Forensic Detection Of Toxic Chatbots, Deepfakes, And Automated Harmful Interactions: A State-Of-The-Art Review

Dr.Kiranbhai R Dodiya[1], Miss Ankita Kumari[2], Sudha Shetty[3], Dr. Parvesh Sharma[4], Dr. Kapil Kumar[5]

[1, 2]Teaching and Research Associate (TRA), Dept of Biochemistry and Forensic Science
[3]Assistant Professor
[4, 5] Assistant Professor, Dept of Biochemistry and Forensic Science
[5]Professor &Coordinator, Dept of Biochemistry and Forensic Science
[1, 2, 4, 5] Gujarat University, Ahmedabad, Gujarat, INDIA
[3]Sigma University, Vadodara, Gujarat, India

Abstract- *The rapid expansion of Generative AI (GenAI) technologies has enabled unprecedented synthesised media and sophisticated automated systems. The existence of advanced deepfake technology, generative large language models (LLMs) that can produce toxic content, and automated AI botnets has created substantial new obstacles in digital forensics. The present review consolidates recent work (2024–2025) concerning the forensics of three colliding AI threats: audio-visual deepfakes, toxic or weaponised chatbots, and the automated facilitation of harm. The review reconstructs the shift from artefact detection systems toward the use of behavioural, semantic, and multimodal forensics. The emerging methodologies, such as biological signal processing, audiovisual correlation, LLM systematisation fingerprinting, and traffic encryption biometrics, as well as the comprehensive integration of these methodologies, reveal vital yet unrefined methodologies that focus on the challenges of adversarial attacks, laundering via paraphrasing, and performance drops on real-world data. The review articulates the necessity of well-founded, transparent, and standardizable forensic systems that are meant to work in adversarial conditions.*

Keywords: Generative AI forensics; deepfake detection; toxic chatbot analysis; multimodal behavioural signatures; adversarial resilience

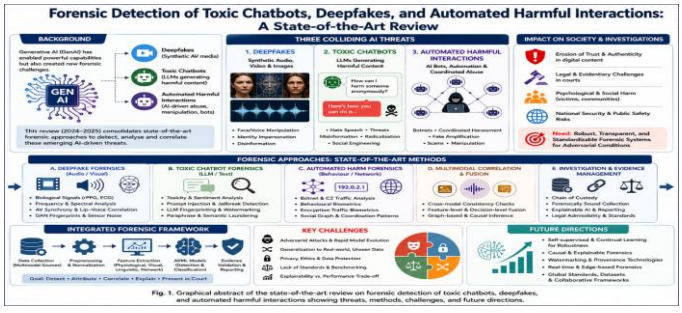

Figure 1. Graphical abstract illustrating the forensic detection framework for deepfakes, toxic chatbots, and automated harmful interactions using multimodal analysis, behavioural profiling, AI-driven forensic techniques, and explainable evidence to support robust and trustworthy digital investigations.

I. INTRODUCTION

Network forensics has always looked at metadata, file headers, log files, and the contents of device memory. With the new generation of GenAI systems, the focus of forensics has, of necessity, shifted to the contents of those data sets. Rather than pulling data and examining whether the files, images, or audio sequence are the result of artificial data generation or manipulation, the focus of forensic analysis needs to shift to whether investigators are pulling data that includes any images, video, audio, or text that has been manipulated or generated by artificial intelligence and whether these files have any evidentiary value that substantiates or supports the focus of investigation. In the case where this is probable, there is a necessity to examine those files[1]. Potentially there the whole gamut of diffusion models will come to play, including the rapidly expanded accessibility of the so-called GPT-4, Llama 3, Claude, and DeepSeek families. Deepfake scams, fabricated Silly-Chat and chatbot personas, automated campaign harassment, and social engineering manipulations have been exacerbated by these technologies, so there is a specialist focus[2].Perhaps more than any other innovation, the rapid development has outpaced forensics integration of computer vision, signal analysis, advanced behaviour analytics, and Linguistic processing. With increased contextual adaptability, LLMs are moving rapidly beyond coherence to coherence at deep, rational levels. Where the focus is primarily audio, visual, and text, Deepfake artefacts, the techniques of the past are waning. This Review is therefore focused primarily on forensics to detect and correlate the more insidious aspects of deepfake audio, chatbots, and their automated, orchestrated behaviours.[3].

II. LITERATURE REVIEW TABLE

Year	Title	Scope & Questions Addressed	Methods / Approaches Reviewed	Key Findings	Relevance to Your Review	Reference
2019	FaceForensic Benchmarking Manipulated Facial Media	Standardising deepfake datasets and evaluating early detectors	CNN-based detectors, Face2Face/FaceSwap/DF/NT pipelines	Introduced a large benchmark dataset; revealed early detection limitations	Foundation for visual deepfake forensics	[4]
2020	A Deepfake Detection Algorithm Based on the Fourier Transform of Biological Signals	Whether biological cues (rPPG) can detect fakes	rPPG extraction, FFT/ICA filtering	Biological signals outperform artefact-based detectors	Core reference for physiological detection	[5]
2020	DFDC – Large-Scale Deepfake Benchmark	Generalisation problem of deepfake detectors	Dataset creation, baseline CNN models	Large dataset shows severe generalisation gaps	Essential for the dataset limitations section	[6]
2020	End-to-end anti-spoofing with RawNet2	Detection of synthetic audio via raw-waveform learning	Sinc filters, CNN-GRU hybrid, ASVspoof	Strong baseline for audio deepfake detection	Supports your audio detection section	[7]
2021	*Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, and Countermeasures*	Reviews deepfake creation and detection; examines limitations and challenges.	GAN/autoencoder-based deepfake generation; CNN/ML-based detection; dataset comparisons.	Identifies major weaknesses in detectors, especially poor cross-dataset generalisation.	Provides foundational understanding of deepfake threats and detection gaps.	[8]
2021	AASIST – Spectro-Temporal Audio Anti-Spoofing	Unified model for synthetic speech detection	Graph attention nets + raw-wave encoder	Achieves SOTA anti-spoofing accuracy	Strong baseline for audio forensics	[9]

2021	Mouse Dynamics for Bot Detection	Whether behaviour patterns detect bots	LSTM/transformer sequence learning	High accuracy bot–human discrimination	Useful for the behavioural forensics section	[10]
2022	Wavelet-Packet-Based Deepfake Analysis	Frequency-domain cues for fake detection	Wavelet transforms, classifier models	Frequency artefacts remain detectable vs diffusion	Supports the frequency-analysis section	[11]
2022	BeCAPTCHA-Mouse – Behavioural Bot Detection	Synthetic mouse trajectory detection	Feature extraction, trajectory profiling	Real vs synthetic behaviour discriminable by ML	Enhances behavioural analysis	[12]
2022	TLS Fingerprinting (JA3/JA3S)	Identify clients & bots using TLS metadata	TLS ClientHello parsing, hashing, fingerprinting	Detects automation even with agent spoofing	Network-layer forensic technique	[13]
2022	*GAN-Generated Faces Detection: A Survey and New Perspectives*	Reviews detection methods for GAN-generated synthetic faces.	Deep-learning detectors, artefact analysis, and dataset comparison.	Highlights generalisation issues and weaknesses in current detectors.	Useful baseline for understanding image-level deepfake forensics.	[14]
2023	Audio Deepfake Detection – Survey	State of the art in synthetic speech forensics	Spectral, prosodic, raw waveform, datasets	Detectors struggle with unseen attacks	Key baseline for the audio section	[15]
2023	Survey on LLM Text Detection	Can AI-generated text be reliably detected?	Perplexity, stylometry, classifiers, watermarking	Paraphrasing breaks many detectors	Core for LLM forensic analysis	[16]
2023	rPPG-Based Deepfake Detection – Review	Robustness of physiological detection methods	rPPG pipelines, artefact suppression	rPPG works where pixel detectors fail	Biological-signal forensics support	[17]
2023	Paraphrasing Evasion of Detectors	How paraphrasing circumvents LLM detectors	DIPPER, back-translation, robustness tests	Paraphrasing ruins detection performance	Supports the counter-forensics section	[18]

2023	Multimodal Deepfake Detection – Survey	Does fusion improve robustness?	Audio-video-text fusion	Fusion greatly reduces false negatives	Supports a multimodal approach	[19]
2023	Localisation of Deepfake Regions	Explainability in forensic detection	Saliency maps, attention, patch models	Regional localisation improves evidentiary value	Useful for the XAI section	[20]
2023	Social Bot Detection – Survey	Coordinated inauthentic behaviour	Graph analysis, temporal patterns	Network + behaviour fusion strengthens detection	Relevant for harmful interaction analysis	[21]
2024	*Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook*	Reviews modern deepfake generation and detection techniques.	GANs, diffusion models, NeRF-based generation; multimodal detection methods.	Shows detectors struggle with new generative models; major robustness gaps remain.	Provides updated SOTA overview for deepfake forensics.	[22]
2024	Adversarial Attacks on Deepfake Detectors	Vulnerability analysis of video detectors	FGSM/BIM/ PGD attacks	Detectors are highly vulnerable	Supports the adversarial forensics section	[23]
2024	*AVT2-DWF: Audio-Visual Deepfake Detection with Dynamic Fusion*	Improves deepfake detection using audio-visual fusion.	Transformer-based audio & visual encoders; dynamic weight fusion.	Achieves higher cross-dataset accuracy.	Supports multimodal forensic detection.	[24]
2024	*Evolving from Single-modal to Multi-modal Facial Deepfake Detection: A Survey*	Reviews multimodal deepfake detection approaches.	Audio-visual fusion, text-visual methods, and dataset comparison.	Multimodal detectors outperform single-modal detectors.	Strong evidence for multimodal pipelines.	[25]
2024	SynthID Text Watermarking – Nature	Watermarking LLM outputs at scale	Token-distribution watermark embedding	Reliable watermarks; adversarial caveats	Attribution & provenance section	[26]
2024	*Adversarial Attacks and Defences for Deepfake*	Examines how	Adversarial attack	Shows detectors are	Critical for forensic	[27]

	Detection: A Comprehensive Analysis	adversarial attacks impact modern deepfake detectors and which defences work.	algorithms (FGSM, PGD, BIM), adversarial training, robust feature learning.	highly vulnerable to small perturbations; adversarial training improves robustness.	evaluation of deepfake systems and understanding counter-forensics.	
2024–25	Large Language Models: A Survey	Distinguish model-generated text	Stylometry, perplexity, and embeddings	Fingerprints degrade with paraphrase	Authorship attribution	[28]
2024–25	Comprehensive GenAI Threat Review	Meta-survey of detection & mitigation	Dataset/meta-analysis, roadmaps	Calls for multimodal + XAI + legal frameworks	For your conclusion chapter	[29]
2025	Multimodal AI Forensics (outlook papers)	Future of multimodal forensics	Cross-modal alignment, physiological fusion	Fusion + provenance yields the highest robustness	Helps define future research gaps	[30]
2025	Countering anti-forensic tactics in cybercrime investigations – a systematic literature review	How attackers evade detectors	Paraphrasing, voice conversion, diffusion rewriting	Evasion methods increasing; need provenance	Critical for threat modelling	[31]
2025	*Unmasking the Unknown: Facial Deepfake Detection in the Open-Set Paradigm*	Detecting fake faces created by **unknown/unseen**manipulation methods (beyond known deepfake techniques)	Supervised contrastive learning; open-set classification distinguishing real, known-fake, and unknown-fake classes	Model successfully detects deepfakes from unseen generation methods — more robust than closed-set detectors	Critical for real-world forensic detection, where new deepfake methods constantly emerge	[32]
2025	Transformation of Forensic Standards Post-GenAI	Legal and evidentiary implications	XAI, provenance, watermarking, benchmarking	Courts require transparency + explainability	Supports policy/legal sections	[33]

Table 1. Summary of key studies (2019–2025) on forensic detection of deepfakes, toxic chatbots, and AI-generated harmful interactions, highlighting their research scope, methodologies, major findings, and relevance to the present review.

Deepfake technology has rapidly transformed from basic face-swapping algorithms to lifelike, photo-realistic, temporally consistent videos that are indistinguishable to forensics and human viewers alike. These forensics have entered the human and erasing technology age and include the evaluation of spatial artefacts, the analysis of temporal coherence, and the acknowledgement of physiological signals. Innovative approaches are needed for computer-based technology forensics.[34].

III. VISUAL FORENSICS SPANS IMAGES AND VIDEOS.

3.1.1 Spatial Analysis.

The first GAN-based systems of deepfake technology demonstrated spatio-temporal flaws in optic signals, such as the checkerboard patterns that resulted from uneven upsizing operations. Although more recent deepfake technology using diffusion systems has greatly improved and almost eliminated such obvious artefacts, several high-frequency patterns associated with human features, such as the micro-textures on the surface of the eyelashes, iris borders, and skin pores, pose considerable challenges. Forensic analysts often use frequency-based methods such as the Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT) for the detection of divergent and attainable spectres of optical signals generated from the patterns of real signals. November more spatial forensics increasingly details the statistical irregularities of the gradients of illumination, the lens distortions, and the various sensor noise levels, which diffusion models have difficulty replicating.[35].

3.1.2 Temporal Analysis

Static images exhibit realism. On the other hand, deep fake videos are characterised by temporal inconsistencies. These inconsistencies can be described as flickering textures, unnatural changes between frames, and

inconsistencies within the facial features of people. Motion coherence and head pose stability across frames are evaluated using Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) units and 3D CNNs. To detect unnatural smoothing or movements that have been artificially interpolated, Optical Flow methods construct motion vectors. Temporal forensics is still the most effective method for detecting video deep fakes, as generative algorithms often prioritise realism for individual frames over the replication of long-term, temporal realism.[36].

3.1.3 Biological Signal Verification (rPPG)

Remote photoplethysmography (rPPG) is predicated on the premise that genuine human faces display infrequent, minute changes to the visibility of their skin, which is their way of showing circulation. In faces, the periodic changes are more perceptible to the eyes. Deepfakes, however, do not replicate this sort of physiology with sufficient spatial-temporal resolution. Forensic rPPG technology first identifies the forensic facial ROI, then isolates the RGB channel fluctuations in a structured way, followed by filtering to determine the existence of periodic biological signals, and determining the existence/absence of a biological beat. This technique is less vulnerable to pixel-level counterforensic movements because it is based on real biological phenomena.[37].

3.2 Audio Forensics

Voice cloning is a system that has advanced in real-time, with varieties of zero-shot and cross-lingual voice synthesis providing the potential of generating real-time speech. That said, real-time speech generation has discrepancies at the forensics level. Any potential for real-time generation has been forensically identified to an empirical margin.

3.2.1 Vocoder-Based Artefact Detection

Neural vocoders such as WaveNet, WaveRNN and HiFi-GAN are known to introduce novel changes in the harmonics, spectral brightness, and microjitter of the files they edit. Systems such as RawNet2 and AASIST are designed to identify these changes in speech and to classify the speech as synthetic. These systems identify regions of spectral smoothness, where holophonic continuum speech is absent and harmonically stable regions, which are indicative of real voice production[38].

3.2.2 Audio-Visual Consistency Forensics

The presence of audio deepfakes, along with video deepfakes, can use cross-modal consistency as a forensic marker. The human speech production system is such that phonemes are tightly coupled with visemes in that when someone speaks a phoneme, the viseme of that phoneme is

shown on the speaker's lips in a corresponding movement. Tools like LipForensics help determine whether there is a match between the lip movements and the speech content. However, in cases where there are very minor discrepancies between the lip movements and the audio cross-resolution patterns and in cases where the audio comprises content that is likely to be manipulated or synthesised, it becomes relatively straightforward to identify or synthesise audio content that is paired with the video deepfake content.[39].

3.3.3. Real-World Dataset Degradation

Recent studies, such as Deepfake-Eval- 2024 benchmark, reveal that the performance of deepfake detectors is deteriorating in the real world, compressed and noisy data, as the system is highly trained. But models that are highly trained on sound and video uncluttered, deepfake data sets suffer performance drops to the tune of 50\% when they are exposed to social media data that has compression artefacts, poor lighting and movement occlusion. This suggests that there is a need for improved and new model training to provide forensic systems that can accommodate real-time streaming and fast-paced data for systems for training in order to provide sufficient updated data.[40].

IV. FORENSIC ANALYSIS OF TOXIC CHATBOTS AND LLMS

LLMs now incorporated in communication applications, customer service interfaces, and social media streams have quickly become areas of potential abuse. Angry users have turned LLMs into hotbeds for generating hate and radicalisation, phishing, and misinformation automation. This abuse has a natural forensic component that requires attention to the triggering prompts for the LLM abuse, definitely, and for the forensic analysis of the LLM output to provide the text system generated by LLM.4.1 Prompt Engineering and Jailbreaking.[41].

While LLMs are developed and deployed with mitigations in place, adversaries are quick to circumvent those mitigations and effect prompt engineering. Scenarios that comprise role playing, emotional persuasion, hypotheticals, and adversarial poems/poetry can lead a model to ignore or circumvent a model’s guardrails. Research in 2025 indicated that craftily deployed prompts, in particular, poems or poetry that used metaphors, attained especially high rates of bypassing safeguards. Forensic investigations of this nature require a comprehensive record of user prompts and model responses, and classification of these pairs via transformer architectures employed for toxicity detection. Such records can determine if a model produced harmful content via prompt injection, an unguarded model, or unintentional user manipulation.[42].

4.1 Authorship Attribution and Stylometric Fingerprinting

Analysts are increasingly required to assess the possible authorship of text samples created by AI models and determine which specific models created these text samples. Text produced by AI models demonstrates predictability in token selection and shows consistency in sentence construction and style, which is why perplexity and burstiness values are low. Systems for authorship attribution analyse n-gram distributions, syntactic structures, and levels of lexical entropy. Recent studies have shown that different large language models (LLMs) leave unique stylistic fingerprints as a result of varied training data and sampling methodologies. These fingerprints improve attribution accuracy, with studies indicating above 95% accuracy in differentiating different models of GPT, Claude, Llama, and variants of DeepSeek.[43].

Watermarking techniques, including Google’s SynthID, aid in attribution by unobtrusively modifying token selection to introduce statistically recognisable patterns. Forensic checks involve examining token patterns to assess insufficient alignment with watermarking parameters. These methods, however, are susceptible to attacks that generate direct copies of samples while removing embedded signatures, known as watermarking.[44].

V. AUTOMATED HARMFUL INTERACTIONS (BOTNETS)

AI agents are turning off safety and security tools protecting against botnet attacks, which involve the application of DDoS attacks, the automation of fraudulent traffic generation, automated engagement, and data scraping. Many new AI tools exhibit human-like traits, making them more difficult to detect than traditional bots.[45].

5.1 Behavioural Biometrics

Authentication of identity through behaviour focuses on the entity's interacting activity rather than on the entity's identity on a computer network. Mouse dynamics analytics reveal bots in user sessions to move their mouse cursor along unnaturally straight, mathematically correct paths, and their mouse cursor movements are highly consistent. Contactless, non-human systems operate in fixed temporal patterns and can be distinguished from human-initiated systems (or use machines in a coordinated manner) through Behavioural Dynamics analytics. A suite of heuristics can assist a human

respondent in a discrete identification of automated malignant interactions, missing the network activity detection and control systems.[46].

5.2 TLS Fingerprinting (JA3)

JA3 type systems create fingerprints that register what has transpired at the level of the SSL and TLS handshakes. Even when botnets attempt to masquerade themselves by using realistic user-agent strings, the 'cryptographic' or algorithmic handshakes tell a different story, revealing the automation framework that lies underneath. The individual mappings of different suites of cyphers, elliptic curve preferences, and the sequences of extensions indicate whether a particular query has come from a true browser or automation tool such as Selenium, Puppeteer, or a Python-scripted library. Thus, TLS Fingerprinting is a highly reliable protocol metric by which some activities can be traced to particular bot infrastructures.[47].

VI. CHALLENGES AND COUNTER-FORENSICS

AI-powered modern threat vectors are adaptive to the systems being put in place to counteract them. The use of deepfake laundering, such as re-encoding, downscaling, or noise injection, applies stylistic filters that can interrupt watermark line signals and eliminate any frequency domain artefacts that are detectable. In the same way, when texts are paraphrased using different algorithms, the stylistic fingerprints are altered, and the reliability of perplexity-based classifiers is minimised. These types of evasion tactics bring a continuous challenge to the field of forensics and the available tools.[48].

Yet another important hurdle is the legal obligation to satisfy the principle of transparency. In the forensic domains, simple numerical predictions do not suffice, as there is a legal demand for interpretative and methodological transparency. This is where automated forensic assessments become validated through Explainable AI.[49]. In the analysis of visual deep fakes, Explainability shows which parts of a given image or video are inconsistent and how these inconsistencies parts are the reason for the concluded manipulation. In audio deep analysis, explainable systems are also required to demonstrate the atypical spectral or temporal features that differentiate artificial human voices from genuine human voices. In the analysis of toxic chatbots, explainability is concerned with the classification of text as either toxic or machine-generated through specific linguistic structures, patterns, token transitions, or situational contexts. These mechanisms of interpretability transform AI-generated probabilities into transparent legal forensic evidence.[50].

VII. CONCLUSION

The ability to accurately detect deepfakes, the workings of AI chatbots, and AI-driven botnets represents a newly developing area of study within the field of digital forensics. Next-generation techniques, including rPPG physiological verification, LLM stylistic fingerprinting, and TLS handshake biometrics, present promising pathways for the forensic world to remain relevant to the detection of AI tools. However, the development of GenAI is outpacing the ability to accurately forensicate the tools. Evasive behaviours of AI actors, consolidation of content, and degradation of real-world datasets to distract detection models have exposed the weaknesses of the AI tools. A combination of continuous model retraining, AI that explains the decision processes, and the application of C2PA might be the best hope for forensic digital systems to retain information true to the original GenAI artefact. The passage and advancements of time in the digital world and the realities of the world outside the virtual sphere make the world's social systems and the technologies that surround them even more intertwined. Synthesised media that appear to have been generated or created by a human make the development of legally defensible forensic tools a necessity.

REFERENCES

[1] G. Bendiab, H. Haiouni, I. Moulas, and S. Shiaeles, "Deepfakes in digital media forensics: Generation, AI-based detection and challenges," *Journal of Information Security and Applications*, vol. 88, p. 103935, Feb. 2025, doi: 10.1016/J.JISA.2024.103935.

[2] P. V. Falade, "Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 9, no. 5, pp. 185–198, Oct. 2023, doi: 10.32628/CSEIT2390533.

[3] J. Loovens and H. Tinmaz, "A systematic literature review of deepfakes in forensic science," *Forensic Imaging*, vol. 43, p. 200647, Dec. 2025, doi: 10.1016/J.FRI.2025.200647.

[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–11, Jan. 2019, doi: 10.1109/ICCV.2019.00009.

[5] Y. Ni, W. Zeng, P. Xia, G. S. Yang, and R. Tan, "A Deepfake Detection Algorithm Based on Fourier Transform of Biological Signal," *Computers, Materials*

and Continua, vol. 79, no. 3, pp. 5295–5312, 2024, doi: 10.32604/CMC.2024.049911.

[6] B. Dolhansky*et al.*, "The DeepFake Detection Challenge (DFDC) Dataset," Jun. 2020, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2006.07397

[7] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2021-June, pp. 6369–6373, Nov. 2020, doi: 10.1109/ICASSP39728.2021.9414234.

[8] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, Feb. 2021, doi: 10.1007/s10489-022-03766-z.

[9] J. W. Jung *et al.*, "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 2405–2409, Oct. 2021, doi: 10.1109/ICASSP43922.2022.9747766.

[10] H. Niu, J. Chen, Z. Zhang, and Z. Cai, "Mouse Dynamics Based Bot Detection Using Sequence Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12878 LNCS, pp. 49–56, 2021, doi: 10.1007/978-3-030-86608-2_6.

[11] M. Wolter, F. Blanke, R. Heese, and J. Garcke, "Wavelet-packets for deepfake image analysis and detection," *Machine Learning 2022 111:11*, vol. 111, no. 11, pp. 4295–4327, Aug. 2022, doi: 10.1007/S10994-022-06225-5.

[12] A. Acien, A. Morales, J. Fierrez, and R. Vera-Rodriguez, "BeCAPTCHA-Mouse: Synthetic mouse trajectories and improved bot detection," *Pattern Recognit*, vol. 127, p. 108643, Jul. 2022, doi: 10.1016/J.PATCOG.2022.108643.

[13] "TLS Fingerprinting with JA3 and JA3S - Salesforce Engineering Blog." Accessed: Dec. 04, 2025. [Online]. Available: https://engineering.salesforce.com/tls-fingerprinting-with-ja3-and-ja3s-247362855967/

[14] X. Wang, H. Guo, S. Hu, M. C. Chang, and S. Lyu, "GAN-generated Faces Detection: A Survey and New Perspectives," *Frontiers in Artificial Intelligence and Applications*, vol. 372, pp. 2533–2542, Feb. 2022, doi: 10.3233/FAIA230558.

[15] "(PDF) Audio Deepfake Detection: A Survey." Accessed: Dec. 04, 2025. [Online]. Available: https://www.researchgate.net/publication/373487442_Audio_Deepfake_Detection_A_Survey

[16] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275–338, Oct. 2023, doi: 10.1162/coli_a_00549.

[17] Q. Xu, H. Qiao, S. Liu, and S. Liu, "Deepfake detection based on remote photoplethysmography," *Multimed Tools Appl*, vol. 82, no. 23, pp. 35439–35456, Sep. 2023, doi: 10.1007/S11042-023-14744-Z; SUBPAGE:STRING: ABSTRACT; JOURNAL:JOURNAL: MTAA; WEBSITE:WEBSITE: DL-SITE; CSUBTYPE:STRING: PERIODICAL;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION.

[18] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," *Adv Neural Inf Process Syst*, vol. 36, Mar. 2023, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2303.13408

[19] X. Wan *et al.*, "One-step Multi-view Clustering with Diverse Representation," *IEEE Trans Neural Netw Learn Syst*, vol. 36, no. 3, pp. 5774–5786, Jun. 2023, doi: 10.1109/TNNLS.2024.3378194.

[20] B. Weng, G. A. Castillo, W. Zhang, and A. Hereid, "On the Adversarial Scenario-based Safety Testing of Robots: the Comparability and Optimal Aggressiveness," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 3299–3318, Apr. 2023, doi: 10.1109/TRO.2023.3267020.

[21] J. Martínez Llamas, K. Vranckaert, D. Preuveneers, and W. Joosen, "Balancing Security and Privacy: Web Bot Detection, Privacy Challenges, and Regulatory Compliance under the GDPR and AI Act," *Open Research Europe*, vol. 5, p. 76, 2025, doi: 10.12688/OPENRESEUROPE.19347.1.

[22] G. Pei *et al.*, "Deepfake Generation and Detection: A Benchmark and Survey," Mar. 2024, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2403.17881

[23] S. Khan, "Adversarially Robust Deepfake Detection via Adversarial Feature Similarity Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14556 LNCS, pp. 503–516, Feb. 2024, doi: 10.1007/978-3-031-53311-2_37.

[24] R. Wang, D. Ye, L. Tang, Y. Zhang, and J. Deng, "AVT2-DWF: Improving Deepfake Detection with Audio-Visual Fusion and Dynamic Weighting Strategies," *IEEE Signal Process Lett*, vol. 31, pp. 1960–1964, Mar. 2024, doi: 10.1109/LSP.2024.3433596.

[25] P. Liu, Q. Tao, and J. T. Zhou, "Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," Jun. 2024, Accessed: Dec. 04,

2025. [Online]. Available: https://arxiv.org/pdf/2406.06965

[26] S. Dathathri*et al.*, "Scalable watermarking for identifying large language model outputs," *Nature 2024 634:8035*, vol. 634, no. 8035, pp. 818–823, Oct. 2024, doi: 10.1038/s41586-024-08025-4.

[27] S. Khan, "Adversarially Robust Deepfake Detection via Adversarial Feature Similarity Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14556 LNCS, pp. 503–516, Feb. 2024, doi: 10.1007/978-3-031-53311-2_37.

[28] S. Minaee*et al.*, "Large Language Models: A Survey," Feb. 2024, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2402.06196

[29] F.-A. Croitoru *et al.*, "Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook," Nov. 2024, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2411.19537

[30] X. Wan *et al.*, "One-step Multi-view Clustering with Diverse Representation," *IEEE Trans Neural Netw Learn Syst*, vol. 36, no. 3, pp. 5774–5786, Jun. 2023, doi: 10.1109/TNNLS.2024.3378194.

[31] S. Surakanti, S. Goundar, and J. Dwight, "Countering anti-forensic tactics in cybercrime investigations – a systematic literature review," *Int J Inf Secur*, vol. 24, no. 5, Oct. 2025, doi: 10.1007/S10207-025-01131-Y.

[32] N. Bahavan, S. Saha, K. Chen, S. Seneviratne, S. Rasnayaka, and S. Halgamuge, "Unmasking the Unknown: Facial Deepfake Detection in the Open-Set Paradigm," Mar. 2025, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.08055

[33] "C2PA in India: Understanding it's Role, Legal Implications, and Business Adoption, 2025 – EPWA." Accessed: Dec. 04, 2025. [Online]. Available: https://epwa.in/publication/c2pa-in-india-understanding-its-role-legal-implications-and-business-adoption-2025/

[34] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," *Alexandria Engineering Journal*, vol. 125, pp. 265–277, Jun. 2025, doi: 10.1016/J.AEJ.2025.04.007.

[35] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting GAN DCT anomalies," *J Imaging*, vol. 7, no. 8, Aug. 2021, doi: 10.3390/jimaging7080128.

[36] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pp. 1205–1207, Oct. 2019, doi: 10.1109/ICCVW.2019.00152.

[37] K. Ikematsu, H. Oshima, R. Eardley, and I. Siio, "Investigating How Smartphone Movement is Affected by Lying Down Body Posture," *Proc ACM Hum Comput Interact*, vol. 4, no. ISS, Nov. 2020, doi: 10.1145/3427320.

[38] J. W. Jung *et al.*, "AASIST: AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 2405–2409, 2022, doi: 10.1109/ICASSP43922.2022.9747766.

[39] Y. Wang and H. Huang, "Audio–visual deepfake detection using articulatory representation learning," *Computer Vision and Image Understanding*, vol. 248, p. 104133, Nov. 2024, doi: 10.1016/J.CVIU.2024.104133.

[40] N. A. Chandra *et al.*, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," Mar. 2025, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.02857

[41] M. Q. Li and B. C. M. Fung, "Security Concerns for Large Language Models: A Survey," May 2025, doi: 10.1016/j.jisa.2025.104284.

[42] J. Xia, S. Shu, and X. Li, "Quantum fluctuation energies over a spatially inhomogeneous field background in a chiral soliton model," Mar. 2025, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.04015

[43] A. A. Najjar, H. I. Ashqar, O. Darwish, and E. Hammad, "Leveraging Explainable AI for LLM Text Attribution: Differentiating Human-Written and Multiple LLMs-Generated Text," *Information (Switzerland)*, vol. 16, no. 9, Jan. 2025, doi: 10.3390/info16090767.

[44] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models," Jul. 03, 2023, *PMLR*. Accessed: Dec. 04, 2025. [Online]. Available: https://proceedings.mlr.press/v202/kirchenbauer23a.html

[45] A. Mahboubi *et al.*, "The evolving threat landscape of botnets: Comprehensive analysis of detection techniques in the age of artificial intelligence," *Internet of Things*, vol. 33, p. 101728, Sep. 2025, doi: 10.1016/J.IOT.2025.101728.

[46] Prof. G. G. Sayyad, S. Kadam, K. Kadam, S. Godage, and R. Zagade, "Assessment of Bot Detection Approaches Using Behavioral Biometrics and Mouse Dynamics," *International Journal on Advanced Computer Engineering and Communication Technology*, vol. 14, no. 1, pp. 746–751, Nov. 2025, doi: 10.65521/IJACECT.V14I1.818.

[47] J. Heino, A. Hakkala, and S. Virtanen, "Categorizing TLS traffic based on JA3 pre-hash values," *Procedia Comput Sci*, vol. 220, pp. 94–101, Jan. 2023, doi: 10.1016/J.PROCS.2023.03.015.

[48] Y. Zha, R. Min, and S. Sushmita, "PADBen: A Comprehensive Benchmark for Evaluating AI Text

Detectors Against Paraphrase Attacks," Nov. 2025, Accessed: Dec. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2511.00416

[49] Z. Khalid, F. Iqbal, and B. C. M. Fung, "Towards a unified XAI-based framework for digital forensic investigations," *Forensic Science International: Digital Investigation*, vol. 50, p. 301806, Oct. 2024, doi: 10.1016/J.FSIDI.2024.301806.

[50] N. Yu, L. Chen, T. Leng, Z. Chen, and X. Yi, "An explainable deepfake of speech detection method with spectrograms and waveforms," *Journal of Information Security and Applications*, vol. 81, p. 103720, Mar. 2024, doi: 10.1016/J.JISA.2024.103720.