# Multimodal Deepfake Audio and Video Detection

*(13 size) A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of AI&DS**

**MULTI MODAL INFORMATION PROCESSING (23ALT3102E)**


Submitted by

**2310080065: K. PHANI TEJA**

**2310080073: D. VINAY**

**2310080074: K. ANANTH KRISHNA**

**2310080077: T. GOPALA KRISHNA**

Under the guidance of

**6218(DR. GANGAMOHAN PAIDI)**




Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075


FEB - 2025.

# Introduction

*(Minimum 200 words)*

**Brief introduction about your project area**

In the digital era, synthetic media or "deepfakes" have become one of the most critical challenges for information authenticity and trust. Deepfakes are generated using advanced deep learning techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, enabling the manipulation of faces, voices, and entire video streams with highly convincing realism. While these techniques have legitimate applications in entertainment, accessibility, and education, their misuse has created severe risks, including misinformation, impersonation, identity theft, and political manipulation.

Traditionally, detection efforts have focused on unimodal approaches—analyzing either the visual frames of videos or the audio speech streams. However, with the increasing sophistication of deepfake generation, unimodal detectors often fail when only one modality is manipulated while the other remains authentic. This has led to the rise of multimodal deepfake detection, which leverages both audio and visual cues. By correlating lip movements with speech, analyzing synchronization patterns, and extracting modality-specific inconsistencies, multimodal detectors can achieve higher robustness.

Recent advances in multimodal learning, particularly with transformer-based architectures, have shown significant promise. Benchmark datasets such as FakeAVCeleb, DFDC, and KoDF provide diverse real and forged content to train and evaluate detection systems.

# Literature Review/ Application Survey

*(Minimum 800 words)*

Rossler et al. [1] introduced the **FaceForensics++** dataset and proposed detection methods based on convolutional neural networks (CNNs) for visual-only deepfake detection. Their work demonstrated that visual artifacts such as unnatural eye blinking, mismatched facial expressions, and lighting inconsistencies could be exploited to detect manipulated videos. However, their approach was limited to the video modality and failed to address audio forgeries.

Dolhansky et al. [2] released the **Deepfake Detection Challenge (DFDC)** dataset, one of the largest multimodal resources, containing more than 100,000 manipulated videos. Their study benchmarked several detection methods and highlighted the importance of developing robust models that can generalize across unseen manipulation

techniques. Although multimodal in nature, most participating methods still focused heavily on visual cues rather than joint audio-visual analysis.

Korshunov and Marcel [3] investigated vulnerabilities of face recognition systems to deepfakes and proposed baseline detection techniques. Their work revealed that simple visual forensic traces can sometimes detect forgeries, but they emphasized the growing sophistication of generative models that reduce such artifacts, thereby demanding more advanced detection strategies.

Hashmi et al. [4] proposed **AVTENet**, an audio-visual transformer-based ensemble network inspired by human cognition. Unlike earlier CNN-based methods, AVTENet integrates three independent transformer classifiers—audio-only, video-only, and joint audio-visual networks—and combines their outputs using different fusion strategies. Trained on the FakeAVCeleb dataset, AVTENet achieved state-of-the-art accuracy, significantly outperforming CNN-based baselines. Importantly, the study demonstrated that transformers could model long-range dependencies and synchronization patterns, making them more robust for multimodal detection.

Katamneni and Rattani [5] introduced **MIS-AVoiDD**, addressing the modality gap issue in audio-visual fusion. They argued that direct feature concatenation of heterogeneous modalities often leads to suboptimal performance. Instead, MIS-AVoiDD projects features into two complementary subspaces—**modality-invariant representations** that capture common patterns across audio and video, and **modality-specific representations** that preserve unique modality details. These are fused using a transformer-based attention mechanism. Experimental results on FakeAVCeleb and KoDF datasets showed that MIS-AVoiDD outperformed existing multimodal detectors by margins of up to 18%, highlighting the importance of representation-level fusion.

Mittal et al. [6] explored emotion-based multimodal detection, proposing that forged audio-visual content often contains emotional mismatches between facial expressions and vocal tone. Their Siamese network leveraged triplet loss to compare perceived emotions across modalities, providing another angle for identifying manipulations. Similarly, Zhao et al. [7] focused on **audio-visual synchronization**, checking consistency between lip movements and speech, which proved effective in detecting lip-synced forgeries generated by models like Wav2Lip.

Taken together, these studies illustrate the progression from unimodal to multimodal detection, from CNN-based visual cues to transformer-based cross-modal analysis. While early methods relied heavily on handcrafted or visual-only artifacts, recent advances leverage **multimodal fusion, attention mechanisms, and cognitive-inspired designs** to address the growing sophistication of deepfake forgeries.

You can add Journals, conference and website links as references

# References


1. Rossler, A. et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” *IEEE ICCV*, 2019.
2. Dolhansky, B. et al., “The Deepfake Detection Challenge Dataset,” *arXiv:2006.07397*, 2020.
3. Korshunov, P., and Marcel, S., “Deepfakes: A New Threat to Face Recognition? Assessment and Detection,” *arXiv:1812.08685*, 2018.
4. Hashmi, A., Shahzad, S.A., Lin, C.W., Tsao, Y., & Wang, H.M., “AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection,” *IEEE TCDS*, 2025.
5. Katamneni, V.S., & Rattani, A., “MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection,” *arXiv:2310.02234v2*, 2023.
6. Mittal, T. et al., “Emotions Don’t Lie: A Multimodal Approach to Detecting Deepfakes,” *Proc. ACM ICMI*, 2020.
7. Zhao, H. et al., “Learning Audio-Video Correlations for Deepfake Detection,” *IEEE TIFS*, 2021.