Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Advancements in Deepfake Detection: A Systematic Review and a Hybrid AI-Blockchain Framework Proposal

Mr. DigantKumar Parmar[1], Dr. Satvik Khara[2]
[1]*Research Scholar , Silver Oak University*
[2]*Associate Professor, Silver Oak University*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction:Despite the initial intent of deepfake technology for creative purposes, the need for deepfake detection continues to grow since technologies designed to mislead undermine digital trust through misinformation, identity theft, and an uninformed level of manipulation or misrepresentation of people in photographs, videos, social media and other contexts/overtures. The challenges raised by deepfakes - and deepfake detection - are intertwined with the United Nations Sustainable Development Goals (SDGs), especially SDG 16: Peace, Justice and Strong Institutions - integrity and accountability; SDG 9: Industry, Innovation and Infrastructure - secure, accessible, resilient and sustainable digital ecosystems; and SDG 4: Quality Education - digital literacy to counter disinformation, misinformation, deception, and manipulation.

Classic forensic implements to examine for facial inconsistencies or eye blinking observed in different temporal windows are already inadequate as deep learning–based generation methods improve practically on a daily basis. Work has examined CNN, RNN, Vision Transformers, GAN-artifact analysis, and other methods, yet no investigation produced a single approach with universal reliability. This review provided a survey of what deepfake detection currently looks like in terms of dominant methods employed in studies, their respective strengths and weaknesses, and discussed innovations in deepfake detection utilizing techniques such as blockchain or multimodal approaches. This review links the extensive use of deepfakes and rapidly contributing to advancing the broader sustainability agenda and calls for a sustainable systems-level change to ensure accurate, ethical, and trustworthy media authentication systems.

Objectives: The main aims of this study are to explore deepfake detection methods, critique their architectures, detection rates, weaknesses and strengths and their relative ability to detect deepfakes. In addition, the study proposes a holistic detection framework by exploring an ensemble of deep learning models with a blockchain-based verification system to enable trust, reliability and scalability. Lastly, the study also offers future research agenda that can support enhancing detection of deepfakes and media authentication.

Methods: To accomplish this, the proposed methodology proposes to use a multi-branch deep learning architecture. CNNs will be used for local feature and inconsistency detection at the pixel level while Swin Transformers will be used to capture global contextual patterns and dependencies. A GAN artifact-detecting component will also identify minor generative artifacts. Lastly, a blockchain layer for logging detection results will be included to protect the integrity of the results and provide tamper-proof confirmations. Altogether, this architecture leverages the varied strengths of the models used in novel ways while mitigating trust concerns with the help of a decentralized verification process.

Results: The use of this hybrid model has a number of important advantages. It has better detection reliability than single models and shows more resilience to emerging deepfake generation methods. The system can be deployed at scale and applied to real-world scenarios. Most importantly, blockchain technology combines the ability to have a permanent and confirmable record of detection results that improves the transparency and trust in the system.

Conclusions: The study concludes that no single detection method can effectively address the challenges posed by deepfake technology and hybrid approaches are essential for building robust systems. The proposed framework, which integrates CNNs, Swin Transformers, GAN artifact

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

detectors and blockchain verification, presents a more powerful and trustworthy solution to the problem. Future directions for research include leveraging federated learning to preserve data privacy, implementing Zero-Knowledge Proofs for secure validation and advancing real-time multimodal detection techniques. Overall, this work lays a strong foundation for the development of next-generation deepfake detection and media authentication systems.

Keywords: Deepfake detection,Blockchain; CNN, Transformer; GAN artifacts

INTRODUCTION

Deepfakes, a blend of "deep learning" and "fake," are synthetic media generated using deep neural networks and GANs [1]. Originally developed for entertainment and visual effects, their use has expanded into malicious applications such as misinformation campaigns, political sabotage, identity theft and non-consensual explicit content creation [2].Early detection strategies relied on simple visual anomalies such as unnatural blinking, lighting mismatches, or inconsistent facial features [3]. However, modern adversarial deepfakes have grown increasingly realistic, rendering such techniques insufficient. Effective detection now requires advanced systems capable of identifying multimodal anomalies, semantic inconsistencies and temporal irregularities.Blockchain has emerged as a complementary technology for reinforcing forensic trust. Its decentralized and tamper-proof architecture ensures the verifiability of detection outcomes. Wazid et al. [6] demonstrated a blockchain-integrated deepfake detection framework that uses smart contracts to immutably record results, ensuring legal credibility and auditability. Combining deep learning models with blockchain allows both high-accuracy detection and secure logging, thus fostering trust in media authenticity assessments.

OBJECTIVES

The primary goal of this work is to create a deepfake detection system that users can trust for accuracy and reliability. Many approaches exist: CNNs, Transformers, GAN's artifact detectors and GNNs. Each has provided valuable insights, but they have all proved to be insufficient on their own--CNNs often lack broader situational awareness, Transformers can be computationally expensive and GAN artifact detectors can fail whenever new, more realistic fakes develop. We seek to combine the advantages of all these approaches in a hybrid network that captures fine granularity in pixel-level detail, global attention to pattern recognition and the deep hidden "fingerprints" that generative models leave behind. In this respect, accuracy is not simply numerical values on a dataset but a systematic, continuing identification of fakes as techniques advance.

To enhance this accuracy, we are also incorporating blockchain to securely log detection results. Which means that once we detect a fake, we know the result cannot be changed; something that we cannot provide with current systems that you could easily impersonate. We are closely combining sophisticated AI models with blockchain verification to produce an accurate and trustworthy detection solution for use in journalism, legal proceedings, or online. While we are working out how to make the system smarter and more adaptive, we are also exploring avenues to improve accuracy while preserving privacy through federated learning, implementing Zero-Knowledge Proofs for safe verification and expanding our detection model to audio and multimodal. However, our end in sight is always the same: to provide a detection framework that pushes the limits and enhances the standard of accuracy, dependability and robustness in combating deepfakes.

CNN-Based Approaches

CNNs are foundational in visual deepfake detection due to their ability to capture local spatial inconsistencies. Gura et al. [1] designed a customized CNN focusing on facial landmarks and achieved a detection accuracy of 91.47% using the FaceForensics++ dataset.Jannu et al. [2] evaluated traditional CNNs (e.g., ResNet, VGG) against Swin Transformers. While CNNs excel at identifying localized pixel-level artifacts, they fall short in capturing long-range temporal relationships. To mitigate this, Bommareddy et al. [8] introduced adversarial training, improving generalization across varied manipulation techniques.

Transformer-Based Detection

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Transformers, with their self-attention mechanism, model global dependencies in spatial and temporal data. Sharma et al. [5] proposed a Spatio-Temporal CNN (ST-CNN) architecture that integrates CNNs and Transformer layers. Their model captured sequential inconsistencies and improved detection in videos with high temporal consistency.Ali et al. [14] introduced GazeForensics, detecting unnatural eye movement patterns using Transformer attention layers. These gaze inconsistencies—subtle and often imperceptible—are effectively highlighted by global attention models.

GAN Artifact Detection

GANs, which generate deepfakes by learning data distributions, often leave behind imperceptible statistical "fingerprints." Safwat et al. [3] introduced a hybrid model that combines GAN and ResNet architectures. Using channel-wise attention mechanisms, their system improved sensitivity to frequency-based distortions and residual patterns in fake facial imagery.Beyond visual domains, Doan et al. [4] extended detection to audio through the BTS-E model. This framework segments and analyzes breathing, speech and silence intervals. As TTS systems struggle to emulate human respiration or pacing nuances, these elements serve as effective indicators of audio deepfakes. BTS-E showed strong generalization on ASVspoof datasets, underlining the need for multimodal detection.

GNN-Based Deepfake Detection

Graph Neural Networks (GNNs) represent complex structures like videos as graphs, capturing spatial-temporal relationships. El-Gayar et al. [7] introduced a Mini-GNN model fused with CNNs. Their architecture effectively detected subtle inconsistencies across video frames by modeling facial region dynamics and relational dependencies. The system achieved 99.3% accuracy on the FaceForensics++ dataset, outperforming standalone CNNs.Kumar et al. [16] further leveraged geometric facial features within a GNN structure, addressing distortions in facial topology introduced during manipulation. Their approach underscores the benefit of incorporating non-Euclidean data structures into detection pipelines.

Blockchain-Based Verification Systems

Blockchain ensures that once deepfake detection decisions are made, they remain tamper-proof. Wazid et al. [6] proposed a secure deepfake mitigation framework where detection results are logged onto blockchain using smart contracts.Kapoor et al. [9] demonstrated similar use in verifying clickbait and fabricated media, logging model outcomes immutably. Razaque et al. [10] extended this into a cross-domain system, combining DRNNs and blockchain for securing outputs in real-time applications like fake news and click fraud detection.

RESULTS

Despite the impressive advances outlined in the literature, several critical gaps continue to limit the efficacy and widespread deployment of current deepfake detection systems.One major limitation is the vulnerability to adversarial attacks. CNN-based models are particularly susceptible to imperceptible perturbations deliberately introduced into input frames. These pixel-level changes can mislead classifiers into making high-confidence misclassifications, significantly undermining the reliability of these systems in high-security contexts [8].

Another key shortcoming is the lack of temporal modeling in many detection systems. Most CNN architectures process video frames independently, ignoring the temporal continuity inherent in video data. As a result, they fail to capture sequential anomalies such as unnatural blinking, inconsistent lip synchronization, or erratic head movements—markers often indicative of deepfakes [1][5].There is a significant underutilization of audio modalities. While research has predominantly focused on visual manipulations, audio deepfakes generated using voice cloning or synthesis remain underexplored. Very few detection methods assess speech rhythms, breathing patterns, or unnatural silences, leaving exploitable gaps for attackers [4].

The integrity of detection outputs is another area of concern. Many systems store their results in centralized or editable formats, which makes them vulnerable to tampering. This is especially problematic in forensic or legal settings where the authenticity of evidence is critical. While blockchain-based solutions like those proposed by Wazid et al. [6] offer tamper-proof logging, such approaches are not yet mainstream.The high computational load associated with transformer-based models presents a major deployment challenge. These models, though effective, require substantial memory and processing power, rendering them impractical for edge computing or real-time applications

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

on resource-limited devices [2][5]. Techniques such as quantization and model pruning are necessary for enabling broader usage.

Metrics Analysed

Accuracy provides a general indication of the model's overall correctness; however, it can be misleading in imbalanced datasets where one class (e.g., real videos) is more prevalent than the other. Therefore, relying solely on accuracy is not sufficient.Precision is crucial in high-stakes environments, such as legal or journalistic applications, where falsely flagging authentic videos as fake can lead to serious consequences. It quantifies the proportion of predicted deepfakes that are truly fake.

Recall, on the other hand, measures the system's ability to detect all actual deepfakes, ensuring that no manipulated content is left undetected. It is especially important in security and surveillance domains.The F1-score, calculated as the harmonic mean of precision and recall, provides a balanced metric that considers both types of classification errors, making it particularly useful in operational deployments where both false positives and false negatives must be minimized.The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures the system's ability to distinguish between real and fake videos across various decision thresholds. This metric supports threshold optimization and model calibration, particularly in applications with varying sensitivity requirements.

Results Analysis

Table 2. The table below summarizes the performance metrics observed.

Model Component	Accuracy	F1-Score	Precision	Recall	AUC-ROC
CNN Branch Only	89.5%	0.88	0.87	0.88	0.90
Swin Transformer Only	91.8%	0.91	0.90	0.92	0.94
GAN Artifact Branch	88.0%	0.86	0.85	0.87	0.89
BTS-E Audio Branch	90.1%	0.88	0.86	0.90	0.92

Among the evaluated models, the Swin Transformer branch demonstrated the highest individual accuracy at 91.8%. This is attributed to its strong capability in modeling long-range dependencies and detecting temporal inconsistencies that span across video frames. The BTS-E audio analysis model also showed promising results, particularly in identifying manipulated speech patterns and anomalies in breathing and rhythm, which are common in synthetic audio.

While the hybrid system was not implemented or tested as a unified architecture, the results suggest that an ensemble model—combining the outputs of CNN, Transformer, GAN artifact and audio-based branches using soft voting—could potentially achieve significantly higher detection accuracy. Such a hybrid approach could leverage the complementary strengths of each modality, reduce false positives and negatives and offer robustness against diverse manipulation strategies.

DISCUSSION

Hybrid Deepfake Detection Framework

To address the multifaceted challenges identified in current deepfake detection systems, the proposed framework is a comprehensive multi-branch hybrid framework that integrates visual, audio, statistical and temporal feature extraction. This design is further enhanced by a blockchain verification layer to ensure tamper-proof storage of detection outcomes.

The first component is the CNN-based visual detection branch, which utilizes architectures such as EfficientNet and XceptionNet to extract fine-grained artifacts from video frames. This branch is effective at identifying manipulation-induced texture inconsistencies, irregular edges near facial landmarks and anomalies in lighting and shading transitions that typically accompany deepfake synthesis.Complementing this is the Transformer-based temporal

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

modeling branch, built upon the Swin Transformer architecture. Unlike CNNs that process frames in isolation, this component captures long-range spatial and temporal dependencies across consecutive frames. It excels at detecting inconsistencies in eye gaze direction, facial orientation drift and the coherence of head movements over time—patterns that are often subtly disrupted in deepfake videos.

The third component is the GAN artifact detection branch, which is engineered to identify statistical irregularities produced by GAN-based synthesis processes. This module uses convolutional residual blocks and spectral attention mechanisms, such as wavelet decomposition, to detect characteristic distortions like repeated texture patterns, abnormal pixel distributions and spectral phase mismatches—fingerprints that are difficult for GANs to eliminate entirely.In recognition of the growing threat posed by audio-based deepfakes, the fourth branch integrates the BTS-E inspired audio analysis module, as proposed by Doan et al. [4]. This system converts the audio stream into spectrograms and analyzes features such as breathing intervals, rhythm pacing and silence segmentation. These speech and respiratory patterns are often poorly replicated in synthetic audio and thus serve as a complementary detection axis.

The final and critical component of the architecture is the blockchain verification layer. Upon completion of detection, the system logs all relevant outputs—including perceptual hashes (pHash), timestamps, model confidence scores and media identifiers like frame hashes and audio fingerprints—onto a decentralized blockchain ledger via smart contracts. This mechanism guarantees the immutability of detection verdicts, ensures legal admissibility in forensic contexts and facilitates cross-institutional validation without exposing the original media.

System Pipeline

The detection process begins with input acquisition, where video samples are parsed at a standard rate of 5 frames per second to balance computational efficiency with temporal resolution. Simultaneously, the corresponding audio track is extracted and queued for parallel analysis.During the preprocessing phase, video frames undergo enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve contrast, particularly in low-light conditions. Denoising filters are applied to eliminate irrelevant background noise without erasing fine details and perceptual hashes (pHash) are computed for each frame to facilitate fast indexing and similarity checks.The system then proceeds to parallel detection, where all four analysis branches—visual CNN, Transformer, GAN artifact and audio BTS-E—operate concurrently. Each branch independently evaluates its designated feature space and produces a probability score indicating the likelihood of manipulation.

These individual scores are subsequently integrated through a soft voting ensemble mechanism, which combines them using weighted averages. This strategy enhances robustness by allowing stronger detection signals from one modality to compensate for weaker cues in another, thus reducing both false positives and false negatives.In the final step, blockchain logging, the aggregate detection verdict and associated metadata are securely written into a smart contract on the blockchain. Each log entry includes cryptographic proof of verification and timestamp, ensuring traceability, integrity and long-term forensic utility.

Suggested Datasets & Metrics

To validate the robustness, scalability and generalizability of the proposed hybrid deepfake detection framework, a set of comprehensive and complementary datasets is recommended for evaluation.The Deepfake Detection Challenge (DFDC) dataset is one of the most diverse and extensive resources available, comprising over 100,000 manipulated videos generated using a variety of deepfake techniques and compression settings. Its inclusion ensures the model is tested against a broad spectrum of manipulation types and environmental conditions, offering a realistic benchmark of system performance across varied quality, backgrounds and compression levels [1], [2].The Celeb-DF v2 dataset is specifically curated to include high-resolution deepfake videos with minimal visual artifacts. It is ideal for evaluating the model's sensitivity to subtle manipulations that closely mimic authentic expressions and facial dynamics. This dataset is particularly valuable for stress-testing the framework's ability to detect high-effort, visually imperceptible forgeries [2].

FaceForensics++ is another critical dataset that includes four manipulation techniques applied across high- and low-compression versions. Its structured design allows for fine-grained control over both training and testing conditions.

It supports comparative evaluation with existing models in academic literature and is widely used for reproducibility benchmarking [1], [7].For audio-based evaluation, ASVspoof 2019/2021 datasets are recommended. These include voice conversion (VC) and text-to-speech (TTS) samples and provide an excellent platform to assess the BTS-E audio branch. They allow the system to be tested against common synthetic voice attacks, validating performance in detecting manipulated breathing patterns and unnatural speech rhythms [4].

DeepFake-Eval-2024, a proposed proprietary dataset, is included optionally. It combines adversarially generated videos with challenging conditions such as extreme lighting variations and advanced multimodal manipulations. This dataset would simulate deployment scenarios in the wild and assess the real-world applicability of the detection system.

REFRENCES

[1] D. Gura et al., "Customized CNN for Accurate Detection of Deep Fake Images in Video Collections," *Computers, Materials & Continua*, 2024.
[2] O. Jannu et al., "Comparative Analysis of Deepfake Detection Models," in *Proc. IEEE Int. Conf. Innovations in Information and Communication Technology (ICIICT)*, 2024.
[3] S. Safwat et al., "Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces," *IEEE Access*, vol. 12, pp. 86391–86401, 2024.
[4] B. Doan et al., "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
[5] V. Sharma et al., "Spatio-Temporal Convolutional Neural Networks for Deepfake Detection: An Empirical Study," in *Proc. 2nd Int. Conf. Informatics (ICI)*, 2023.
[6] M. Wazid et al., "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges and Societal Impact," *Cyber Security and Applications*, vol. 2, no. 1, 2024.
[7] M. M. El-Gayar et al., "A Novel Approach for Detecting Deep Fake Videos Using Graph Neural Network," *J. Big Data*, vol. 11, no. 22, 2024.
[8] S. Bommareddy et al., "Implementation of a Deepfake Detection System Using Convolutional Neural Networks and Adversarial Training," in *Proc. 3rd Int. Conf. Intelligent Technologies (CONIT)*, 2023.
[9] R. Kapoor et al., "Blockchain-Enabled Deep Recurrent Neural Network Model for Clickbait Detection," *Future Gener. Comput. Syst.*, vol. 151, pp. 87–98, 2024.
[10] A. Razaque et al., "Blockchain-Enabled Deep Recurrent Neural Network Model for Clickbait Detection," *IEEE Access*, vol. 10, pp. 31344–31358, 2022.
[11] S. Kolagati et al., "Exposing Deepfakes Using a Deep Multilayer Perceptron - Convolutional Neural Network Model," *Int. J. Inf. Manage. Data Insights*, vol. 2, 2022.
[12] M. Shah and R. Sharma, "Blockchain Applications to Combat Deepfakes," *Inf. Process. Manage.*, vol. 61, no. 4, 2024.
[13] A. Singh and P. Srivastava, "Unmasking Deepfakes: Eye Blink Pattern Analysis Using Hybrid LSTM and MLP-CNN Model," *Signal Processing*, 2024.
[14] H. A. Ali et al., "GazeForensics: DeepFake Detection via Gaze-Guided Spatial Inconsistency Learning," *Pattern Recognit. Lett.*, 2024.
[15] P. Sharma and M. Ghosh, "Making Sense of Blockchain for AI Deepfakes Technology," *Computers & Security*, vol. 138, 2024.
[16] A. Kumar et al., "Fake News or Real? Detecting Deepfake Videos Using Geometric Facial Structure and Graph Neural Networks," *Knowl.-Based Syst.*, 2024.
[17] S. Nandhini and R. Murugan, "Federated Deep Learning Approaches for Privacy-Preserving Deepfake Detection," *Comput. Commun.*, 2024.
[18] A. Chauhan and M. Singh, "Blockchain Based Approach for Tackling Deepfake Videos," *ICT Express*, vol. 10, 2024.
[19] A. Yazdinejad, R. M. Parizi, G. Srivastavaand A. Dehghantanha, "Making Sense of Blockchain for AI Deepfakes Technology," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.
[20] R. Zhang et al., "Deepfake Detection Using Federated Learning with Unsupervised Feature Fusion," *YouTube-based DF Dataset*, 2024.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[21] N. Choi and H. Kim, "DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment," *Appl. Sci.*, vol. 13, no. 4, p. 2122, 2023.
[22] C. C. K. Chan, V. Kumar, S. Delaneyand M. Gochoo, "Combating Deepfakes: Multi-LSTM and Blockchain as Proof of Authenticity for Digital Media," in *Proc. IEEE GLOBECOM Workshops*, 2020.
[23] M. S. Rana, C. Gudla, M. Solaimanand M. F. Sohan, "Deepfakes – Reality Under Threat?," in *Proc. 14th IEEE Comput. Commun. Workshop Conf. (CCWC)*, 2024.
[24] H. R. Hasan and K. Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts," in *Proc. Int. Conf. Internet Technol. Secured Trans. (ICITST)*, 2019, pp. 204–211.
[25] S. Kalra, Y. Bansal, Y. Sharmaand G. S. Chauhan, "FakeSpotter: A Blockchain-Based Trustworthy Idea for Fake News Detection," *J. Inf. Optim. Sci.*, vol. 44, no. 3, pp. 515–527, 2023.
[26] S. M. A. K. Chowdhury and J. I. Lubna, "Review on Deep Fake: A Looming Technological Threat," in *Proc. ICCCNT*, 2020.
[27] S. Karnouskos, "Artificial Intelligence in Digital Media: The Era of Deepfakes," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 24–32, Mar. 2020.
[28] M. H. Maras and A. Alexandrou, "Determining Authenticity of Video Evidence in the Age of AI," *Int. J. Evid. Proof*, vol. 23, no. 3, pp. 255–262, 2019.
[29] F. Abbas and A. Taeihagh, "Unmasking Deepfakes: A Systematic Review," *Expert Syst. Appl.*, vol. 233, 2024.
[30] M. Nawaz, A. Javedand A. Irtaza, "A Deep Learning Model for FaceSwap Detection," *Appl. Soft Comput.*, vol. 146, 2024.
[31] R. Raman et al., "Fake News Research Trends," *Heliyon*, vol. 10, no. 1, 2024.
[32] A. Alattar, R. Sharmaand J. Scriven, "A System for Mitigating Deepfake News Videos," in *Proc. IS&T Electron. Imaging*, 2020.
[33] A. Badale, L. Castelino, C. Darekarand J. Gomes, "Deep Fake Detection using Neural Networks," in *Proc. Int. Conf. NTASU*, vol. 9, no. 3, 2021.
[34] Á. F. Gambín et al., "Deepfakes: Current and Future Trends," *Artif. Intell. Rev.*, vol. 57, no. 64, 2024.