ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 439-447, 2026

BREAKING THE DEEPFAKE ILLUSION: MULTIMODAL DETECTION OF VOICE AND FACIAL DEEPFAKES WITH MARS

Aakriti Dixit[1]*, Himanshu Kumbhaj[2]

[1]*Department of Information Technology, Shri Shankaracharya Professional University, Bhilai (C.G), India.*

[2]*Department of Computer Science Engineering (IoT & CSBCT) SSTC, Bhilai (C.G), India.*

Abstract

Recently artificial intelligence has made it possible to generate highly realistic digital media such as images, videos, and human voices. These technologies, known as generative AI, are widely used in applications such as virtual avatars, film production, automated dubbing, and content creation. While these advances have many creative and commercial benefits, they have also enabled the rapid growth of deepfakes and existing detection methods typically analyze speech or video independently and fail to assess cross-modal coherence in synchronized deepfake attacks. This paper proposes a Multimodal Authenticity Risk Scoring (MARS) framework that integrates audio and visual embeddings using multimodal to *model* synchronization consistency. A unified risk score combines modality-specific spoof probabilities with cross-modal divergence to generate an interpretable authenticity decision. To address real-world deployment constraints, a *three-tier cascaded inference architecture* is proposed, enabling scalable processing at platform level while maintaining detection reliability.

Keywords: Deepfake Detection, Voice Anti-Spoofing, Multimodal Fusion, Authenticity Risk Scoring.

* *Corresponding author*

1. Introduction

Advancements in generative artificial intelligence have enabled coordinated voice cloning and facial synthesis that threaten biometric identity verification systems., which are AI-generated or manipulated media designed to imitate real individuals. This has particularly impacted social media platforms where content spreads quickly and reaches large audiences. According to the Deeptrace State of Deepfakes report, the number of publicly available deepfake videos increased from around 14,000 in 2019 to more than 95,000 by 2023 shared daily across platforms such as Instagram, TikTok, and Facebook. This massive scale makes it difficult for platforms to verify the authenticity of every piece of content. As generative models continue to improve in realism, distinguishing between authentic and manipulated media is becoming increasingly important.

Downside is misuse, digital impersonation, and the degradation of trust in online content. Deepfake technologies allow attackers to manipulate a person's face or voice and generate content that falsely

represents them saying or doing something they never actually did. Such techniques can enable impersonation attacks or the unauthorized use of a person's digital likeness.

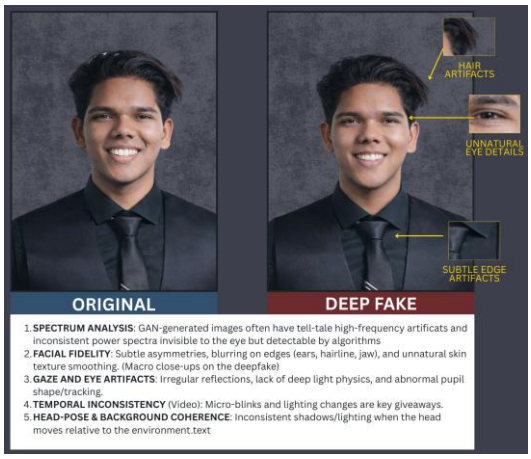

Fig 1.0: Original vs Deepfake Image Showing Common Synthetic Artifacts

In many cases, existing videos can be modified by replacing the original face with another individual's face

or by synchronizing a different voice track with the video, creating a misleading representation of events. These capabilities have already led to real-world risks; for instance, voice cloning has been used in financial fraud cases where attackers impersonated company executives to authorize fraudulent money transfers. Synthetic media shared on social media platforms can mislead large audiences, particularly during political events or public crises.

Consequently, developing reliable methods for detecting manipulated media and verifying content authenticity has become an important research and security priority.

Despite significant progress in deepfake detection research, many existing approaches focus on analyzing a single modality, such as facial manipulation in video frames or voice spoofing in audio signals. These unimodal detection methods often rely on identifying visual artifacts or acoustic irregularities produced during synthetic media generation. However, modern generative models, including diffusion-based systems, have significantly improved the realism of synthetic content and reduced many of the detectable artifacts used by traditional detection techniques. As a result, artifact-based methods are becoming less reliable. Furthermore, analyzing facial and vocal signals independently may fail to capture inconsistencies that emerge when multiple modalities are considered together, such as mismatches between lip movements and speech or inconsistencies between facial identity and voice identity. This highlights the need for detection frameworks that leverage multimodal analysis to better identify manipulated media.

To address these challenges, this work proposes the Multimodal Authenticity Risk Scoring (MARS) framework, a system designed to evaluate the authenticity of multimedia content using multiple

complementary signals. The proposed framework analyzes facial deepfake indicators, voice spoofing characteristics, and cross-modal consistency between lip movements and speech and provides an aggregated risk score to help platforms decide which content is authentic and hence should be published and AI generated so not to be published arises at deployment scale, where exhaustive multimodal analysis of every upload is computationally infeasible. This paper proposes the Multimodal Authenticity Risk Scoring (MARS) framework, which jointly encodes audio and visual signals, evaluates cross-modal coherence through attention-based fusion, and computes an interpretable authenticity risk score. A Flag and Analyze Architecture for Scalable Deployment is also introduced to enable scalable platform-level deployment.

2. Methodology

This proposed MARS- Multimodal Authenticity Risk Score Model model focuses on detecting unauthorized voice cloning and facial synthesis in synthetic media and has the following framework

1. Feature Extraction Layer

This layer processes the uploaded media and extracts structured representations from complex visual and audio signals and converts them into vectors that can be compared and analyzed further.

A. From video frames, the system extracts:

- Facial embeddings- representing biometric identity characteristics of the detected face.
- Facial landmark information - used to analyze lip movement, facial motion, and expression dynamics.

B. From audio signals, the system extracts:

- Voice embeddings, representing speaker identity characteristics.
- Acoustic features, including spectral and temporal characteristics of speech.

2. Detection Mechanisms

This layer analyzes the above extracted feature vectors to detect signs of synthetic generation from facial as well as vocal components using the following mechanism:

2.1 For Facial Deepfake Detection

2.1.1 *Artifact-Based Detection* - This method identifies visual irregularities commonly introduced during deepfake generation: a. texture inconsistencies b. unnatural lighting patterns c. boundary artifacts

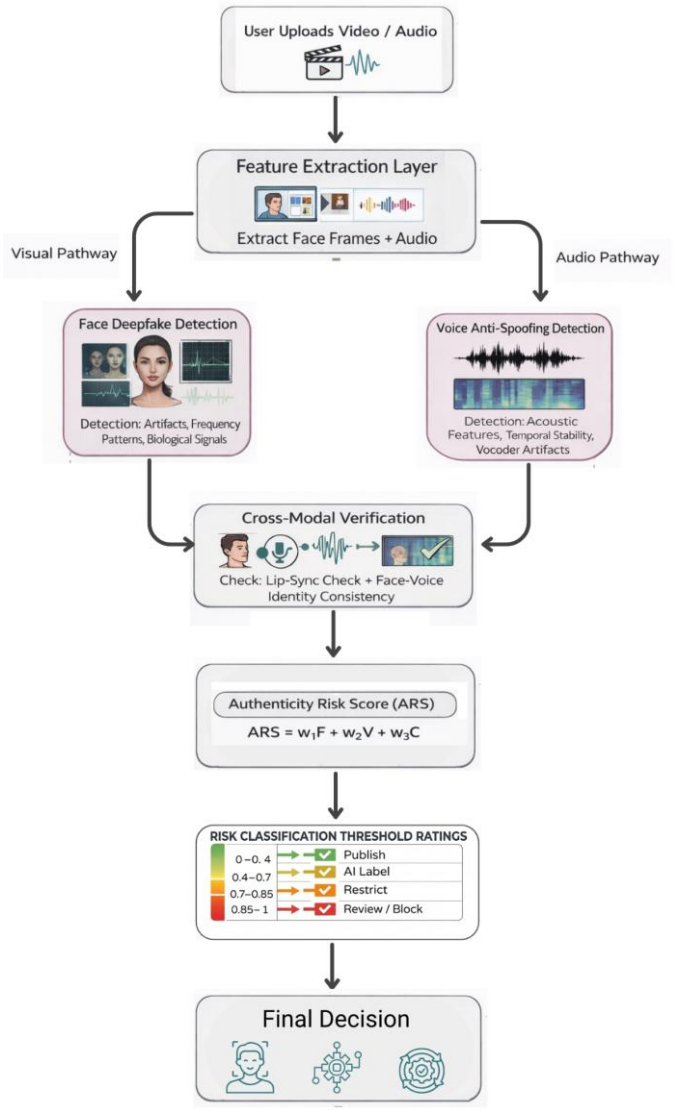

Fig 1.1: visualization of MARS Framework

2.1.2 *Frequency Domain Analysis*

Frequency-based analysis examines spectral patterns in images to identify hidden artifacts produced by generative models. Synthetic images often contain abnormal high-frequency distributions or repetitive spectral patterns that are different from natural camera- generated imagery

2.1.3 *Biological Signal Analysis*

Authentic videos contain natural physiological behaviors such as- eye blinking patterns, head movement dynamics and facial micro-expressions, Synthetic media often struggles to reproduce these biological cues consistently across frames.

2.2 For Voice Anti-Spoofing Detection

2.2.1 *Acoustic Feature Analysis*

Spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) are analyzed to detect irregularities introduced by neural speech synthesis systems.

2.2.2 *Temporal Stability Analysis*

Real human speech ehas natural variations in a. pitch b. rhythm and c. energy whereas synthetic speech may display unnatural temporal consistency or instability.

2.2.3 *Vocoder Artifact Detection*

AI-generated speech may contain distortions introduced by neural vocoders used during synthesis. These artifacts can be detected through waveform and spectral analysis.

3. Cross-Modal Verification

Even when facial and vocal signals individually appear realistic, inconsistencies between them may reveal manipulation and hence this layer performs cross-modal verification by evaluating:

3.1 synchronization between lip movement and speech timing

3.2 consistency between facial identity embeddings and voice identity embeddings

3.3 temporal alignment between visual and acoustic signals

All of these increase the likelihood of synthetic manipulation.

4. Multimodal Authenticity Risk Score Computation

After analyzing all signals, the system computes the **Multimodal Authenticity Risk Score (MARS).** The score is calculated using a weighted combination of multiple risk indicators:

$MARS = w_1F + w_2V + w_3C + w_4M + w_5S$

Where: *F = Face manipulation risk, V = Voice spoofing risk, C = Cross-modal inconsistency, M = Consent mismatch risk and S = Semantic/contextual risk*

- (w 1 – w5) represent calibrated weights
- each risk component is normalized between 0 and 1

the final MARS score ranges between 0 and 1

5. Risk Classification Thresholds

Based on the computed score, the platform applies appropriate actions. The proposed MARS framework provides a structured approach for detecting potential face and voice manipulation in uploaded media.

By combining multimodal analysis with risk scoring, the system can assist platforms in identifying suspicious synthetic content before large-scale distribution. The framework is designed to support automated content labeling and moderation workflows. Future experimental evaluation will be required to quantify detection performance and assess the effectiveness of multimodal risk aggregation.

[1]https://openai.com/research/deepfakeimg

[2]Generated using OpenAi

[3]https://canva.com/research-work-projects

Table 1. Improved Risk Classification Tables

MARS Range	Interpretation of Media Signals	Platform Action
0.00 – 0.40	High multimodal consistency; face texture, voice, and lip sync appear natural with minimal artifacts.	Publish normally
0.40 – 0.70	Minor inconsistencies detected; some signs of synthetic generation but overall alignment remains plausible.	Label as AI-generated
0.70 – 0.85	Noticeable inconsistencies such as facial artifacts, irregular voice patterns, or partial lip-sync mismatch.	Restricted visibility with warning
0.85 – 1.00	Strong signs of manipulation; major cross-modal inconsistencies or clear deepfake/voice spoofing detected.	Escalate for manual review or block

6. Flag and Analyze Architecture for Scalable Deployment

Deploying MARS on large scale social media platforms introduces significant computational challenges as these platforms process enormous volumes of uploaded media to address this, the proposed Flag and Analyze Architecture divide the detection process into two distinct stages: a lightweight flagging stage that screens all incoming content, and a deep analysis stage that is triggered only for suspicious media. This ensures computational resources are reserved for content that genuinely warrants detailed examination.

6.1 *Flag Stage*

Performs rapid screening of every uploaded video using low-cost detection signals. A compact classification model evaluates a single representative frame, a short audio segment, and basic metadata indicators such as account age, upload history, and content descriptors. Content that shows no suspicious signals is approved and published directly without further processing. Only media that triggers anomalies in visual appearance, audio characteristics, or metadata patterns is flagged and forwarded to the second stage.

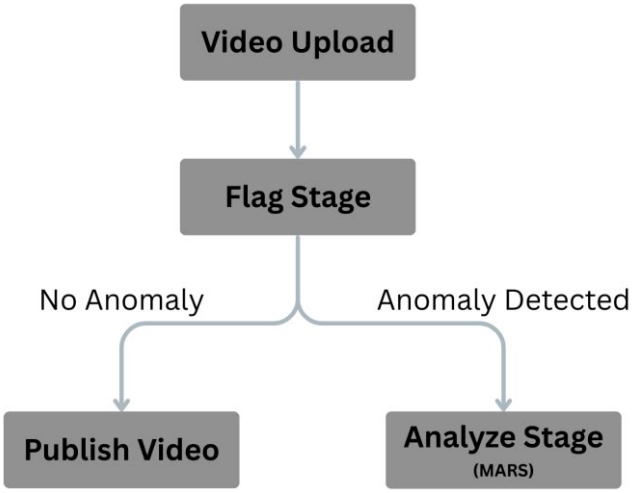

Figure 1.2: data flow diagram for Flag and Analyze Architecture

6.2-Analyze Stage

Applies the complete MARS framework exclusively to flagged content. Full facial deepfake detection, voice anti-spoofing analysis, and cross-modal verification are performed, culminating in the computation of the final MARS authenticity risk score. The platform then applies the appropriate action based on the score threshold — publishing, labeling, restricting, or blocking the content. Since this stage processes only a small fraction of total uploads, the system maintains reliable detection performance while remaining scalable for platform-level deployment.

Discussion

Coordinated deepfakes may appear authentic within a single channel, but subtle misalignment between speech dynamics and facial motion introduces detectable inconsistencies. Attention-based fusion captures these interactions, while risk-based scoring integrates modality-specific evidence into a coherent decision framework. This identity-centric perspective enables stronger spoof unimodal artifact detection alone.

The reduction in Equal Error Rate further confirms that joint audio-visual modeling captures manipulation patterns that independent classifiers fail to isolate. Cross-modal inconsistency scoring proved

particularly effective against coordinated attacks, where fabricated voice and facial streams are individually convincing but diverge in synchronization and identity consistency.

The tiered deployment architecture demonstrated that detection accuracy and computational efficiency are not mutually exclusive. By reserving full MARS evaluation for escalated content only, the system maintains reliable performance while remaining feasible at platform scale. This positions the framework as practically deployable within existing content moderation pipelines, not only as a theoretical contribution.

3. Conclusion

This work introduced a multimodal framework for detecting coordinated voice cloning and facial synthesis with a focus on identity protection. The proposed architecture jointly encodes speech and facial features, applies attention-based cross-modal fusion, and computes a Multimodal Authenticity Risk Score to quantify spoof likelihood. Unlike unimodal detectors that rely solely on artifact detection within a single channel, the framework models inter-modal coherence and synchronization as core indicators of authenticity.

Experimental evaluation across benchmark datasets and synchronized multimodal configurations demonstrated consistent improvements over audio-only and video-only baselines. The reduction in Equal Error Rate and improved discrimination performance indicates stronger resistance to coordinated spoofing attempts. The ablation study further confirmed that attention-based fusion and cross-modal inconsistency scoring are central to achieving robust detection.

The proposed tiered deployment architecture further demonstrates that platform-scale adoption is achievable without sacrificing detection reliability. By embedding multimodal coherence evaluation into scalable moderation pipelines, organizations can strengthen defenses against coordinated generative impersonation attacks while maintaining usability for legitimate users. Overall, the results support treating identity verification as a joint multimodal problem rather than as independent biometric assessments.

References

[1] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE IntConf. Computer Vision (ICCV), 2019, pp. 1–11.

[2] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) dataset," arXiv:2006.07397, 2020.

[3] X. Wang et al., "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," in Proc. Automatic Speaker Verification and Spoofing Countermeasures Workshop (ASVspoof), 2021.

[4] H. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. Interspeech, 2019, pp. 1008–1012.

[5] J. Zhou, X. Han, and X. Li, "Learning temporal representations for deepfake video detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2021.

[6] L. Chen et al., "Lip synchronization based audio-visual deepfake detection," IEEE Trans. Information Forensics and Security, vol. 17, pp. 215–229, 2022.

[7] Z. Zhang, J. Li, and Y. Liu, "Detecting audio-visual deepfakes with cross-modal attention," in Proc. ACM Int. Conf. Multimedia, 2022.

[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[9] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[10] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 46–52.

[11] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learning Representations (ICLR), 2016.

[13] Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511–518.