real world situations. To overcome these challenges, a multimodal and explainable deepfake detection framework AV-Fake detection is proposed. The AV-Fake Detection is an advanced deepfake detection framework developed by using multimodal approach and explainable AI to boost the accuracy and explainability of deepfake detection. The proposed framework mitigates the shortcomings of extent unimodal system and enhancing the reliability of forensic tasks, social media management and validation of digital content. This innovation architecture establishes a scalable framework for future intelligent multimedia forensic systems.

Keywords: Deepfake detection, Multimodal approach, Explainable AI, Audio-Visual forensics

Introduction

The deepfake, a manipulated image, video or audio created with deeplearning technologies like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), g Generative Adversarial Networks (GANs) and more recently diffusion-based models. It is a serious threat to security and society as it is difficult to distinguish the real content and the manipulated content. With the high quality deepfakes being produced, there are serious security issues arising, especially on social media platforms such as rapid spread of misinformation, identity theft, blackmail, and creation of fake evidence. This leads to a situation that is challenging to keep accurate data in the digital world and has sparked a renewed sense of the need to create a fake content detection methods [1][2].

Even though there is lot progress in development of detection methods, the majority of the existing methods work on either images, or video or audio signals. These methods have their own technical challenges. In image-based detection, current generative models can remove some of the artefacts, making it harder to extract discriminative features. Modelling both the spatial and temporal dependency is difficult challenge, especially in case of subtle manipulation across the video sequence.
Recent studies have been conducted to improve the accuracy of the system by combining CNN and Long Short-Term Memory (LSTM)-Transformer and Vision. This architecture yields better results, though generalization and robustness continues to be a challenge [3][4]. The robustness is strongly influenced by detection, environmental noise, compression artefacts and variations in recording conditions in the audio-based detection methods [5]. These unimodal systems are less effective under real-world scenarios.

In addition to their high detection performance, the multimodal approaches raise new challenges, especially for feature fusion, cross-modal alignment and model interpretability. The fusion of heterogeneous features is still a challenging task and it results loss of information or model bias [6]. The artificial intelligence models that make the detection process incomparable to users, particularly in the fields such as security, forensic computing, and public policy. Thus the Explainable Artificial Intelligence (XAI) tools have gained importance in making the decision making process understandable [7]. There is a lack of friendly systems that combine high detection performance with interpretable and accessible outputs. The restriction emphasizes the importance of multi-modal, transparent and cohesive deepfake detection.

Related work

A framework called YOLO[8] is used to detect faces in video frames, utilizing Local Binary Patterns Histogram as a temporal features and the model is trained on FeceForensics++ and CelebDF datasets with the model backbone EfficientNet-B6. A Image Quality Assessment (IQA) metric system is proposed [9] to detect AI generated or manipulated videos, that includes the key components namely, Curvelet Quality Assessment, Blind Image Quality Assessment, Natural Image Quality Assessment and Fourier Transform. It achieves the accuracy of 99%.

A hybrid strategy[10] is proposed by merging the generative capabilities of GAN with REsNet for distinguishing real from fake. By leveraging the generative aspect of GAN and the discriminative poewer of ResNet, the hybrid model achieved an accuracy rate of 83% and an AUC score of 82.5%. A system known as the fake spotter was introduced for deepfake detection. It is monitoring neuron behaviour via the identification of activated neurons for classification purposes and reported 90.6% of accuracy [11]. In another study, an alternative Deepfake detection system was developed [12], using local features extracted using CNN, which was further improved by the addition of transformer blocks with retention for enhanced detection performance and achieved an accuracy of 97.66%.

Methods

A deepfake detection model, AV-Fake Detection is proposed to overcome the issues in the existing methods. The Figure-1 clearly shows the architecture of the proposed system.

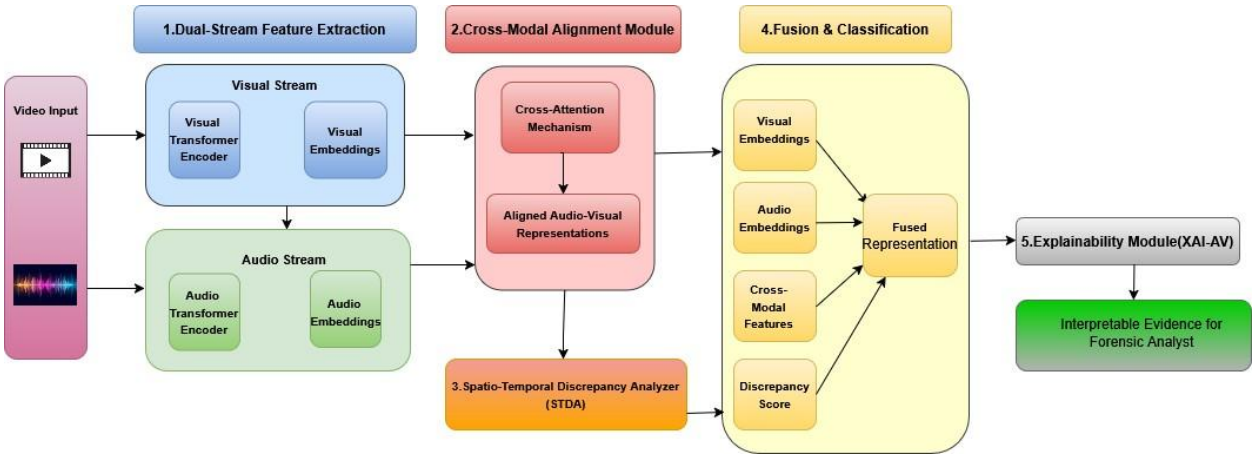

Figure-1: Architecture of AV-Fake Detection

The model consists of five modules namely, Dual-Stream Feature Extraction, Cross-Modal Alignment, Spatio-Temporal Discrepancy Analyzer, Fusion and Classification and XAI-AV. The Dual-Stream Feature Extraction module extracts both audio and video features separately. The video is divided into facial frames. The spatial artefacts such as blending, irregularities, lightening distortions and texture anomalies are captured through CNN. The identified spatial patterns are forwarded to Transformer encoder to capture the motion dynamics and facial expressions. In the same way, audio features are also extracted and stored.

These extracted features go to the input of the second module called Cross-Modal Alignment. By using these features this module lean synchronization between speech signals and associated lip movements. A cross-attention mechanism is used to learn the temporal communication between spoken phonemes and lip movement dynamics. The primary component of AV-Fake Detection is Spatio-Temporal Discrepancy Analyzer. It identifies the conflicts in both spatial appearance and temporal motion patterns. This allows the detection of forced transitions, unwanted movements or abnormal expression changes that are not visible in a single frame.

The Fusion and Classification module combines the multimodal information from audio, visual and spatio-temporal inconsistency representations to produce a final genuine estimation. This information is passed through an explainability module called XAI-AV. This module provides interpretable evidence to aid in the decision of deepfakes.

Experiments and Results

The study aims to present the performance evaluation of proposed system AV-Fake detection. It is compared with benchmark dataset FakeAVCeleb and a Real-Time Social Media dataset. The assessment is carried out using standard classification metrics such as Accuracy, Precision, Recall, F1 Score and Area Under the ROC Curve (AUC). The performance result shown in Table-1. The result depicts the generalization capability of the model in the real-time scenarios.

Table 1: Performance evaluation

Method	**Accuracy (%)**		**Precision (%)**		**Recall (%)**		**F1-score (%)**		**AUC (%)**	
	FakeAV Celeb	Real-Time	FakeAV Celeb	Real-Time	FakeAV Celeb	Real-Time	FakeAV Celeb	Real-Time	FakeAV Celeb	Real-Time
AV-Fake Detection	97.36	96.12	96.92	95.81	97.10	95.54	97.01	95.67	98.85	97.94

Discussions

The research work introduces a novel multimodal deepfake detection architecture AV-Fake detection. It is assigned to exploit audio-visual cues for enhancing the accuracy of forgery detection. The experimental finding shows that the proposed model surpasses existing state-of-the-art models.

In the world of digital forensics, the model significantly authenticates the multimedia evidences. Moreover, its capability extends to cyber security systems for detecting synthetic media attacks such as audio and video spoofing. Beyond its multimodal nature, the framework is well-suited for integration into

surveillance system requiring audio-visual consistency. Thus, the framework serves as a foundational tool for developing effective real-time authentication solutions in today's world.

Conclusions

The deepfake detection becomes more critical, as the deepfake generation methodologies increased sophisticatedly. The study presents a multimodal deepfake detection framework called AV-Fake Detection. It combines the audio and visual features with XAI to overcome the limitations of existing unimodal approaches. The framework leverages a comprehensive analysis of facial attributes, motion dynamics, speech traits and their temporal correlation to detect the modifications that may be missed by conventional detection methods. Moreover, its explainability promotes the reliability by providing explainable rationale for each prediction.

References

1. Kawa P, Plata M, Czuba M, Szymanski ´ P, Syga P, "Improved DeepFake Detection using Whisper Features," Proc Annu Conf Int Speech Commun Assoc Interspeech, pp. 4009–4013 August 2023.
2. Almars AM, "Deepfakes Detection Techniques using Deep Learning: a Survey," J Comput Commun., Vol.9, ISS.5, pp. 20–35, 2021.
3. Soudy AH, et al., " Deepfake detection using convolutional vision transformers and convolutional neural networks," Neural Comput Appl, Vol.36, Iss.31, pp. 19759–75, 2024
4. Petmezas G, Vanian V, Konstantoudakis K, Almaloglou EEI, Zarpalas D, " Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification," Multimed Tools Appl, Vol.84, Iss.33, pp. 40617–36,2025.
5. Zhao B, et al., "Generalized audio deepfake detection using frame-level latent information entropy," Proc - IEEE Int Conf Multimed Expo, 2025.
6. Y. Li, Z. Wang, and S. Liu, "Enhance carbon emission prediction using bidirectional long short-term memory model based on text-based and data-driven multimodal information fusion," J. Clean. Prod., vol. 471, May 2024.
7. Momin MS, Sufian A, Barman D, Leo M, Distante C, Damer N, "Explainable deepfake detection across different modalities: an overview of methods and challenges," Image vis Comput, Vol.163, Aug 2025.
8. Š. Hubálovský, P. Trojovský, N. Bacanin, and K. Venkatachalam, ''Evaluation of deepfake detection using YOLO with local binary pattern histogram,'' PeerJ Comput. Sci., vol. 8, p. e1086, Sep. 2022.
9. S. Kiruthika and V. Masilamani, ''Image quality assessment based fake face detection,'' Multimedia Tools Appl., vol. 82, no. 6, pp. 8691–8708, Mar. 2023.
10. S. Safwat, A. Mahmoud, I. E. Fattoh, and F. Ali, ''Hybrid deep learning model based on GAN and RESNET for detecting fake faces,''IEEE Access, vol. 12, pp. 86391–86402, 2024.
11. R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, ''FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces,'' 2019, arXiv:1909.06122.
12. T. Wang, H. Cheng, K. P. Chow, and L. Nie, ''Deep convolutional pooling transformer for deepfake detection,'' ACM Trans. Multimedia Comput., Commun., Appl., vol. 19, no. 6, pp. 1–20, Nov. 2023.