SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V13I8P101

Volume 13 Issue 8, 1-7, August 2026

Original Article

Preventing the Impact of AI Deepfakes on Journalistic Integrity: A Hybrid Approach to Detection and Prevention

Anish Chelliserikatil Pavithran

Senior Software Engineer, The Associated Press, USA.

Corresponding Author : anishcp@gmail.com

Received: 17 June 2026 Revised: 25 July 2026 Accepted: 12 August 2026 Published: 30 August 2026

Abstract - *Deepfakes pose a direct challenge to journalistic verification because convincing synthetic video and audio can circulate faster than manual fact-checking, while many high-performing detectors remain costly or lose reliability under compression and domain shift. This study presents a newsroom-oriented hybrid framework that combines lightweight visual screening, temporal consistency analysis, selective frequency-domain verification, optional audio analysis, content-provenance checks, and human editorial review. MobileNetV3 provides low-cost frame triage, a shallow GRU/LSTM models short-term temporal inconsistencies, and the frequency-aware stage is reserved for ambiguous content to limit computational overhead. Available branch scores are combined through late fusion, and uncertain or high-risk items are escalated to editors. The reported model comparison yields 91.8% accuracy, 89.4% precision, 90.5% recall, and an 89.9% F1-score for the MobileNet-LSTM configuration, exceeding the three internal baselines reported in this study. The contribution is not a claim of superior benchmark accuracy over all state-of-the-art detectors; rather, it is an operational design that integrates efficient detection, provenance evidence, multimodal checks, and editorial judgment for resource-constrained newsroom workflows.*

Keywords - *Content Provenance, Deepfake Detection, Hybrid Human-Ai Systems, Journalistic Integrity, MobileNetV3.*

1. Introduction

Deepfakes are synthetic or manipulated images, audio, and video produced with artificial-intelligence techniques such as generative adversarial networks and diffusion models—improvements in.

The generative quality has made manipulated media increasingly difficult to distinguish from authentic content, creating a verification problem for journalism and other trust-sensitive domains [1].

The journalistic risk is not limited to detector accuracy. Deepfakes can exploit the speed of social-media distribution, create uncertainty during high-impact events, and weaken public confidence in authentic reporting. Survey evidence shows substantial public concern about the effect of deepfakes on truth and media trust [3], while recent reporting and international guidance have highlighted synthetic-media risks in geopolitical and health-information settings [4]-[6]. Existing research demonstrates strong benchmark performance from CNN, recurrent, transformer, and multimodal detectors, but the literature also identifies persistent weaknesses: cross-dataset degradation, sensitivity to compression and post-processing, high computational cost, and limited treatment of provenance and editorial context [9], [16], [19]-[21]. These limitations create a research gap between laboratory-oriented detection and newsroom verification, where media arrives from heterogeneous sources, time is limited, and publication decisions require both technical evidence and contextual judgment.

The novelty of the proposed framework lies in the integration of four operational elements that are usually studied separately: lightweight MobileNetV3 triage, selective frequency-aware verification, shallow GRU/LSTM temporal modeling, and provenance-aware human escalation. In contrast with heavier transformer-centric systems that emphasize peak benchmark accuracy [10]-[15], the proposed design prioritizes staged computation and editorial accountability. The study, therefore, evaluates whether a resource-conscious spatial-temporal detector can improve over simple internal baselines while providing a practical path for provenance checks, optional audio analysis, and human review.

The remainder of the paper reviews current detection research, summarizes common limitations, defines the hybrid workflow and its decision logic, reports the available

comparative results, and discusses dataset, deployment, and validation constraints that must be addressed before production use.

2. Literature Review

Recent advances in deepfake detection reveal a more complicated scenario where methods are getting progressively complex to address realistic synthetic media. There are several comprehensive surveys that have been published, which provided summary of the state-of-the-art in detection technologies and grouped them by modality, methodology, and performance. For example, Sunil et al. present a review of automatic deep fake detection mechanisms in image, video and audio streams, introducing progress on slightly less influenced categorization models to distinguish faked videos with human from authentic ones [7].

Singh (2025) carries out an integrative review of deepfake generation and detection and underlines the challenges that current biometric systems have to cope with, dynamic, multimodal deepfake content, but also the growing importance of interdisciplinary approaches such as explainable AI and federated learning [8]. Alrashoud (2025) also reviews recent works on deepfake video detection approaches, including the application of visual, audio and multimodal ones, while emphasizing the unresolved issues such as generalization and real-world implementation [9].

Recent specific technical studies showed that the agile detection models are evolving. Soudy (2024) proposed an end-to-end deep learning system that integrates Convolutional Neural Network (CNN) and Vision Transformer (ViT) in boosting spatial and contextual analysis, hence detecting robustness when compared to a traditional CNN-only-based method [10]. Hussien (2024) obtains a similar level of accuracy, ~98%, for spotting deepfakes by using Vision Transformer architectures tuned to extract face features [11]. Petmezas (2025) presented a hybrid model by incorporating 3D Morphable Models with CNN-LSTM and transformer structures to capture the temporal and spatial inconsistencies occurring in manipulated videos and showed impressive improvements on the benchmark [12].

Multimodal approaches to detection are also becoming popular. Thakur et al. (2025) propose a transformer-based LLM framework, which is able to fuse visual, auditory and textual features at the same time for detection performance boosting on various types of deepfakes, demonstrating that integrating multiple modalities can enhance the resistance against complicated manipulation [13]. Recent transformer-driven survey works, including Liu et al. (2024) and Zhang et al. (2024), review deepfake detection based on ViT architectures and hybrid transformer models that can capture global context dependencies that traditional CNNs may overlook [14], [15].

Systematic reviews converge on three unresolved issues. First, detectors often lose accuracy under cross-dataset testing, compression, or adversarial manipulation [16], [19], [21]. Second, many studies optimize a single modality even though real misinformation can combine visual, audio, and textual cues [13], [14], [17]. Third, operational factors such as compute cost, provenance evidence, and the role of human decision-makers receive less attention than benchmark accuracy [20]. These findings motivate a hybrid design in which machine outputs are treated as evidence within a broader verification workflow rather than as an autonomous publication decision.

Compared with prior work, the proposed study occupies a different point in the accuracy-efficiency tradeoff. Vision-transformer and hybrid CNN-transformer systems report approximately 97-98% benchmark accuracy in the cited literature [10]-[12], and attention-based models in Table 1 reach 98.5% on FaceForensics++. The proposed MobileNet-LSTM result of 91.8% is therefore not presented as a new state-of-the-art benchmark. Its contribution is the integration of lower-cost spatial-temporal screening with selective verification, provenance checks, optional audio analysis, and human escalation for newsroom use. This distinction clarifies the novelty claim and avoids equating deployment practicality with maximum benchmark accuracy.

3. Detection Methods and Limitations

The landscape of deepfake detection has undergone major advances in the last few years with a wide diversity of machine learning and deep learning techniques, but due to fundamental limitations, significant challenges remain. Most of the detection-related work is on deep neural networks to learn visual and temporal distortions in manipulated media.

Fig. 1 Deepfake dataset visualization

Table 1. Deepfake Detection Model Performance

Detection Model	Dataset	Accuracy
Hybrid CNN-RNN (PSO-tuned)	Celeb-DF	97.26%
Hybrid CNN-RNN (PSO-tuned)	DFDC	94.2%
CNN + Vision Transformer	FaceForensics++	~97%
Attention-based Visual Model	FaceForensics++	98.5%
Attention-based Visual Model	FaceForensics++ + DFDC Combined	92.33%
Xception Network	DFDC	89.2%
RNN-based Model	DFDC	77.5%
CNN + Transformer (Audio)	ASVspoof 2019 (audio deepfake)	92%

3.1. Deep Learning-Based Approaches

Convolutional Neural Networks (CNNs) have played a pivotal role in the field of deepfake detection owing to their capability in capturing spatial features from images and frames. Hybrid models that include CNNs along with recurrent ones, such as LSTM or CNN-RNN architectures, target learning of both spatial and temporal artifacts in manipulated videos, which is usually high on standard benchmarks. For instance, Al-Adwan et al. (2024) introduced a CNN-RNN hybridized detector optimized with Particle Swarm Optimization (PSO) based fine-tuning and showed that the obtained models outperform simple CNNs [18].

3.2. Transformer and Vision Transformer Models

Modern methods rely on Vision Transformers (ViTs) and hybrid CNN-ViT architectures, which help to better capture global context across video frames. Hussien (2024) demonstrated that ViT-driven models can reach high detection performance over ~98% for face swap manipulations on benchmark data by detecting fine-grained texture patterns [11]. There are still hybrid models that combine CNN feature extractors and transformer encoders, suitable for long-range temporal dependencies, which could also improve robustness, as in the CNN-LSTM-Transformer model proposed by Petmezas et al. (2025) that combines spatial and sequence analysis to improve the deepfake detection [12].

3.3. Generalization Problems

Despite good performance on preprocessed data, most models face significant problems of generalization. Systematic studies show that detectors trained using academic benchmarks lose accuracy on new or 'in the wild' deepfakes, as a result of domain shift, compression artifact, and adversarial manipulation. These problems are described by Alrashoud [9] as a fundamental challenge in present-day detection research. Furthermore, benchmarks including Deepfake-Eval-2024 and other real-world evaluations observe significant drops in performance for open-source detectors, where the accuracy of models diminishes when they are tested against deepfakes generated from a wide variety of social media sources.

3.4. Limitations of Current Methods

3.4.1. Training Data Dependency

One problem with existing deepfake detection techniques is that they perform well on the datasets they were trained on in artificial conditions. However, they do not generalize well on new or unseen data distributions or novel real-world distortions such as video compression, noise and low-resolution [19].

3.4.2. Computation Cost

There are many deep learning models that can achieve high accuracy, but their architecture is complex and requires a huge amount of computational power. This renders them challenging to deploy in real-world scenarios, at edge devices with a constrained-resource context [20].

3.4.3. Susceptibility to Adversarial and Post-Processing

Methods that compress or resize video, apply other simple edits, etc., may cause a noticeable drop in detection performance, highlighting a discrepancy between research settings and real-world deployability [21].

These limitations demonstrate that though state-of-the-art deep learning models, i.e., CNNs, RNNs, and transformers, are effective in controlled environments, their dependence on targeted training data and their complexity can result in a lack of robustness in the uncertain and unpredictable settings encountered by journalism or social media. This underscores the importance of hybrid detection approaches and adaptive systems that jointly use different techniques with the help of human supervision.

4. Proposed Methodology

The proposed system uses a MobileNet-LSTM hybrid detector as the central low-cost component of a broader newsroom verification pipeline. The design combines lightweight deep learning, selective robust verification, provenance inspection, optional audio analysis, and human editorial review. The purpose of the staged architecture is to reserve additional computation and editorial effort for content that cannot be resolved confidently by early checks.

4.1. System Overview

The proposed framework follows a multi-stage pipeline consisting of (i) media ingestion and provenance checking, (ii) low-cost automated triage, (iii) selective robust verification, and (iv) human-in-the-loop decision making. This staged design ensures that computationally expensive operations are

applied only when necessary, while high-risk or ambiguous content receives additional scrutiny.

4.2. Media Ingestion and Pre-Processing

Input media may include video, audio, and images received from social-media platforms, user submissions, or external news sources. For video, faces are detected, and 8-16 frames per clip are sampled uniformly to reduce redundancy while preserving temporal coverage. Detected faces are cropped and resized to 224 x 224 pixels. When speech is available, audio is segmented into short windows and represented as log-mel spectrograms. These preprocessing choices define the reproducible parameters stated in the source manuscript; the exact training optimizer, learning rate, batch size, train/validation/test split, and random seed were not preserved in the available experiment description and therefore are not fabricated here.

4.3. Dataset Scope, Diversity, and Bias

The manuscript references FaceForensics++, DFDC, Celeb-DF, and ASVspoof 2019 because these benchmarks cover manipulated faces, video sequences, and synthetic speech. However, the currently available experiment record does not specify the exact composition of the subset used for the proposed MobileNet-LSTM result, nor does it document a newsroom-specific corpus.

Consequently, demographic balance, manipulation-type balance, source-platform balance, and class balance cannot be quantified from the source material. This is a limitation of the current evaluation record and motivates future reporting of per-class counts, demographic attributes where ethically appropriate, compression levels, manipulation families, and cross-domain test splits.

Domain shift is particularly relevant to journalism because broadcast clips, messaging-app videos, re-encoded social-media uploads, and user-generated footage can differ substantially from curated benchmark data. The proposed workflow, therefore, treats benchmark performance as preliminary evidence and uses selective frequency analysis, provenance checks, and human review as safeguards against overconfidence when content falls outside the training distribution.

4.4. Provenance and Content Authentication Gate

Before model inference, the workflow checks available metadata and content credentials, including cryptographic signatures or capture claims when present. Provenance is treated as an evidence gate rather than as proof of truth: consistent credentials lower the risk score, missing credentials do not by themselves imply manipulation, and conflicting credentials increase the need for model-based and editorial verification. This ordering reduces unnecessary inference on content with strong origin evidence while preserving escalation when provenance is absent or inconsistent.

1. Media ingestion + provenance check
2. MobileNetV3 visual triage ↓
3. Selective frequency verification ↓
4. GRU/LSTM temporal analysis ↓
5. Optional audio analysis ↓
6. Late-fusion risk score ↓
7. Human editorial review/decision

Fig. 2 Editable block diagram of the proposed newsroom verification pipeline

Figure 2 shows the component integration. Provenance evidence is evaluated first; visual triage then supplies a low-cost manipulation score. Ambiguous items can trigger frequency-aware verification and temporal analysis, while speech content can additionally enter the audio branch.

The available branch scores are fused into a single risk score, and the final operational action is determined by editorial escalation rules rather than by a model output alone.

4.5. Stage-1 Automated Triage Detector (Lightweight Visual Analysis)

In order to perform scalable screening, the first detection stage utilizes a MobileNetV3-based lightweight CNN as it has a good tradeoff between accuracy and computation. The model works at the frame level and outputs a manipulation score p_v in [0,1] to indicate each manipulated frame. Given a video clip with N sampled frames, the video-level score is calculated as:

$$p\_v(video) = (1/N)\, \Sigma\_(i = 1)\text{^}N\, p\_v(i) \qquad (1)$$

This stage is CPU-optimized inference for deployment in resource-limited newsrooms and fast processing of terabytes of user-generated content.

4.6. Stage-2 Robust Verification (Frequency-Aware Analysis)

Content flagged by the triage detector is submitted to a selective frequency-aware verification stage. The design premise is that manipulation traces in transform or frequency representations can remain informative when pixel-space cues are weakened by resizing or recompression. To control cost, the stage is applied only to 4-8 representative keyframes per video. Its output is a probability p_f that contributes to the late-fusion score.

The source manuscript does not preserve the exact transform, frequency-band selection, or classifier hyperparameters; those implementation details should be recorded in the next experimental run before claims of reproducibility are made.

4.7. Lightweight Temporal Consistency Modeling

To capture short-term temporal irregularities without the cost of a 3D CNN, frame embeddings from the MobileNetV3 backbone are passed to a shallow single-layer GRU or LSTM. The recurrent unit models change across sampled frames, including inconsistent facial motion or blinking patterns.

This branch complements frame-level spatial evidence because it can detect sequence inconsistencies that are not visible in an isolated image. The source record specifies the shallow recurrent structure but does not provide hidden-state width, dropout, optimizer, or training schedule, so those values are not inferred in this revision.

4.8. Optional Audio Deepfake Detection Branch

An optional audio analysis step is enabled for speech-based deepfake analysis. Log-mel spectrograms are analyzed by a lightweight convolutional neural network to identify synthesized or tampered speech. The produced probability p_a completes the visual examination to provide multimodal detection, if necessary.

4.9. Decision Fusion and Risk Scoring

Outputs from the available branches are combined using late fusion:

$$p\_final = w\_1\, p\_v + w\_2\, p\_f + w\_3\, p\_a \quad (2)$$

where p_v, p_f, and p_a denote the visual triage, frequency-aware verification, and audio scores. The nonnegative weights satisfy w1 + w2 + w3 = 1 and are normalized over the branches that are available for a given item. Because the source experiment does not report learned or fixed numerical weights, the manuscript does not assign unsupported values. Operationally, low-risk items with consistent provenance may be cleared for routine verification, clearly high-risk items are escalated, and intermediate or conflicting cases are sent to human review. Newsroom-specific score thresholds should be calibrated on validation data to reflect the relative cost of false positives and false negatives.

4.10. Human-in-the-Loop Editorial Review

Human review is the final safeguard for high-risk, ambiguous, or evidentially conflicting media. Editors can compare the automated scores with source history, provenance claims, contextual facts, and independent reporting before publication. This step is designed to reduce the risk of treating a detector score as a definitive truth judgment.

Human review also introduces a scalability constraint: during high-volume events, large numbers of escalations can create editorial bottlenecks. Threshold calibration, queue prioritization by risk, and clear escalation policies are therefore required for production deployment.

5. Results and Discussion

5.1. Experimental Results

The available experiment summary compares the proposed MobileNet-LSTM configuration with three internal baselines: standalone MobileNet, standalone LSTM, and a CNN baseline. Accuracy, precision, recall, and F1-score are reported in Table 2. The source record does not identify the exact dataset split used for this comparison, so the results are interpreted as an internal model comparison rather than as a cross-benchmark state-of-the-art claim.

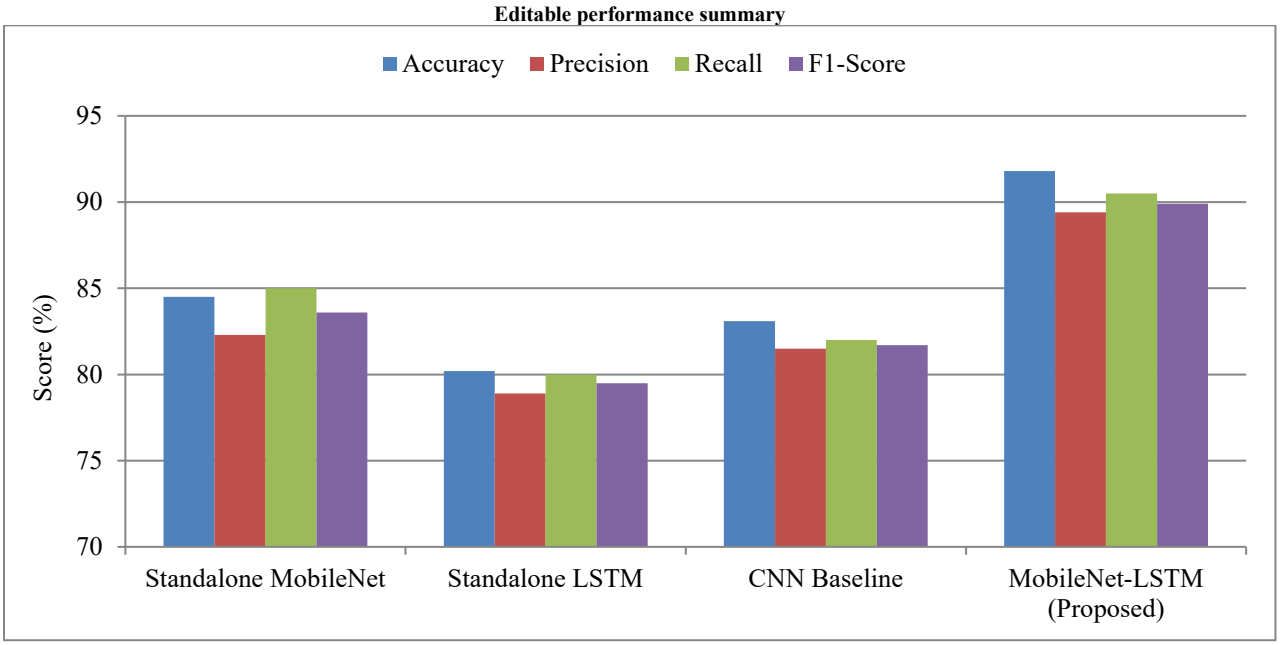

Fig. 3 Performance comparison of the different baseline deepfake detection models with the proposed model

Table 2. Performance Comparison of Deepfake Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Standalone MobileNet	84.5	82.3	85.0	83.6
Standalone LSTM	80.2	78.9	80.0	79.4
CNN Baseline	83.1	81.5	82.0	81.7
MobileNet-LSTM (Proposed)	91.8	89.4	90.5	89.9

Within that comparison, MobileNet-LSTM achieves 91.8% accuracy, 89.4% precision, 90.5% recall, and an 89.9% F1-score. Accuracy is 7.3 percentage points higher than standalone MobileNet, 11.6 points higher than standalone LSTM, and 8.7 points higher than the CNN baseline. Figure 3 reproduces these values in editable form. The gain supports the value of combining spatial features with temporal sequence modeling under the reported setup, but it should not be interpreted as outperforming all published detectors, several of which report higher benchmark accuracy under different datasets and protocols (Table 1).

5.2. Discussion

The internal comparison indicates that spatial-temporal integration is the main source of improvement. Standalone MobileNet can capture frame-level appearance artifacts but lacks explicit sequence modeling, whereas standalone LSTM has limited value when temporal modeling is not supported by strong spatial embeddings. Combining the two allows the recurrent layer to operate on compact visual features rather than raw frames, which explains the consistent improvement across all four reported metrics.

The result should be read alongside the literature rather than in isolation. The cited CNN-ViT and transformer-oriented studies report approximately 97-98% accuracy in benchmark settings [10]-[12], which is higher than the 91.8% reported here. The proposed framework, therefore, trades some peak benchmark performance for a design that can reduce compute through MobileNetV3 triage, selective frequency verification, and shallow recurrence, while also incorporating provenance and human oversight. That operational combination is the principal novelty of the work.

Production deployment would require additional validation. False positives could delay legitimate breaking-news media, while false negatives could allow synthetic content to pass into editorial workflows. Domain shift from re-encoding, low-light capture, unusual cameras, and emerging generation methods may reduce reliability. Human escalation can also become a bottleneck at peak volume. A practical deployment plan is therefore phased: first, validate on a newsroom-representative holdout set; then calibrate risk thresholds and branch weights; next, run the system in shadow mode without affecting publication; and finally introduce editor-facing alerts with audit logs, provenance evidence, and periodic error analysis. These steps would provide the real-world evidence that is not available in the present experiment record.

5.3. Limitations

The current study has four principal limitations. First, the experiment record does not preserve the exact dataset composition, train/validation/test split, or demographic and manipulation-type balance for the proposed model result. Second, the frequency-module hyperparameters and late-fusion weights are not specified. Third, the reported metrics do not quantify false-positive and false-negative costs under newsroom-specific operating conditions. Fourth, no production newsroom trial or throughput study is reported. These limitations restrict claims about generalization, reproducibility, and human-review scalability and define the priorities for follow-up validation.

6. Conclusion

Deepfakes remain a persistent threat to journalistic verification because realistic synthetic media can spread quickly while detector performance can deteriorate under domain shift, compression, and unfamiliar manipulation methods. The proposed framework addresses this operational gap by combining lightweight visual triage, shallow temporal modeling, selective frequency-aware verification, content-provenance checks, optional audio analysis, and human editorial review.

The reported MobileNet-LSTM configuration achieves 91.8% accuracy, 89.4% precision, 90.5% recall, and an 89.9% F1-score in the available internal comparison, outperforming the three reported baselines. The result does not establish state-of-the-art benchmark accuracy; instead, it supports the feasibility of combining efficient spatial and temporal analysis within a broader newsroom workflow that preserves human judgment.

Further work should document full dataset composition and training hyperparameters, validate cross-dataset behavior, calibrate fusion weights and escalation thresholds, measure false-positive and false-negative costs, and evaluate editor workload in production-like settings. Deployment should also address provenance interoperability, auditability, changing deepfake-generation techniques, and the scalability of human review during high-volume news events.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding Statement

No external funding was reported for this research.

References

[1] Guangyu Lin et al., "Fit for Purpose? Deepfake Detection in the Real World," *arXiv preprint*, pp. 1-43, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nadia Naffi, Deepfakes and The Crisis of Knowing: As Deepfakes Blur Reality, Education must go Beyond Detection, Teaching Students to Navigate Truth, Knowledge, and AI-mediated Uncertainty, UNESCO, 2025. [Online]. Available: https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing

[3] Tvesha Sippy et al., "Behind the Deepfake: 8% Create; 90% Concerned. Surveying Public Exposure to and Perceptions of Deepfakes in the UK," *arXiv preprint*, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Taylor Herzlich, AI Deepfakes of Nicolás Maduro Flood Social Media Depict Venezuelan Dictator in Jail with Diddy, Among other Vids, New York Post, 2026. [Online]. Available: https://nypost.com/2026/01/06/business/ai-deepfakes-of-nicolas-maduro-flood-social-media-depict-venezuelan-dictator-in-jail-with-sean-diddy-combs/

[5] Olivia Le Poidevin, UN Report Urges Stronger Measures to Detect AI-Driven Deepfakes, Reuters, 2025. [Online]. Available: https://www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11/

[6] Denis Campbell, AI Deepfakes of Real Doctors Spreading Health Misinformation on Social Media, The Guardian, 2025. [Online]. Available: https://www.theguardian.com/society/2025/dec/05/ai-deepfakes-of-real-doctors-spreading-health-misinformation-on-social-media

[7] Reshma Sunil et al., "Exploring Autonomous Methods for Deepfake Detection: A Detailed Survey on Techniques and Evaluation," *Heliyon*, vol. 11, no. 3, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sonam Singh, and Amol Dhumane, "Unmasking Digital Deceptions: An Integrative Review of Deepfake Detection, Multimedia Forensics, and Cybersecurity Challenges," *MethodsX*, vol. 15, pp. 1-30, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mubarak Alrashoud, "Deepfake Video Detection Methods, Approaches, and Challenges," *Alexandria Engineering Journal*, vol. 125, pp. 265-277, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ahmed Hatem Soudy et al., "Deepfake Detection using Convolutional Vision Transformers and Convolutional Neural Networks," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19759-19775, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Shereen Ahmed Hussien, and Seif N. Mohamed, "DeepFake Video Detection using Vision Transformer," *International Journal of Intelligent Computing and Information Sciences*, vol. 24, no. 1, pp. 55-68, 2024. [Google Scholar]

[12] Georgios Petmezas et al., "Video Deepfake Detection using a Hybrid CNN-LSTM-Transformer Model for Identity Verification," *Multimedia Tools and Applications*, vol. 84, no. 33, pp. 40617-40636, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Kutub Thakur et al., "Multimodal Deepfake Detection using Transformer-based Large Language Models: A Path Toward Secure Media and Clinical Integrity," *The American Journal of Engineering and Technology*, vol. 7, no. 5, pp. 169-177, 2025. [Google Scholar]

[14] Ping Liu, Qiqi Tao, and Joey Tianyi Zhou, "Evolving from Single-Modal to Multi-Modal Facial Deepfake Detection: Progress and Challenges," *arXiv preprint*, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yaning Zhang et al., "Distilled Transformers with Locally Enhanced Global Representations for Face Forgery Detection," *Pattern Recognition*, vol. 161, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Arash Heidari et al., "Deepfake Detection using Deep Learning Methods: A Systematic and Comprehensive Review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 2, pp. 1-45, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tony Mathew Abraham et al., "Leveraging Data Analytics for Detection and Impact Evaluation of Fake News and Deepfakes in Social Networks," *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Aryaf Al-Adwan et al., "Detection of Deepfake Media using a Hybrid CNN-RNN Model and Particle Swarm Optimization (PSO) Algorithm," *Computers*, vol. 13, no. 4, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ramcharan Ramanaharan, Deepani B. Guruge, and Johnson I. Agbinya, "DeepFake Video Detection: Insights into Model Generalisation - A Systematic Review," *Data and Information Management*, vol. 9, no. 4, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ismael Balafrej, and Mohamed Dahmane, "Enhancing Practicality and Efficiency of Deepfake Detection," *Scientific Reports*, vol. 14, no. 1, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Viacheslav Pirogov, and Maksim Artemev, "Evaluating Deepfake Detectors in the Wild," *arXiv preprint*, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]