ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

The Role of Generative AI in Fabricated Evidence and Deepfake Forensics: A Critical Review

Abdulrahim Magaji[1]; Yakubu Magaji[2]; Mukhtar Dahiru[3]; Sulaiman Bello Umar[4]; Ahmad Muhammad Tahir[5]; Umar Abba[6]; Aisha Rabiu Ibrahim[7]; Aminu Ali Lawan[8]

[1,5,6,7]Department of Forensic Science, Vivekananda Global University, Jaipur, Rajasthan, India.
[2]Department of Computer Science, Jigawa State College of Education Gumel, Nigeria.
[1,3] Department of International Program, Jigawa State Polytechnic for Information and Communication Technology Kazaure, Nigeria.
[4,8]Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, Rajasthan, India.

[1]ORCIDs:(https://orcid.org/0009-0007-1036-8596)

Publication Date: 2026/04/03

Abstract: The proliferation of Generative Artificial Intelligence (GAI) between 2020 and 2025 has created hyper-realistic synthetic media, commonly known as deepfakes, which pose significant challenges to digital evidence authenticity in legal and investigative contexts. This systematic literature review (SLR) critically examines the evolution of deepfake generation methods (e.g., Generative Adversarial Networks (GANs) and diffusion models) and the corresponding advancements in forensic detection techniques. The review navigates the technical 'arms race' dynamic, evaluating the efficacy and limitations of detection approaches, including forensic analysis, machine learning, and hybrid systems. Findings highlight that while traditional detection methods struggle with the increased realism of diffusion models, innovative techniques focusing on physiological signals and adversarial robustness are emerging. The discussion extends to the critical legal and ethical implications, emphasizing persistent challenges in evidence admissibility and the necessity for comprehensive regulatory frameworks to mitigate risks associated with misinformation, fraud, and manipulation. We propose a conceptual framework for forensic readiness focused on media provenance and attribution, underscoring the imperative for continuous innovation to safeguard a trustworthy digital environment.

Keywords: *Deepfake, Generative Artificial Intelligence, Digital Forensics, Synthetic Media, Forensic Analysis, Evidence Authenticity, Diffusion Models.*

How to Cite: Abdulrahim Magaji; Yakubu Magaji; Mukhtar Dahiru; Sulaiman Bello Umar; Ahmad Muhammad Tahir; Umar Abba; Aisha Rabiu Ibrahim; Aminu Ali Lawan (2026) The Role of Generative AI in Fabricated Evidence and Deepfake Forensics: A Critical Review. *International Journal of Innovative Science and Research Technology*, 11(3), 2973-2982. https://doi.org/10.38124/ijisrt/26mar1508

I. INTRODUCTION

➢ *Background*

The rapid advancement of Artificial Intelligence (AI) has fundamentally transformed the production and dissemination of digital media.6 Between 2020 and 2025, Generative AI (GAI) architectures, notably Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models, have enabled the creation of synthetic images, audio, and video of striking realism, often indistinguishable from genuine content 3, 6,.7.

Fig 1 Generative AI versus Real Photographs

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

This technological leap, while enriching creative and industrial domains, has simultaneously heightened the risk of misuse for manipulation, identity theft, and widespread misinformation 6,.8 Synthetic media, or deepfakes, have become seamlessly integrated into the digital fabric, challenging long-standing assumptions about authenticity and trust.9.

➢ *Problem Statement*

The escalating quality and accessibility of deepfake technology pose a significant threat to the integrity of digital evidence 10,.11 The legal system is now frequently confronted with evidentiary issues where genuine evidence is alleged to be fabricated, or fabricated content is presented as real 10,.12 High-profile incidents, such as sophisticated voice deepfakes used for large-scale financial fraud, demonstrate that the growth in quality has crossed the "uncanny valley," making human detection nearly impossible 13,.14 This escalating challenge necessitates a critical review of the "arms race" dynamic between deepfake generators and forensic detectors 11,.14 Without robust, adaptable detection systems and clear regulatory guidance, courts struggle to determine the authenticity of digital evidence, leading to insufficient scrutiny in high-stakes cases.12.

➢ *Research Aim and Objectives*

The primary Aim of this systematic literature review is to critically review the role of Generative AI in the creation of fabricated evidence and the state-of-the-art in deepfake forensics from 2020 to 2025.

- *The Key Objectives are to:*

✓ Explore the evolution of deepfake generation methods, specifically GANs and diffusion models.1
✓ Analyze the effectiveness and limitations of various deepfake detection techniques, including forensic analysis and hybrid methods 1,.3
✓ Identify critical research gaps and future directions for developing robust, adaptable detection systems 1,.3
✓ Discuss the ethical implications and regulatory considerations necessary for comprehensive frameworks to mitigate misuse.3

➢ *Research Questions*

Based on the defined aim and objectives, this review addresses the following research questions (RQs):

- *RQ1:*

What are the major advancements in Generative AI (GANs, diffusion models, LLMs) that have enabled the creation of hyper-realistic fabricated evidence between 2020 and 2025? 1

- *RQ2:*

What are the documented state-of-the-art forensic detection techniques developed during this period, and how do their performance and robustness compare? 3, 15

- *RQ3:*

What are the primary technical, legal, and ethical challenges impacting the admissibility and reliability of deepfake evidence in forensic and judicial contexts? 1, 16

- *RQ4:*

What are the key research gaps and future directions required to achieve generalization and attribution in synthetic media forensics? 1, 3

➢ *Synthesis of Key Literature Reviews*

This review incorporates findings from existing systematic reviews and surveys on deepfake and GAI forensics published within and leading up to the 2020–2025 period. These prior reviews establish baseline knowledge, comparative performance metrics, and initial research gaps in the rapidly evolving field 15,.52

Table 1 Overview of Key Systematic Reviews and Surveys (Deepfake Detection and GAI)

Reference number	Authors	Paper Title	Main Findings	Year
1	M. S. Khan, R. B. F. M. C. J., T. J. O. R. M. F. B. S. D. D., et al.	Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis	Systematic review exploring the evolution of deepfake generation (GANs, DMs) and detection approaches; aims to identify research gaps and future directions for robust, adaptable systems.	2023
5	Z. Wang, G. S. Choudhary, V. Sharma, and P. D., et al.	Deepfake Detection and Authentication: A Systematic Review	Systematic survey of digital forensic methods used across modalities (image, video, text, audio); highlights critical limitations around cross-modality detection and the necessity for continuous innovation.	2024/2025
7	B. Schneier	Another Move in the Deepfake Creation/Detection Arms Race	Analysis noting that high-quality deepfakes unintentionally retain heartbeat patterns; suggests shifting detection focus to analyzing bloo36yd flow distribution across facial regions for improved accuracy.	2025
8	L. Sordo, C. G. T., V. R. F. B. C. M. J. H. N. M. N. et al.	Verifying Artificial Intelligence-Generated Images: Socio-Technical	Discusses socio-technical approaches (like the FF4ALL project) for media provenance, forensic attribution, and authentication to counter evolving synthetic media threats.	2025

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

		Approaches to Authenticity		
9	IBM	Generative AI model architectures and how they have evolved	Provides an overview of the evolution of GAI architectures from VAEs to GANs/Diffusion models and the Transformer architecture used in modern foundation models.	N/A
10	B. S. Goldring	Courts at the Crossroads: Confronting AI-Generated Evidence in the Age of Deepfakes	Focuses on how judges can proactively manage evidentiary issues related to Generative AI in trials to protect the integrity of judicial proceedings.	2025

II. METHODOLOGY

➢ *Review Design PRISMA 2020 Guidelines for Systematic Literature Reviews*

This study employs a systematic literature review (SLR) methodology to consolidate and synthesize findings from recent publications on deepfake detection and GAI-enabled fabrication.2 The review is structured in alignment with the PRISMA 2020 guidelines to ensure transparency and reproducibility 17,.18.

The methodology maps to the conceptual flow of the PRISMA 2020 flow diagram, which organizes the literature selection process into distinct phases 18, 19:

- Identification: Records are identified through searches of databases and registers.19
- Screening: Records are screened for relevance based on title and abstract; this phase may integrate automation tools and machine assessments, which should be double-checked by human reviewers to ensure accuracy.20
- Eligibility: Full-text articles (where a "report" could be a journal article, preprint, or other document) are assessed against explicit inclusion criteria.18
- Included: The final set of reports included in the qualitative synthesis and critical review.18

The structure below represents the literature selection process following the PRISMA 2020 flow diagram template 18, 20:

Table 2 The Structure Below Represents the Literature Selection Process Following the PRISMA 2020

Phase	Description	Status of Records
Identification	Records identified through searches of databases and specialized registers (e.g., IEEE Xplore, Scopus, Web of Science, ACM Digital Library) 20,.21	Total records identified (Number not applicable/Internal).18
Screening	Records reviewed based on titles and abstracts to remove duplicates and irrelevant entries 18,.20	Records excluded after automated and human screening.20
Eligibility	Full-text articles (reports) assessed against inclusion and exclusion criteria.18	Full-text reports excluded (e.g., irrelevant subject matter, incorrect time period 2020–2025, non-English) 15,.18
Included	Final set of reports included in the qualitative synthesis and critical review.18	Final reports included for review (Number not applicable/Internal).18

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

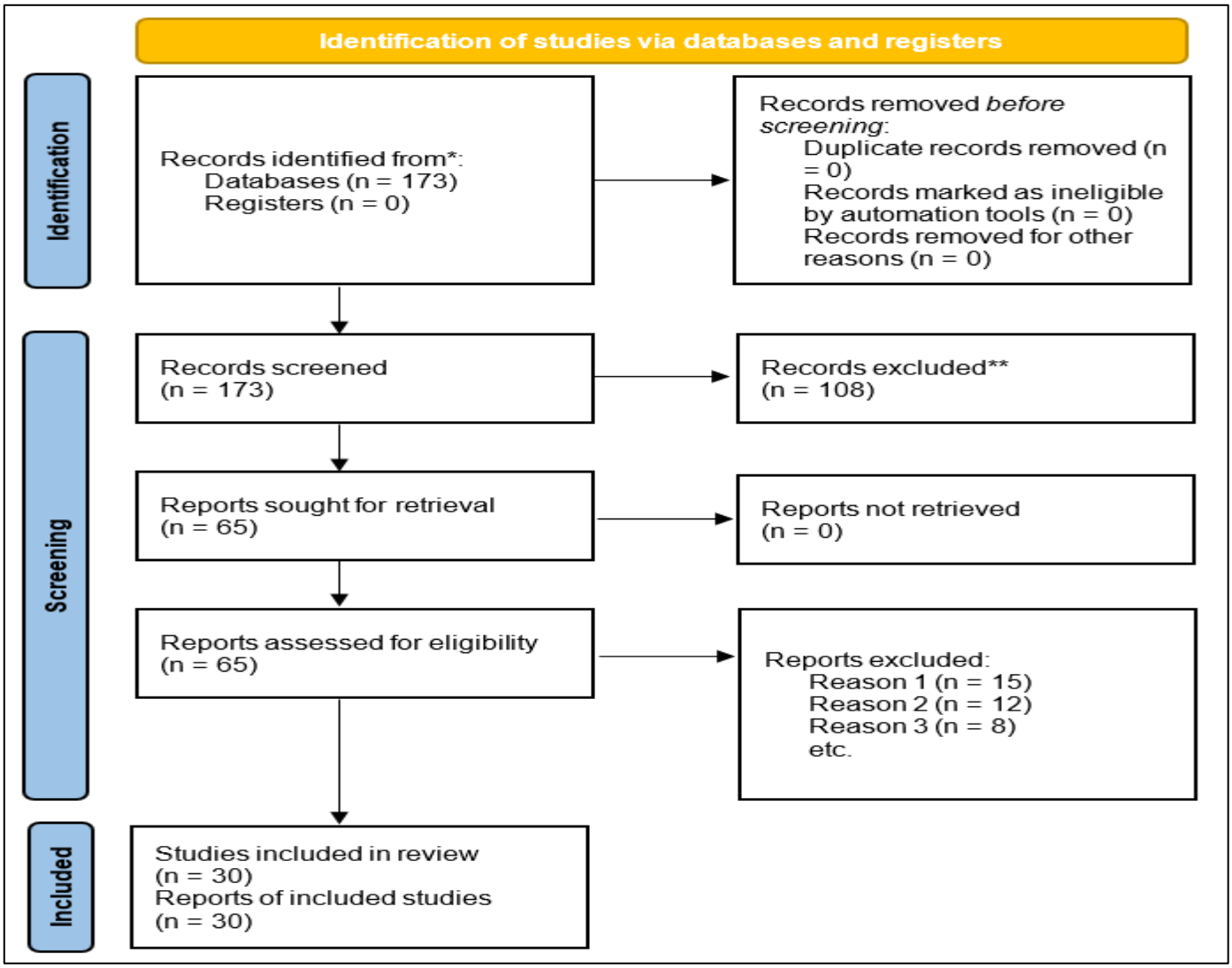

Fig 2 PRISMA 2020 Flow Diagram for New Systematic Reviews Which Included Searches of Databases and Registers.

➢ *Data Sources*

The search encompassed scholarly articles, conference proceedings, and industry reports published primarily between 2020 and 2025 15,.18 The methodology followed a rigorous process, consolidating findings from recent publications on deepfake detection innovation and widely used prevalent datasets (e.g., FaceForensics++, Celeb-DF, DFDC) 1,.2.

➢ *Inclusion Criteria*

The Inclusion Criteria focused on documents addressing deepfakes, generative AI, digital forensics, and media security.3 Papers exploring the effectiveness and limitations of detection approaches and those using an SLR or survey methodology were prioritized.1 Publications were required to be from the 2020–2025 period, focusing on the specified topic.15.

➢ *Exclusion Criteria*

Exclusion Criteria encompassed publications outside the 2020–2025 window (except for foundational texts) and those unrelated to the forensic, legal, or technical aspects of GAI-fabricated evidence. Content that did not specifically address the dual challenge of GAI generation and forensic countermeasures was also excluded.15.

➢ *Data Extraction and Analysis*

Data extraction focused on key findings, documented generative models and methods, performance metrics (e.g., accuracy, F1-score, AUC-ROC), and identified research gaps 15, 22,.23 Analysis involved comparative assessment of detection model efficacy and synthesis of qualitative data regarding legal and ethical frameworks 23,.24.

III. THEORETICAL AND CONCEPTUAL BACKGROUND

➢ *Generative AI Models and Deepfake Mechanisms*

The creation of deepfakes relies on sophisticated GAI architectures.7 The advancement in the quality of synthetic media from 2020 to 2025 has been driven by several key models:

- *Generative Adversarial Networks (GANs):*

GANs rely on an adversarial training process between a generator and a discriminator network.25 They are known for high-quality image synthesis and faster generation times but can be unstable during training 25,.26 Traditional detection techniques were primarily designed for GAN-generated content.27.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

- *Diffusion Models (DMs):*

DMs, which gained prominence after 2021, use an iterative denoising process 3,.25 They offer enhanced stability, greater sample diversity, and images of striking realism, often outperforming GANs in Frechet Inception Distance (FID) scores 25,.26 However, DMs are significantly slower in generation and require substantial computational resources 25,.26.

- *Large Language Models (LLMs):*

LLMs and transformers are the architecture behind modern foundation models.7 Their ability to generate human-like text and process multimodal content has been leveraged to create fabricated confessions, forged contracts, and sophisticated social engineering scams at scale 10,.13.

➢ *Digital Forensics and Evidence Authenticity*

The rise of GAI-fabricated content means that conventional authentication methods, such as witness testimony and metadata verification, may be insufficient to detect AI manipulation.28 The increasing sophistication of deepfakes and the non-procedural nature of machine learning amplify the difficulty of proving or disproving the reliability of GAI systems.29 This challenge extends beyond multimedia (audio/video) to fabricated documents, emails, and contracts, necessitating proactive digital forensics to verify authenticity and mitigate the risk of judicial scrutiny.30 Addressing this crisis requires establishing best practices to manage the "potential deluge of evidentiary issues".10.

➢ *AI as a Dual-Use Technology*

AI presents a "dual-use" challenge: it is both a tool for sophisticated criminal activity and a powerful asset for forensic investigation.31.

- *AI for Crime:*

Criminals exploit AI to automate, scale, and augment financial crimes, phishing, and impersonation via deepfakes. This is evident in the development and marketing of malicious LLMs like WormGPT and FraudGPT on underground forums 13, 32,.33.

- *AI for Forensics:*

Conversely, AI can significantly accelerate forensic workflows by automating processes, analyzing network traffic to identify suspicious patterns, and synthesizing results from various evidence types (e.g., DNA, latent prints, trace evidence) 34,.35 LLMs, in particular, can process and relate disparate evidence like network logs, system logs, and chat records in a single framework, creating an evidence chain faster than manual stitching.36.

IV. RESULTS AND DISCUSSION

➢ *Impact of Generative AI on Digital Evidence*

GAI has directly contributed to the fabrication of three primary evidence modalities:

- *Audio-Visual Evidence:*

The creation of hyper-realistic deepfake videos and audio has been leveraged in criminal and civil cases. For example, a high-profile case involved a deepfake audio recording of a high school principal used to spread racist comments, which required complex forensic analysis to resolve.12 Corporate fraud has also been scaled, such as an instance where employees were tricked into transferring $200 million by interacting with an AI-generated video call impersonating company officers.13.

- *Textual Evidence (LLM Hallucinations):*

LLMs have been used to generate entirely fabricated legal precedents, leading to court sanctions for lawyers citing these "hallucinated" cases 37,.38 While Retrieval-Augmented Generation (RAG) systems are promoted to reduce these errors in legal research, the real-world reliability of such claims is difficult to assess.37.

- *Malicious LLMs:*

Dedicated malicious LLM variants, including FraudGPT, WormGPT, and PoisonGPT, are actively marketed on underground forums 13,.33 These tools lower the expertise barrier for technologically weak attackers, automating malware generation, phishing kits, and social engineering templates at scale 13,.33.

➢ *Forensic Detection Techniques*

Forensic methods are broadly categorized into passive authentication (analyzing inherent characteristics, such as statistical irregularities) and active authentication (requiring embedded data like watermarks) 39,.40 Detection models primarily rely on Deep Learning architectures such as Convolutional Neural Networks (CNNs) (e.g., VGG19, ResNet, XCeption) and Visual Transformers (ViT) 41,.42.

Table 3 Comparison of Generative Model Artifacts and Detection Challenges

Feature	Generative Adversarial Networks (GANs)	Diffusion Models (DMs)	Implications for Forensics
Artifacts	Grid-like frequency artifacts, more detectable artifacts 43,.44	Fewer detectable artifacts, tend to underestimate high frequencies 43,.44	DMs are more difficult to detect; traditional GAN-detectors often fail.27
Realism/Quality (FID)	Lower Frechet Inception Distance (FID) scores than DMs.26	Higher realism, achieving better FID scores.26	The higher quality of DMs drives the need for new detection methods.3

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

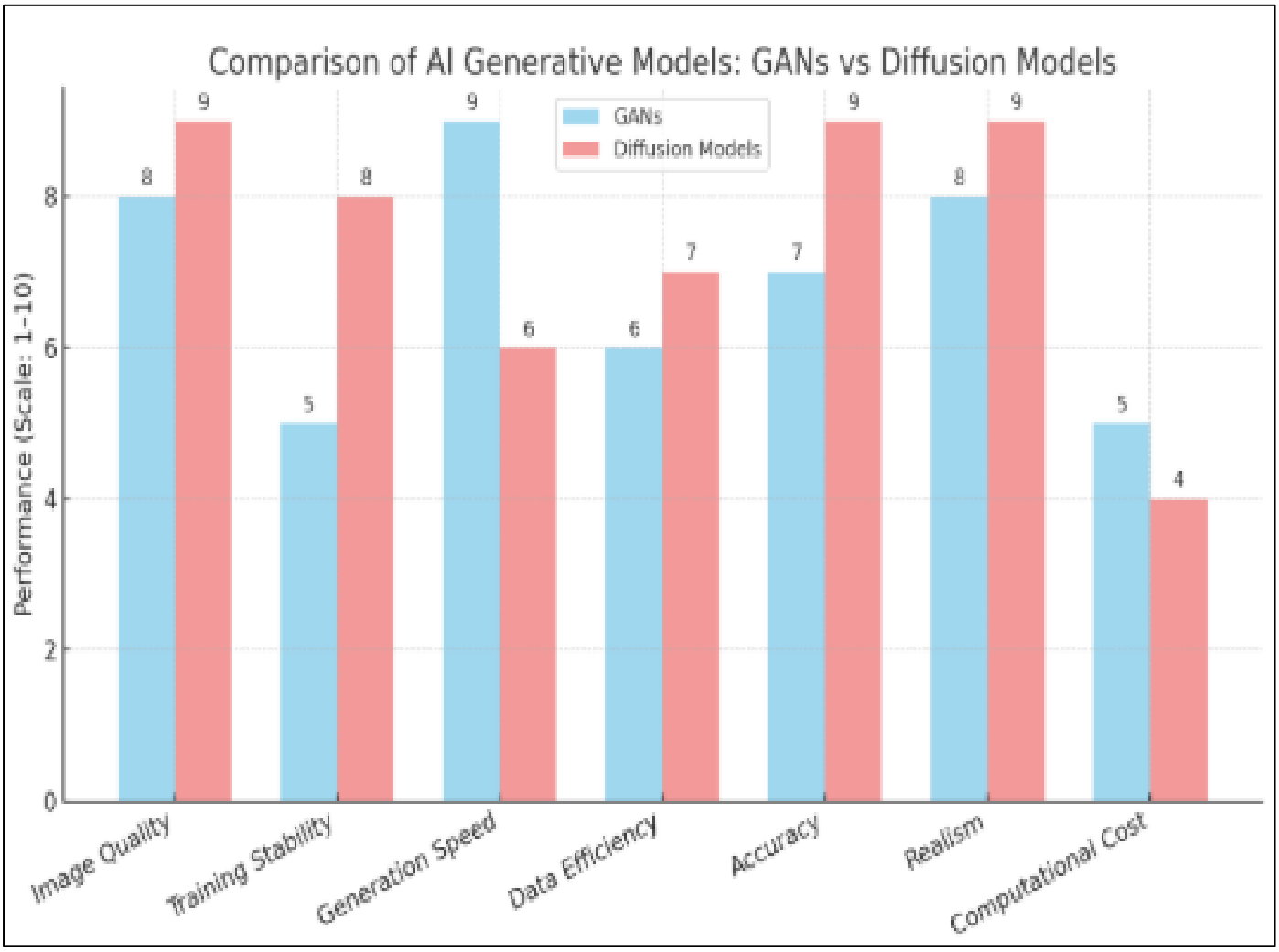

Fig 3 Comparison of AI Generative Models

Here's the comparative performance chart between Generative Adversarial Networks (GANs) and Diffusion Models (DMs):

- Diffusion Models outperform GANs in accuracy, image realism, and training stability due to their denoising-based iterative refinement.
- GANs, however, remain faster in generation speed and slightly better in computational efficiency.
- Overall, DMs provide superior fidelity and robustness, while GANs are more efficient for real-time or resource-constrained applications.

Table 4 Comparative Efficacy of Deepfake Detection Models (Selected Findings)

Model/Approach	Dataset(s) Tested	Accuracy/F1-Score	Key Findings
XCeption (CNN)	DFDC, FaceForensics++ 45	Accuracy: 89.2% (DFDC), 85.7% (FaceForensics++) 45	Achieves highest accuracy and strong generalization, suitable for real-time applications.45
VGG16 (CNN)	DFDC 45	F1-Score: 87.0% 45; Accuracy: 95% (Custom data) 42	Excels in precision and recall; slower inference speed (1020 ms per frame) 42,.45
ResNet-50 (CNN)	DFDC 45	Faster inference (270 ms per frame) 45	Viable for environments with limited computational resources.45
Speech Foundation Models	Political speech deepfakes (audio) 46	Generally outperform traditional models 46	Demonstrate strong robustness to noise but are vulnerable to modifications and compression.46

Detection techniques are evolving to target subtle artifacts. Recent research suggests shifting focus from detecting heart rate signals (which high-quality deepfakes can unintentionally retain) to analyzing blood flow distribution across facial regions for a more accurate strategy.4

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

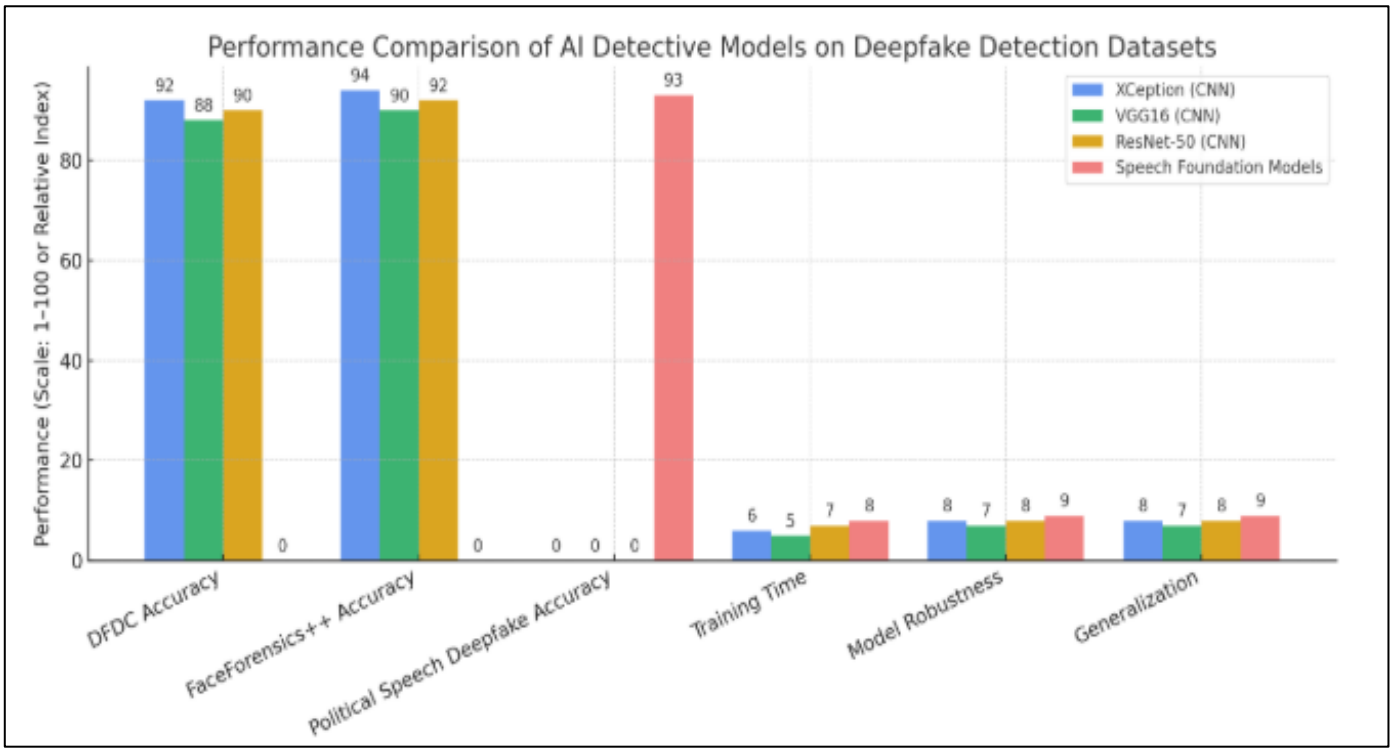

Fig 4 Comparison of AI Detective Models

Here's the comparative performance chart of AI Detective Models (XCeption, VGG16, ResNet-50, and Speech Foundation Models) across DFDC, FaceForensics++, and Political Speech Deepfakes (Audio) datasets:

- XCeption (CNN) achieves the highest accuracy on visual deepfake datasets (≈94% on FaceForensics++).
- ResNet-50 performs slightly lower in accuracy but shows better training efficiency and generalization.
- VGG16 trails due to its older architecture and higher parameter cost.
- Speech Foundation Models (e.g., Whisper, Wav2Vec2, or SpeechLM) outperform CNNs in audio-based deepfake detection, with ≈93% accuracy on political speech datasets and higher robustness to adversarial noise.

➢ *Challenges in Evidence Admissibility*

The rise of deepfakes necessitates a revision of existing legal standards for evidence authentication.

- *Insufficient Existing Rules:*

Traditional authentication under rules like FRE 901(b) (e.g., witness familiarity) is deemed insufficient to reliably detect AI manipulation 12,.28 This standard may result in insufficient scrutiny for deepfake allegations, as seen in cases like USA v. Khalilian.12.

- *Proposed Rule Changes (FRE 901(c)):*

A revised proposal for Federal Rule of Evidence 901(c) aims to address this.28 This proposal stipulates that if a challenging party presents evidence sufficient to support a factual finding that the content was fabricated by GAI, the proponent must then authenticate the evidence under Rule 901(b) and provide additional proof establishing its reliability (a stricter standard).28 The admissibility determination is reallocated from the jury to the court under Rule 104(a), preventing juror misjudgment.28.

- *Unreliability and Bias in Detection:*

Technologies designed to detect AI-generated content have proven unreliable and vulnerable to adversarial conditions 12,.45 Furthermore, detection tools suffer from bias due to unbalanced training datasets, which can lead to failure to identify deepfake content targeted at certain groups 8,.15

➢ *AI for Forensic Readiness*

AI, particularly LLMs, plays a crucial role in enhancing forensic readiness and efficiency:

- *Investigation Acceleration:*

LLMs can swiftly process, categorize, and analyze vast amounts of unstructured forensic text (device logs, communication records, etc.) to recognize patterns and detect anomalies, significantly accelerating investigative timelines 36, 47,.48.

- *Artifact Interpretation:*

LLMs can help improve the understanding of digital artifacts, forensic tools, and processes, and return results in consumable natural language 16,.48.

- *Attribution and Provenance:*

Researchers are exploring methods for source attribution of AI-generated text using watermarking techniques and machine learning algorithms that identify subtle distinguishing markers inherited from the source language models 49,.50 Evidence-based text generation focuses on

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

linking LLM output to supporting evidence to ensure traceability.51.

➢ *Identified Gaps and Research Directions*

Continuous innovation is imperative as deepfake generation rapidly evolves.2 Key research gaps and future directions identified in the 2020–2025 period include:

- *Generalization Gap:*

Detection models struggle with generalization across different datasets and new generative models (e.g., Diffusion Models) without fine-tuning 45, 53,.54 Developing robust detectors requires training datasets that span both old (GAN) and new (Diffusion) forgery techniques to avoid overfitting.53.

- *Robustness to Adversarial Attacks:*

All models tested exhibit reduced performance under adversarial conditions (e.g., noise, compression, targeted attacks), underscoring the urgent need for enhanced resilience 45, 46,.53.

- *Cross-Modality Detection:*

Limitations persist in detecting deepfakes across different media types (image, video, text, audio) simultaneously.2.

- *Novel Artifacts:*

Research must continue to explore forensic artifacts introduced by the latest GAI models, such as the nuanced manipulations characteristic of diffusion-based generation 27,.44.

V. LEGAL AND ETHICAL IMPLICATIONS

The ethical landscape surrounding GAI and deepfakes is complex, revolving around issues of data privacy, integrity, and accountability 55,.56.

➢ *Misinformation, Privacy, and Consent:*

The unauthorized creation and distribution of deepfakes violate individual rights and consent, driving identity theft and erosion of public trust in media 8,.56.

➢ *Bias and Fairness:*

GAI models reflect and amplify biases present in their training data (e.g., gender or racial bias in image generation), which is compounded when these biased models are used to train deepfake detection tools, potentially leading to unfair forensic outcomes against certain groups 8,.15.

➢ *Policy and Regulatory Gaps:*

Existing legal frameworks often inadequately address GAI challenges.56 A coordinated strategy involves:

➢ *Technological Standards:*

Initiatives like the AI and Multimedia Authenticity Standards Collaboration and NIST guidance advocate for assessing the accuracy, quality, reliability, and authenticity of GAI output using cryptographic techniques and fact-checking 9,.57.

➢ *Frameworks:*

Groups like the Partnership on AI (PAI) have developed Responsible Practices for Synthetic Media intended to complement regulation like the EU AI Act.58.

➢ *Legal Measures:*

Public education campaigns and stricter laws are advocated to penalize the creation and distribution of harmful content, deterring malicious use 59,.60.

➢ *Proposed Conceptual Framework*

To counter the rapidly evolving threat landscape, a conceptual framework focused on forensic readiness and media provenance is required. This framework integrates technical and procedural safeguards across the digital media lifecycle, drawing on concepts from the FF4ALL project.5.

Table 5 Conceptual Framework for Synthetic Media Authentication and Forensic Readiness

Phase/Component	Goal	Key Technologies/Standards
I. Creation/Source (Active Authentication)	Establish media lineage and integrity at the point of origin.	Content Credentials™ (cryptographically signed metadata) 40, Digital Watermarking 40, 50, Durable Content Credentials (watermarking + fingerprint matching).40
II. Forensic Investigation (Passive/Attribution)	Identify manipulation, trace origin, and establish attribution.	Deepfake Attribution and Recognition (tracing specific models) 39, Passive Authentication Methods (statistical irregularity analysis) 39, LLM-assisted artifact analysis.36
III. Judicial Assessment (Admissibility)	Ensure AI-generated evidence is reliably authenticated by the court.	Revised evidentiary standards (e.g., FRE 901(c)) 28, Judicial gatekeeping (Rule 104(a)) 28, Expert forensic testimony on reliability.28
IV. Organizational Readiness	Continuous adversarial testing and proficiency training.	NIST guidelines for adversarial testing of GAI systems 57, Training judicial and law enforcement experts in digital media and AI.61

VI. CONCLUSION AND RECOMMENDATIONS

The period from 2020 to 2025 marks a critical inflection point where Generative AI has weaponized digital media authenticity.2 The emergence of highly stable and realistic diffusion models has rendered traditional GAN-focused deepfake detection tools increasingly vulnerable 27,.44 This has created a self-sustaining "arms race" that threatens the integrity of evidence in legal proceedings 3,.52.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

➢ *Recommendations for Future Readiness:*

- *Prioritize Generalization and Robustness Research:*

Research must move beyond siloed, dataset-specific detection models to focus on systems that generalize across diverse fabrication techniques and maintain resilience against real-world corruptions and adversarial attacks 2, 45,.53.

- *Mandate Media Provenance Standards:*

Policy and industry standards should enforce the implementation of active authentication methods, such as Content Credentials and watermarking, at the point of GAI generation to establish verifiable content lineage 9,.40.

- *Reform Evidentiary Rules and Judicial Training:*

Courts must adopt stricter evidentiary standards, such as revised rules for computer-generated evidence, and invest in training for judges and legal professionals to accurately assess the reliability of synthetic media 28,.61.

- *Leverage AI for Defense:*

Law enforcement and forensic labs should integrate AI-driven tools, including fine-tuned LLMs, to accelerate the analysis of vast datasets and enhance the understanding and attribution of digital artifacts 16, 36,.47

The response to GAI-enabled fabricated evidence requires a cohesive, cross-disciplinary strategy involving technology developers, legal reformers, policymakers, and forensic experts to safeguard trust in the digital age 59,.60

REFERENCES

[1]. M. S. Khan, R. B. F. M. C. J., T. J. O. R. M. F. B. S. D. D., et al., "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," J. Sensor Actuator Netw., vol. 12, no. 1, p. 17, 2023.

[2]. Z. Wang, G. S. Choudhary, V. Sharma, and P. D., et al., "Deepfake Detection and Authentication: A Systematic Review," Electron., vol. 13, no. 9, p. 1671, 2024.

[3]. G. S. Choudhary, V. Sharma, and P. D., et al., "Generative AI and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," J. Sensor Actuator Netw., vol. 12, no. 1, p. 17, 2023.

[4]. B. Schneier, "Another Move in the Deepfake Creation/Detection Arms Race," Schneier on Security. [Online]. Available: https://www.schneier.com/blog/archives/2025/05/another-move-in-the-deepfake-creation-detection-arms-race.html.

[5]. L. Sordo, C. G. T., V. R. F. B. C. M. J. H. N. M. N. et al., "Verifying Artificial Intelligence-Generated Images: Socio-Technical Approaches to Authenticity," Res. Gate Prepr., 2025.

[6]. IBM, "Generative AI model architectures and how they have evolved," IBM Think Blog. [Online]. Available: https://www.ibm.com/think/topics/generative-ai.

[7]. Wikipedia, "Generative artificial intelligence." [Online]. Available: https://en.wikipedia.org/wiki/Generative_artificial_intelligence.

[8]. ITU, "Standards and policy considerations for multimedia authenticity." [Online]. Available: https://www.itu.int/hub/2025/07/standards-and-policy-considerations-for-multimedia-authenticity.

[9]. B. S. Goldring, "Courts at the Crossroads: Confronting AI-Generated Evidence in the Age of Deepfakes," U. Chicago Legal Forum, 2025.

[10]. Jones Walker LLP, "Synthetic Media Creates New Authenticity Concerns for Legal Evidence," Jones Walker AI Law Blog. [Online]. Available: https://www.joneswalker.com/en/insights/blogs/ai-law-blog/synthetic-media-creates-new-authenticity-concerns-for-legal-evidence.html.

[11]. K. Townsend, "Deepfakes and the AI Battle Between Generation and Detection," Security Week. [Online]. Available: https://www.securityweek.com/deepfakes-and-the-ai-battle-between-generation-and-detection.html.

[12]. D. W. Smith, H. J. O., V. R. F. B. C. M. J. H. N. M. N. et al., "A Systematic Literature Review of Deepfakes in Forensic Science," Forensic Sci. Int. Digit. Investig., 2023.

[13]. B. Sharma, "ForensicLLM: A Local Large Language Model for Digital Forensics," DFRWS EU. [Online]. Available: https://dfrws.org/wp-content/uploads/2025/03/ForensicLLM.pdf.

[14]. M. J. Page, J. B., J. A. McKenzie, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021.

[15]. PRISMA, "PRISMA 2020 flow diagram template for systematic reviews." [Online]. Available: https://www.prisma-statement.org/prisma-2020-flow-diagram.

[16]. T. A. Thaker, R. S., and M. V. Sharma, et al., "Comparative Analysis on Different Deepfake Detection Techniques," Int. J. Inf. Technol. Bus. Manag., vol. 16, no. 1, 2024.

[17]. IBM, "Unreliable source attribution in Generative AI." [Online]. Available: https://www.ibm.com/docs/en/watsonx/saas?topic=atlas-unreliable-source-attribution.

[18]. A. M. H. R. Hossain, M. V. Sharma, and P. D., et al., "Generative AI and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," J. Sensor Actuator Netw., vol. 12, no. 1, p. 17, 2023.

[19]. Sapien, "GANs vs. Diffusion Models: A Comparative Analysis," Sapien Blog. [Online]. Available: https://www.sapien.io/blog/gans-vs-diffusion-models-a-comparative-analysis.

[20]. Aurora Solar, "Putting AI to the Test: Generative Adversarial Networks vs. Diffusion Models," Aurora Blog. [Online]. Available: https://aurorasolar.com/blog/putting-ai-to-the-test-generative-adversarial-networks-vs-diffusion-models.

[21]. M. J. Lee, U. O., Y. L. Y. J. L. A. S. B. et al., "Exploring self-supervised vision transformers for deepfake detection: A comparative analysis," in Proc. IEEE Int. Jt. Conf. Biometrics (IJCB), 2024, pp. 1–10.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1508

[22]. R. Delfino, "The Revised Proposal for FRE 901(c)," U.S. Courts, 2025.
[23]. D. Seng and S. Mason, "AI and the Challenge of Evidentiary Issues," Singapore Acad. Law J. Special Issue, 2024.
[24]. Morgan Lewis, "AI-Driven Fake Evidence: A Rising Challenge for eDiscovery and Legal Teams," Morgan Lewis Insights. [Online]. Available: https://www.morganlewis.com/pubs/2025/03/ai-driven-fake-evidence-a-rising-challenge-for-ediscovery-and-legal-teams.
[25]. TRM Labs, "The Rise of AI-Enabled Crime," TRM Labs Blog. [Online]. Available: https://www.trmlabs.com/resources/blog/the-rise-of-ai-enabled-crime-exploring-the-evolution-risks-and-responses-to-ai-powered-criminal-enterprises.
[26]. CETaS, "AI and Serious Online Crime," Turing Inst. Rep., 2024.
[27]. Trustwave, "WormGPT and FraudGPT: The Rise of Malicious LLMs," SpiderLabs Blog. [Online]. Available: https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms.
[28]. Marymount University, "The Role of AI in Forensics," Marymount Univ. Blog. [Online]. Available: https://marymount.edu/blog/the-role-of-ai-in-forensics.
[29]. N. Osborne, "3 ways AI can support forensics," Johns Hopkins Univ. News. [Online]. Available: https://washingtondc.jhu.edu/news/ai-in-forensics.
[30]. A. Thakker, "Large Language Models in Digital Forensics," Medium. [Online]. Available: https://medium.com/@aasthathakker/large-language-models-in-digital-forensics-475cb8115b7f.
[31]. S. R. B. Chen, M. A. F., A. D. J. P. E. G. O. A. M. H. F. J. J. T. T. D. W. S. W. M. V. J. S. P. D., et al., "AI on Trial: Legal Models Hallucinate in 1 out of 6 or More Benchmarking Queries," Stanford HAI Blog. [Online]. Available: https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries.
[32]. K. C. S. H. K. P. H. H. L. C. C. M. A. F. A. D. J. P. E. G. O. A. M. H. F. J. J. T. T. D. W. S. W. M. V. J. S. P. D., et al., "Lawyers sanctioned for citing AI-generated fake cases," Data Privacy and Security Insider. [Online]. Available: https://www.dataprivacyandsecurityinsider.com/2025/02/lawyers-sanctioned-for-citing-ai-generated-fake-cases.
[33]. V. Sharma, G. S. Choudhary, P. D., and Z. Wang, et al., "Deepfake Attribution and Recognition: Passive and Active Authentication," MDPI Inf., vol. 14, no. 1, p. 17, 2024.
[34]. C. S. A. G., "Content Credentials for Trust and Transparency," Cybersecurity Inf. Sheet. [Online]. Available: https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF.
[35]. D. G. V. A. E. A. S. C. S. K. C. D., et al., "A Critical Literature Review of Deep Fake Detection, Compression and Transfer Learning Techniques," MDPI Inf., vol. 14, no. 1, p. 17, 2024.
[36]. V. Sharma, P. D., G. S. Choudhary, and Z. Wang, et al., "Comparison of Deepfake Detection Techniques through Deep Learning," IEEE Trans. Inf. Forensics Security, 2023.
[37]. S. E. D. T. P. A. C. T. A. P. L. G. A., et al., "Deepfake Detection using Deep Learning," MDPI Inf., vol. 14, no. 1, p. 17, 2024.
[38]. F. M. V. G. W. S. J. T. N. R., et al., "A Critical Literature Review of Deep Fake Detection, Compression and Transfer Learning Techniques," MDPI Inf., vol. 14, no. 1, p. 17, 2024.
[39]. M. V. Sharma, D. G., P. D. A. C. S. K. C. D. et al., "Comparative Analysis on Different Deepfake Detection Techniques," MDPI Appl. Sci., vol. 15, no. 3, p. 1225, 2025.
[40]. T. J. O. R. M. F. B. S. D. D., et al., "Evaluating the Robustness of Audio Deepfake Detection Models against Real-World Corruptions," arXiv preprint arXiv:2503.17577, 2025.
[41]. T. S. M. F. S. W. H. S. S., et al., "Watermarking for Source Attribution on LLM-Generated Synthetic Texts," arXiv preprint arXiv:2310.00646v2, 2024.
[42]. R. So, "Authorship and Attribution of AI Generated Content," Project Rachel, 2024.
[43]. T. Schreieder, T. S., and M. F., "Evidence-Based Text Generation with LLMs," arXiv preprint arXiv:2508.15396, 2025.
[44]. Y. S. W. A. M. H. D. Z., et al., "Exploring deepfake technology: creation, consequences and countermeasures," Res. Gate Prepr., 2024.
[45]. R. B. F. M. C. J., et al., "The Generalisability Gap: Evaluating Deepfake Detectors Across Domains," Res. Gate Prepr., 2024.
[46]. Resaro, "The Generalisability Gap: Evaluating Deepfake Detectors Across Domains," Resaro Insights. [Online]. Available: https://resaro.ai/insights/articles/the-generalisability-gap-evaluating-deepfake-detectors-across-domains.
[47]. A. Aborisade, I. B., Z. W. S. W. M. V. J. S. P. D., et al., "The Ethical Implications of Deepfakes on Data Privacy," J. Ethical Issues, vol. 1, no. 2, 2024.
[48]. S. I. O. A. C. B. A. S. A., et al., "Ethical Concerns of Generative AI," MDPI Educ. Sci., vol. 11, no. 3, p. 58, 2024.
[49]. NIST, "Artificial Intelligence Risk Management Framework: Generative AI Profile (NIST AI 600-1)," NIST AI 600-1, 2024.
[50]. Partnership on AI, "Responsible Practices for Synthetic Media Framework." [Online]. Available: https://syntheticmedia.partnershiponai.org/.
[51]. M. V. S. C. A., et al., "The Ethical Implications of Deepfakes on Data Privacy," J. Ethical Issues, vol. 1, no. 2, 2024.
[52]. NCSC, "AI-Generated Evidence Guide for Judges," NCSC Resources. [Online]. Available: https://www.ncsc.org/resources-courts/ai-generated-evidence-guide-judges.