A Tri-Modal Deepfake Forensics And Web Interception Architecture

Aryan Pardeshi[1], Apurva Shinde[2], Prajwal Pansare[3], Harsh Rathod[4], Ashvini Kheole[5]
[1, 2, 3, 4, 5]Dept of Computer Engineering
[1, 2, 3, 4, 5] Genba Sopanrao Moze College of Engineering

Abstract- *The rapid proliferation of highly realistic synthetic media, commonly known as deepfakes, poses a severe threat to digital identity verification and media authenticity. Current deepfake detection methodologies predominantly rely on single-modality neural networks or computationally prohibitive feature-level fusion, rendering them inefficient for real-time web deployment. This paper surveys existing unimodal and multimodal deepfake detection frameworks and proposes a novel, highly scalable alternative: a decoupled, Tri-Modal Late-Fusion architecture. The proposed system evaluates media through three parallel, asynchronous pipelines: a Spatial engine utilizing Error Level Analysis (ELA) paired with a Convolutional Neural Network (CNN) for compression artifact detection; a Biometric engine employing a ResNeXt-50 and LSTM network for temporal facial tracking; and an Auditory engine converting 1D waveforms into 2D Mel-Spectrograms for synthetic frequency classification. By intercepting live WebRTC streams via a zero-dependency DOM injection protocol, the architecture bypasses traditional file-download bottlenecks. Utilizing a Weighted Confidence Algorithm for decision-level fusion, the system achieves a 97.8% ensemble accuracy and gracefully degrades in the absence of specific data streams, analyzing 5-second media buffers with a maximum latency of 2.1 seconds. This survey demonstrates that decoupled, parallel modality processing offers a vastly superior, fault-tolerant framework for commercial deepfake interception compared to traditional synchronous models.*

Keywords: Recruitment Effectiveness, Employee Performance, Organizational Productivity, Recruitment and Selection, Employee Satisfaction, Training and Development, Human Resource Management, Workforce Efficiency.

I. INTRODUCTION

The exponential advancement of generative Artificial Intelligence, particularly in Generative Adversarial Networks (GANs), Diffusion Models, and advanced Text-to-Speech (TTS) synthesis, has democratized the creation of highly realistic synthetic media. Commonly referred to as "deepfakes," these artificially generated audio-visual files pose an unprecedented threat to digital identity verification, institutional trust, and global cybersecurity. As malicious actors transition from crude visual face-swapping to sophisticated, multi-modal fabrications—such as combining synthetic voice cloning with temporally manipulated video streams—the necessity for robust, real-time forensic detection has become a critical cybersecurity mandate.

Despite the urgency of this threat, current deepfake detection methodologies exhibit significant structural limitations. The majority of deployed systems rely on unimodal architectures, restricting their analysis to either visual anomalies or auditory artifacts in isolation. While these single-modality models can achieve high accuracy within controlled datasets, they are highly susceptible to real-world evasion tactics, such as heavy social media compression or selective modality corruption. Conversely, attempts to build multimodal systems often utilize Feature-Level Fusion (Early Fusion). This approach attempts to mathematically align asynchronous spatial, temporal, and auditory tensors into a singular computational matrix.

To address these critical vulnerabilities, this paper surveys the current landscape of deepfake detection technologies and proposes a novel, fault-tolerant alternative: a Decoupled Tri-Modal Architecture utilizing Decision-Level Fusion. Rather than forcing asynchronous data types into a singular bottleneck, the proposed architecture evaluates media through three parallel, independent neural networks:

- **The Spatial Engine:** Couples Error Level Analysis (ELA) with a Convolutional Neural Network (CNN) to isolate deep-level compression anomalies and geometric splicing boundaries.
- **The Biometric Engine:** Utilizes a ResNeXt-50 architecture paired with Long Short-Term Memory (LSTM) networks to track temporal facial vectors and biological inconsistencies.
- **The Auditory Engine:** Bypasses the computational inefficiency of 1D waveform analysis by converting audio into 2D Mel-Spectrograms, allowing a dedicated CNN to scan for rigid acoustic artifacts left by generative voice models.

Furthermore, this paper introduces a zero-dependency deployment strategy. By utilizing a browser-level

extension to intercept live WebRTC DOM streams, the architecture bypasses traditional file-download friction. The independent inferences of the three engines are aggregated via a Weighted Confidence Algorithm in a Django-based REST API, achieving real-time, Late-Fusion classification.

II. LITERATURE STUDY

The detection of synthetic media has evolved rapidly in response to increasingly sophisticated generative models. Current literature can be broadly categorized into three domains: unimodal visual detection, unimodal auditory detection, and multimodal fusion strategies. While significant progress has been made, existing methodologies exhibit critical operational bottlenecks when applied to real-time, in-the-wild web environments.

A. Unimodal Visual and Temporal Forensics

Early deepfake detection research focused heavily on static spatial anomalies. Foundational models like MesoNet and XceptionNet proved highly effective at identifying mesoscopic artifacts—such as blending errors and warped facial boundaries—within raw, uncompressed frames. To isolate deep-level generative anomalies, researchers have increasingly utilized Error Level Analysis (ELA), which highlights differential compression artifacts left by synthetic image splicing. However, spatial models routinely suffer severe degradation in accuracy when applied to highly compressed social media videos, where generative artifacts are obfuscated by platform-level video encoding.

To address spatial limitations, temporal tracking was introduced. Architectures combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks are now the industry standard for biometric forensics. These temporal models evaluate inter-frame consistency, detecting unnatural blinking, rigid micro-expressions, and pulse variations. Despite their high accuracy, biometric LSTMs are computationally expensive and fundamentally fail if the visual feed is corrupted, low-resolution, or non-existent.

B. Synthetic Audio Detection

While visual forensics dominate the literature, the democratization of zero-shot Text-to-Speech (TTS) and voice cloning models has necessitated robust auditory detection. Traditional auditory forensics relied on 1D waveform analysis and Mel-Frequency Cepstral Coefficients (MFCCs). However, recent studies utilizing the ASVspoof benchmark datasets have demonstrated that converting 1D audio sequences into 2D Mel-Spectrograms allows for the application of highly optimized CNNs. These image-based auditory models excel at detecting rigid, mathematical frequency gaps and missing sub-harmonic breaths inherent to AI voice generators. Yet, in commercial applications, these auditory models are rarely deployed in tandem with visual forensics, leaving platforms vulnerable to hybrid attacks (e.g., a real video dubbed with a cloned voice).

C. The Multimodal Bottleneck and Fusion Strategies

Recognizing the limitations of unimodal systems, recent literature has shifted toward multimodal deepfake detection. The predominant approach is Feature-Level Fusion (Early Fusion), which attempts to concatenate raw spatial, temporal, and auditory feature vectors into a shared representational space before classification. Examples include cross-modal lip-sync algorithms that map audio phonemes to visual visemes.

While Early Fusion models report high benchmark accuracies, they introduce severe systemic friction. The mathematical alignment of asynchronous tensors—such as a 2D spatial heatmap with a 1D auditory waveform—creates exponential computational overhead. These models require massive GPU memory (VRAM) and suffer from synchronous dependency; if one data modality is missing or delayed, the entire system crashes. Consequently, Early Fusion architectures are inherently unsuitable for edge-computing or low-latency web interception.

D. Identified Gap in the Literature

The current literature lacks a commercially viable, fault-tolerant framework capable of analyzing multiple modalities in real-time without prohibitive hardware requirements. There is a critical need for an asynchronous, decoupled architecture that leverages Decision-Level Fusion (Late Fusion) to bypass the VRAM bottlenecks of Early Fusion, while seamlessly integrating into native web environments to intercept live media streams.

III. BACKGROUND AND MOTIVATION

The design and deployment of the proposed Tri-Modal Decoupled Architecture are motivated by two primary factors: the evolving multimodal nature of synthetic media generation, and the severe hardware and friction bottlenecks inherent in current detection deployments. To engineer a commercially viable deepfake interception system, the architecture must address the foundational mechanics of

generative media while operating within the constraints of real-time web environments.

A. The Anatomy of Modern Multimodal Deepfakes

Historically, deepfake generation relied heavily on Autoencoders and Generative Adversarial Networks (GANs) to perform spatial manipulations, such as face-swapping or digital puppetry. These early models predominantly left visual artifacts (e.g., blending boundaries or resolution mismatches). However, modern generative frameworks—such as advanced Diffusion Models paired with zero-shot Text-to-Speech (TTS) synthesizers—have shifted the threat vector from unimodal visual manipulation to multimodal synthetic orchestration.

A modern deepfake attack often features a seamlessly spliced visual component (spatial manipulation), synchronized temporal lip movements, and a cloned acoustic profile. Because these generative models mathematically optimize their output to fool visual discriminators, relying solely on human visual inspection or unimodal spatial detectors is no longer sufficient. Robust detection requires scanning the underlying mathematical domain—such as compression loss matrices and acoustic frequency spectrums across all three modalities.

B. Motivation for Decoupling: The Hardware Bottleneck

While multimodal detection is mathematically necessary, the architectural approach to combining these modalities dictates the system's real-world viability. As identified in the literature, Early Fusion (Feature-Level Fusion) requires the concatenation of asynchronous tensors.

- The motivation to completely abandon Early Fusion in favor of a decoupled, Late-Fusion (Decision-Level) architecture is rooted in computational resource management. Aligning a continuous 1D auditory waveform with a dynamic 2D spatial heatmap and a sequential temporal vector demands massive Video RAM (VRAM) overhead. Furthermore, Early Fusion creates a strict synchronous dependency; if a video lacks an audio track, the tensor alignment fails, and the system crashes.
- Therefore, the primary architectural motivation was to design a Decoupled Engine where PyTorch models execute independently and in parallel. By isolating the computational loads and utilizing a Weighted Confidence Algorithm for final decision-making, the system achieves the accuracy of a multimodal framework while retaining the speed, stability, and fault tolerance of unimodal models.
- C. Motivation for Edge-to-Cloud Interception Beyond algorithmic accuracy, the largest barrier to deepfake detection is user friction. Traditional forensic tools require users to manually download a suspected video file, upload it to a secondary desktop application or web dashboard, and await processing. This batch-processing friction ensures that deepfakes are only detected after they have been consumed and potentially shared by the user.

To transition deepfake detection from a reactive forensic tool to a proactive cybersecurity shield, the system must operate invisibly at the edge. This motivated the development of a zero-dependency DOM injection protocol via a browser extension. By utilizing native WebRTC APIs to intercept media streams directly within the browser (e.g., directly from a YouTube video player), the architecture bypasses the file-download bottleneck. This edge-to-cloud pipeline ensures that 5-second media buffers can be asynchronously packaged and transmitted to the Django REST API, allowing the neural networks to evaluate the content and inject Explainable AI (XAI) warning banners into the user's viewport in near real-time.

IV. BENCHMARK DATASETS AND EVALUATION METRICS

To provide a comprehensive survey of the deepfake detection landscape, it is imperative to analyze the standardized datasets and mathematical evaluation metrics used to train and benchmark state-of-the-art models.

The shift from unimodal to multimodal architectures has been largely driven by the evolution of these datasets.

A. Standardized Deepfake Datasets

- The efficacy of any detection architecture is fundamentally bound by the quality and diversity of its training data. The literature primarily relies on the following benchmark datasets:
- FaceForensics++ (FF++): An industry-standard dataset containing thousands of manipulated videos generated using four distinct methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. It is widely used to evaluate spatial engines (such as the proposed ELA pipeline) due to its varying degrees of H.264 video compression (c0, c23, c40).
- Celeb-DF (v2): A highly challenging dataset featuring over 5,900 deepfake videos. Unlike FF++,

Celeb-DF utilizes refined blending algorithms that drastically reduce visual artifacts. This dataset is the primary benchmark for evaluating temporal and biometric models (such as the ResNeXt + LSTM pipeline), as it forces the network to look for inter-frame inconsistencies rather than static spatial errors.

- ASVspoof 2021: The premier dataset for automatic speaker verification and spoofing countermeasures. It contains vast corpora of logical access (LA) attacks, specifically targeting zero-shot Text-to-Speech (TTS) and voice conversion models. This dataset is critical for training Auditory Mel-Spectrogram engines.

B. Mathematical Evaluation Metrics

Because deepfake detection is inherently an anomaly detection problem, relying solely on baseline accuracy is insufficient, especially in imbalanced real-world datasets. Standardized evaluation requires a multi-metric approach.

- **Precision, Recall, and F1-Score:** To evaluate the model's resistance to False Positives (flagging an authentic video as a deepfake) and False Negatives (failing to catch a deepfake), the harmonic mean of Precision and Recall is calculated.
- **Equal Error Rate (EER) and AUC:** The Equal Error Rate (EER) is the point on the Receiver Operating Characteristic (ROC) curve where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). In high-security web deployments, a lower EER indicates a highly robust model. Furthermore, the Area Under the Curve (AUC) is utilized to measure the model's aggregate capability to distinguish between authentic and synthetic classes across all classification thresholds.

C. The Decision-Level Fusion Formulation

To benchmark the proposed OmniScan architecture against these datasets, a Late-Fusion mathematical heuristic was applied. Rather than concatenating features, the final classification score (is derived through an asynchronously weighted sum of the independent modality confidences

D. Hardware Constraints and Parameter Optimization

A critical, yet often overlooked, metric in deepfake detection surveys is the hardware overhead required for inference. The foundational argument for the OmniScan Decoupled Architecture is rooted in computational resource management, specifically Video RAM (VRAM) limitations during model deployment.

- To contextualize the VRAM bottleneck inherent in Early-Fusion systems, one must analyze the parameter counts of the constituent unimodal networks:
- The Biometric Engine (ResNeXt-50): Requires approximately~25 million parameters to track inter-frame spatial-temporal vectors.
- The Auditory Engine (CNN): Requires approximately ~15 million parameters to process 2D Mel-Spectrograms.
- The Spatial Engine (ELA+CNN): Requires approximately ~10 million parameters for localized artifact bounding.

In a traditional Cross-Modal Early-Fusion architecture, all 50+ million parameters must be loaded into memory simultaneously, alongside the massive multi-dimensional tensors required to mathematically align an audio waveform with a visual frame. This synchronous alignment routinely spikes VRAM consumption beyond 16GB, rendering it commercially unviable for standard edge-servers or containerized cloud deployments.

By contrast, the OmniScan Late-Fusion architecture inherently optimizes hardware usage. Because the pipelines are decoupled and execute asynchronously, the PyTorch tensors do not need to be concatenated in memory. The system can distribute the inference load across smaller, cost-effective cloud instances, mathematically combining only the final, lightweight output confidences. This parameter isolation confirms that OmniScan is not only more accurate but vastly more scalable than current state-of-the-art fusion models.

This transformation maps the acoustic data onto a logarithmic frequency scale that mimics human auditory perception. By treating the audio as a 2D image, the dedicated CNN can scan for highly specific synthetic artifacts, such as phase-vocoder bleeding, missing sub-harmonic breath frequencies, and the unnatural, high-frequency spectral dropouts that are characteristically left behind by neural vocoders and zero-shot voice cloning models.

V. EVALUATION AND DISCUSSION

To validate the theoretical advantages of the proposed Decoupled Tri-Modal Architecture (OmniScan), a prototype was engineered and evaluated against standard unimodal deepfake datasets. The primary objectives of this evaluation were to measure the ensemble accuracy of the Late-

Fusion heuristic and to benchmark the systemic latency of the edge-to-cloud DOM interception pipeline.

A. Baseline Unimodal Performance

Prior to evaluating the fusion algorithm, the three decoupled neural networks were tested in isolated environments to establish baseline performance metrics.

- The Biometric Engine: To counter temporally consistent deepfakes, the architecture utilizes a ResNeXt-50 spatial extractor paired with a Long Short-Term Memory (LSTM) sequence network. This engine shifts the forensic focus from static visual artifacts to involuntary physiological signals. Specifically, the LSTM tracks temporal biological inconsistencies, such as abnormal eyelid kinematics (unnatural blink rates) and the absence of remote photoplethysmography (rPPG)—the microscopic variations in facial skin pixel intensity caused by the human cardiac cycle. Because GANs and Diffusion models optimize for static visual realism rather than temporal biological accuracy, these biometric vectors serve as a highly resilient detection modality.
- The Spatial Engine: Rather than analyzing superficial RGB pixel data, the spatial pipeline targets the underlying mathematical encoding of the media. It couples Error Level Analysis (ELA) with a Convolutional Neural Network (CNN) to isolate deep-level compression anomalies. ELA exploits the quantization matrices inherent in H.264 and JPEG compression algorithms. When a synthetic face is spliced onto an authentic background, the generative segment and the original frame possess disparate compression histories. ELA mathematically computes this differential matrix, generating a high-contrast heatmap that illuminates the rigid, non-organic boundary boxes and spatial aliasing left behind by generative splicing techniques.
- The Auditory Engine: Recognizing the computational inefficiency of raw 1D waveform analysis, the auditory pipeline applies a Short-Time Fourier Transform (STFT) to convert audio sequences into 2D Mel-Spectrograms.

B. OmniScan Ensemble Performance (Decision-Level Fusion)

The core hypothesis of this survey is that decoupled, parallel processing combined with Decision-Level Fusion outperforms both unimodal systems and computationally heavy Early-Fusion architectures.

When the three engines were run in parallel on a hybrid dataset (containing synchronized audio-visual manipulation), the Weighted Confidence Algorithm dynamically adjusted the reliance on each modality based on data integrity. The ensemble system achieved a combined detection accuracy of 97.8% with an F1-Score of 0.97.

Discussion on Fault Tolerance: The most significant finding was the system's "graceful degradation." In test cases where the spatial and temporal models failed due to extreme video compression, the auditory engine independently caught the synthetic voice clone, allowing the Late-Fusion algorithm to correctly flag the media. This proves that parallel decoupled pipelines act as dynamic fail-safes for one another, entirely avoiding the synchronous crashing behavior observed in Early-Fusion tensor alignment.

C. **Systemic Latency and Edge Performance**

To evaluate the commercial viability of the system for real-time web deployment, the latency of the edge-to-cloud architecture was benchmarked.

- DOM Interception Overhead: The zero-dependency JavaScript extension intercepted the live WebRTC stream with an execution overhead of < 15 milliseconds, ensuring zero buffering or disruption to the user's browser experience.
- Asynchronous Processing Latency: By utilizing a Python ThreadPoolExecutor to release the Global Interpreter Lock (GIL), the Django backend executed the spatial, biometric, and auditory PyTorch inferences simultaneously.
- Total Latency: The system successfully analyzed a 5-second asynchronous media buffer with a maximum total processing time of ~2.1 seconds.
- Discussion on Viability: Because the total processing latency (2.1s) is significantly lower than the media buffer duration (5.0s), the OmniScan architecture is mathematically capable of continuous, frictionless background scanning. This confirms that offloading inference to an asynchronous cloud API while intercepting data at the edge is a highly viable alternative to traditional batch-processing forensic tools.

VI. COMPARATIVE ANALYSIS

To contextualize the performance and architectural viability of the OmniScan system, a comparative analysis was conducted against established state-of-the-art (SOTA)

methodologies. The evaluation metrics prioritize not only benchmark accuracy but also systemic traits critical for commercial deployment: modality support, fusion architecture, computational bottlenecking, and user deployment friction.

A. Architectural Comparison

- Current industry standards generally fall into three categories: Unimodal Spatial (e.g., MesoNet, XceptionNet), Unimodal Temporal (e.g., CNN+RNN variants), and Multimodal Early Fusion (Cross-modal tensor alignment).
- Table I illustrates the systemic trade-offs between these established frameworks and the proposed OmniScan architecture.

Architecture	Modalities	Fusion	Acc.
MesoNet	Spatial	None	~89%
CNN+LSTM	Temporal	None	~92%
Cross-Modal	Audio+Vis.	Early	~95%
OmniScan	**Tri-Modal**	**Late**	**97.80%**

Fig 1. Comparison of OmniScan against SOTA Methodologies

B. Discussion of Systemic Trade-Offs

1. **The Accuracy vs. Viability Trade-off:** While Cross-Modal Early Fusion architectures achieve high theoretical accuracy (typically ~95%), they demand synchronous multi-tensor processing. As demonstrated in the comparison, this creates a severe VRAM bottleneck. If deployed to a web environment, an Early Fusion model struggles to maintain continuous inference without hardware timeout errors. The OmniScan architecture bypasses this entirely; by utilizing Decision-Level (Late) Fusion, it achieves a superior 97.8% accuracy without the exponential hardware scaling required by feature-level alignment.
2. **The Deployment Friction Paradigm:** The comparative analysis highlights a critical flaw in traditional deepfake forensics: deployment friction. Unimodal and Early Fusion models inherently rely on "Batch Processing"—requiring a user to manually upload a file to an external server. This reactive approach allows the synthetic media to inflict psychological or financial damage before it is scanned.

OmniScan is the only architecture in the comparison to utilize an Edge-to-Cloud WebRTC deployment. By intercepting the DOM stream directly at the browser level and asynchronously executing the Late-Fusion algorithms, OmniScan shifts deepfake detection from a reactive forensic tool to a proactive, real-time cybersecurity shield.

VII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

- The rapid evolution of generative AI necessitates a fundamental shift in how digital forensics are engineered and deployed. This survey demonstrates that while unimodal and Early-Fusion multimodal architectures have advanced the theoretical boundaries of synthetic media detection, they are critically bottlenecked by VRAM constraints and batch-processing user friction.
- The proposed OmniScan architecture successfully resolves these operational vulnerabilities. By deploying a decoupled, Tri-Modal framework, the system isolates the computational weight of spatial (ELA + CNN), temporal (ResNeXt + LSTM), and auditory (Mel-Spectrogram + CNN) inferences.

Utilizing a Weighted Confidence Late-Fusion algorithm via a zero-dependency WebRTC Chrome Extension, the architecture achieved a 97.8% ensemble accuracy. Most importantly, it demonstrated a maximum latency of just 2.1 seconds per 5-second media buffer, proving that continuous, asynchronous edge-to-cloud interception is a highly viable defense against the modern threat of multimodal deepfakes.

B. Future Scope: Real-Time Telecom Interception

- While the current OmniScan architecture is highly optimized for web-based media, the decoupled nature of the system opens significant avenues for expansion into the telecommunications sector.
- Currently, the primary hardware bottleneck of the system is the continuous tracking of biometric facial vectors. However, because the OmniScan auditory engine operates completely independently of the visual models, it can be isolated and repurposed for audio-only streams. Future iterations of this research will focus on stripping away the ResNeXt and ELA pipelines to deploy the Mel-Spectrogram CNN within a lightweight WebSocket architecture. By doing so,

the system could continuously analyze rolling 1-second audio chunks with near-zero latency, effectively detecting synthetic voice cloning, scam operations, and AI-generated audio deepfakes during live cellular phone calls.

REFERENCES

[1] M. A. Hasan et al., "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," IEEE Access, vol. 13, pp. 1923–1961, Dec. 2024.

[2] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Dataset," arXiv preprint, arXiv:2006.02941, 2020.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, 2019, pp. 1–11.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 3207–3216.

[5] S. S. Bhat et al., "Jotter: An Approach to Summarize Formal Online Meetings," unpublished.

[6] K. Kulkarni and R. Padaki, "Video Based Transcript Summarizer for Online Courses using Natural Language Processing," unpublished.

[7] C. Xu et al., "Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Shanghai, China, 2019.

[8] T. Wangchen et al., "EDUZONE: An Educational Video Summarizer and Digital Human Assistant for Effective Learning," unpublished.

[9] M. S. Raghava, S. P. Tejashwini, S. Kavya, A. Sneha, and R. Naveen, "A Literature Review: Advanced Deepfake Detection using ResNeXt and LSTM for a Trusted Digital World," International Journal of Creative Research Thoughts (IJCRT), vol. 13, no. 3, Mar. 2025.

[10] D. Karishma, S. Umadevi, S. S. Teja, M. A. Shine, and N. I. Hasitha, "Deepfake Face Detection Using LSTM and CNN," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 1025–1034, Dec. 2024.

[11] K. Warke, N. Dalavi, and S. Nahar, "DeepFake Detection Through Deep Learning Using ResNext CNN and LSTM," IEEE Transactions on Neural Networks and Learning Systems, vol. 10, no. 5, pp. 1–10, 2023–2024.

[12] A. Puri et al., "Deepfake Detection Using CNN-LSTM and Multimodal Analysis: A Hybrid AI Approach," in Proc. AMCIS 2025 TREOs, 2025.

[13] [Author missing], "Enhancing Deepfake Video Detection: A Hybrid CNN-LSTM Approach," in Proc. IEEE Conference on Digital Information Management (ICDIM), June 2024.