Deepfake Detection Using Multimodal AI

Lalit Kumar Joshi[1], Dr. Sangeeta Joshi[2]*

[1]System Administrator Mata Gujri College, Fatehgarh Sahib, Punjab, India

[2]Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India

***Corresponding author**

DOI: https://doi.org/10.51584/IJRIAS.2025.100500033

Received: 12 May 2025; Accepted: 17 May 2025; Published: 05 June 2025

ABSTRACT

Deepfakes, synthetic media generated using deep learning techniques, have grown rapidly in quality and prevalence, posing serious threats to digital trust, personal security, and political integrity. Traditional detection methods, primarily focused on single modalities such as image or audio analysis, have become increasingly ineffective against advanced generation techniques. This paper explores the use of multimodal AI systems, which integrate visual, audio, and textual cues, to enhance the accuracy and robustness of deepfake detection. We present a comprehensive overview of current multimodal detection techniques, compare their performance against unimodal approaches, and highlight challenges and future directions in building reliable, real-time detection systems [4].

Keywords: Deepfake, Multimodal AI, Deep Learning, Fake Media Detection, Adversarial AI, Video Forensics

INTRODUCTION

The proliferation of deepfake technology has created a critical need for effective detection strategies. Deepfakes are forged media content—often videos—produced using techniques like Generative Adversarial Networks (GANs) that convincingly mimic human appearance and behavior. While the technology has applications in entertainment and education, its malicious uses, such as misinformation, identity theft, and harassment, have drawn global concern. Traditional detection approaches, which rely on single-source analysis (e.g., facial artifacts or audio inconsistencies), struggle to cope with the sophistication of modern deepfakes [1]. This paper discusses the evolution of detection methods and proposes the integration of multimodal AI systems for more effective and holistic deepfake identification.

DEEPFAKE DETECTION METHODS AND TECHNIQUES

Unimodal Detection Techniques

- Visual-based Detection: These methods analyze spatial and temporal inconsistencies in videos, such as unnatural blinking, facial warping, and lighting mismatches. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are often used [2].
- Audio-based Detection: Focuses on identifying inconsistencies in voice such as pitch, cadence, and background noise. Spectrogram analysis and audio fingerprinting are typical techniques [5].
- Textual Analysis: In cases where deepfakes include spoken or written text, NLP techniques analyze semantic coherence and speaker consistency [8].

Multimodal Detection Techniques

- Fusion Models: Combine visual and audio modalities using attention mechanisms or feature-level fusion. These models outperform unimodal systems by capturing cross-modal inconsistencies [4].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue V May 2025

- Multimodal Transformers: Leverage pre-trained models like CLIP or VideoBERT, capable of learning contextual embeddings across different modalities [9].
- Temporal Multimodal Analysis: Use LSTM or Transformer-based models to detect inconsistencies across time in both audio-visual synchronization and textual coherence [4].

COMPARISON OF DETECTION METHODS

Unimodal techniques are computationally less expensive and easier to deploy but fall short in detecting well-crafted deepfakes. They are vulnerable to adversarial attacks tailored to exploit specific weaknesses [6]. Multimodal methods, while more complex, provide a significant boost in detection accuracy. Studies show that multimodal systems can achieve over 90% detection accuracy on benchmark datasets like DFDC (DeepFake Detection Challenge) compared to ~75% for unimodal systems [7]. Moreover, multimodal models are better at generalizing to unseen types of deepfakes and offer improved robustness in real-world scenarios [9].

Table 1: Comparison of Deepfake Detection Methods

Detection Method	Modalities Used	Accuracy (Approx.)	Strengths	Limitations
Visual-based	Image/Video	~75%	Simple to deploy; detects visual inconsistencies	Weak against realistic manipulations
Audio-based	Audio	~70%	Effective in voice tampering detection	Vulnerable to high-quality synthesis
Textual-based	Text/NLP	~68%	Useful for transcript or speech analysis	Needs accurate speech-to-text
Multimodal Fusion	Audio + Video	~91%	Detects cross-modal inconsistencies	Requires complex architecture
Multimodal Transformers	Audio + Video + Text	~93%	High contextual understanding	High computational resources needed

CONCLUSION

Deepfake content poses a multifaceted threat to digital society, requiring equally sophisticated detection techniques. Multimodal AI presents a promising direction by integrating complementary data streams to expose synthetic media [4]. Despite challenges such as high computational cost, data scarcity, and the need for interpretability, multimodal approaches show superior performance and adaptability. Future research should focus on lightweight, real-time multimodal detection systems and explore ethical implications of automated deepfake identification [9].

BIBLIOGRAPHY

1. Korshunov, P., & Marcel, S. (2018). Deepfakes: A New Threat to Face Recognition? Assessment and Detection. arXiv preprint arXiv:1812.08685.
2. Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW).
3. Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-Stream Neural Networks for Tampered Face Detection. CVPR Workshops.
4. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of Multimodal Biometrics for the Detection of Deepfakes. arXiv preprint arXiv:1907.06559.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue V May 2025

5. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. arXiv preprint arXiv:1905.00582.
6. Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS).
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:1910.08854.
8. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. NeurIPS.
9. Jiang, H., Li, J., Wu, Y., & Kankanhalli, M. (2020). Defending Deepfakes with Adaptive Multimodal Learning. ACM Multimedia.