International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Efficient Hybrid CNN-Transformer Models for Benchmark-Driven Image Deepfake Detection and Analysis

Priya Mishra[1], Dr. Deepika Pathak[2]
[1]Research Scholar, [2]Vice Chancellor
[1,2]Faculty of Computer Sciences & Application, Dr. A. P. J. Abdul Kalam University, Indore, India

Abstract- The rapid evolution of deepfake generation technologies has created significant challenges for digital media authentication, cybersecurity and misinformation prevention. Modern deepfake images generated using advanced artificial intelligence techniques exhibit highly realistic visual characteristics, making manual identification increasingly difficult. This paper presents a comprehensive review of efficient hybrid CNN-transformer models for benchmark-driven image deepfake detection and analysis. The study critically examines recent advances in Convolutional Neural Networks (CNNs), transformer-based architectures, hybrid feature fusion frameworks and explainable AI techniques for identifying manipulated images, and analyses benchmark-driven evaluation strategies, multimodal feature representation, manipulation localisation and attention-based contextual learning. Twelve studies published in 2025 are reviewed and organised into three themes covering hybrid architectures, benchmark-driven feature analysis, and cybersecurity challenges. The reviewed evidence is consolidated into comparative tables that map each study to its focus, approach, principal finding and limitation, alongside a component-level comparison of CNN, transformer and hybrid designs and an assessment of modality coverage across the corpus. Major challenges including adversarial attacks, computational complexity, cross-domain generalisation, multimodal deepfakes and benchmark inconsistency are discussed, and seven research gaps are identified and mapped to corresponding future research directions focusing on explainable AI integration, multimodal fusion, computational optimisation and generalised feature learning.

Keywords- Deepfake detection; hybrid CNN-transformer models; benchmark-driven evaluation; vision transformers; deep learning; image forensics; explainable AI; multimodal learning; artificial intelligence; feature fusion.

I. INTRODUCTION

The advancement of artificial intelligence and deep generative models has transformed digital content creation and manipulation. Deepfake technology, which uses deep learning algorithms such as Generative Adversarial Networks (GANs), autoencoders and transformer-based generation models, can create highly realistic synthetic images and videos that are difficult to distinguish from authentic media. Although these technologies have beneficial applications in entertainment, education, virtual

communication and digital media production, their misuse has created serious concerns regarding misinformation, cybercrime, identity theft, political manipulation and social engineering attacks.

The widespread circulation of manipulated media across social networking platforms has made reliable and automated detection an urgent requirement. Traditional image forensic approaches based on handcrafted features and statistical analysis demonstrated limited effectiveness against modern synthetic media because they cannot capture complex manipulation patterns and contextual inconsistencies. Deep learning frameworks have therefore emerged as effective solutions, learning hierarchical and discriminative feature representations directly from visual data.

CNNs have been extensively used to extract local spatial features and texture-level artifacts associated with manipulated images, but often struggle to capture long-range contextual dependencies and global semantic information. Transformer-based architectures have gained attention because they model global relationships through self-attention. Hybrid CNN-transformer frameworks combine the strengths of both, integrating local spatial extraction with global contextual learning to improve robustness and classification accuracy.

Beyond architectural advances, benchmark-driven evaluation and feature analysis have become essential components of detection research. Benchmark datasets enable fair comparison across manipulation scenarios and real-world conditions, and advanced frameworks incorporate feature fusion, manipulation localisation, multimodal learning and explainable AI to improve interpretability and reliability.

Despite substantial progress, current systems continue to face challenges including computational complexity, adversarial attacks, poor cross-domain generalisation, benchmark inconsistency and limited scalability for real-world deployment. The increasing sophistication of multimodal generation involving images, text, speech and video introduces further complexity. There is therefore an increasing need for efficient hybrid CNN-transformer architectures capable of achieving high accuracy while maintaining computational efficiency and robustness across diverse benchmarks.

1. Objectives and Contributions of this Review

This review provides a comprehensive analysis of benchmark-driven hybrid CNN-transformer models for image deepfake detection. It makes four contributions. First, it consolidates twelve recent studies into a structured thematic review supported by comparative tables rather than narrative description alone. Second, it compares CNN, transformer and hybrid designs component by component, stating what each contributes and what it costs, so that the case for hybridisation rests on evidence rather than assertion. Third, it assesses the modality coverage of the corpus and the transferability of each study to the image domain, which materially affects how far the reviewed evidence supports conclusions about image detection. Fourth, it identifies seven research gaps and maps each explicitly to a corresponding future research direction.

2. Organisation of the Paper

Section 2 reviews the literature across three themes and consolidates it in comparative tables. Section 3 sets out the identified research gaps. Section 4 concludes, and Section 5 details future research directions.

II. LITERATURE REVIEW

1. Hybrid CNN-Transformer Architectures for Deepfake Detection

Hybrid models integrating CNNs and transformer-based frameworks have demonstrated superior performance in detecting manipulated visual content. Soundarya et al. (2025) proposed a CNN-based

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

framework using deep feature extraction for identifying manipulated images, and demonstrated that CNNs effectively capture spatial inconsistencies and texture-level artifacts in synthetic images.

Zafar et al. (2025) introduced a hybrid framework combining temporal and spatial feature learning, integrating CNN-based spatial extraction with sequential learning modules to improve robustness and classification performance, and showed that hybrid frameworks outperform standalone CNN models by capturing both local and contextual manipulation patterns. Alsolai et al. (2025) proposed Guardian-AI, a deep learning image deepfake detection model designed for robustness against highly realistic manipulations, achieving improved accuracy through optimised feature extraction and hybrid learning. Huang et al. (2025) introduced SIDA, a large multimodal model for social media image deepfake detection, localisation and explanation, integrating transformer-based contextual learning with explainable AI mechanisms to identify manipulated regions. This is the most complete study in the corpus with respect to the objectives of the present review, since it addresses detection, localisation and explanation together rather than treating interpretability as a separate concern, and it evaluates on social media content rather than curated data.

Singh et al. (2025) reviewed recent advances in AI algorithms for deepfake detection and concluded that hybrid CNN-transformer architectures offer improved generalisation, scalability and robustness relative to traditional models. Collectively these studies demonstrate that efficient hybrid frameworks provide significant improvements in feature representation, contextual analysis and classification accuracy.

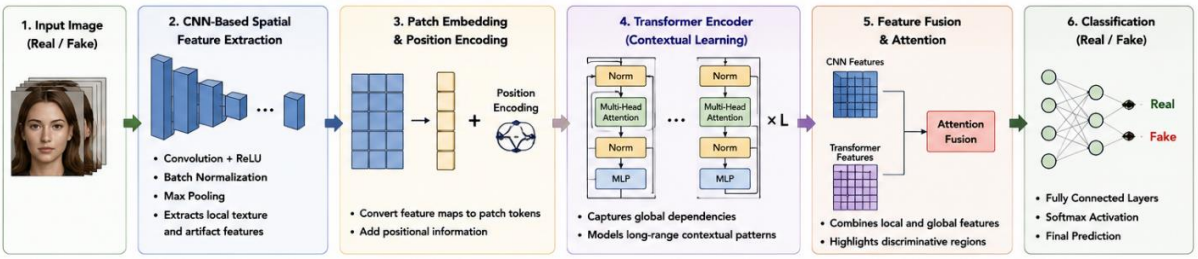

Figure 1: Efficient hybrid CNN-transformer architecture for image deepfake detection

Figure 1 illustrates the hybrid CNN-transformer architecture designed for efficient image-based deepfake detection. The framework includes image input, CNN-based spatial feature extraction, transformer-based contextual learning, feature fusion, attention mechanisms and final classification. It demonstrates how integrating CNNs and transformers improves local and global feature representation, detection robustness and classification accuracy while maintaining computational efficiency.

2. Benchmark-Driven Feature Analysis and Detection Techniques

Benchmark-driven evaluation and feature analysis are essential for developing reliable and generalised detection systems. Gamage et al. (2025) conducted a comparative study of detection models for misinformation prevention and demonstrated that benchmark-driven evaluation enables fair comparison of performance across datasets and manipulation scenarios.

Soundarya et al. (2025) highlighted that deep feature extraction enhances the capability of CNN-based models to identify manipulation traces and synthetic inconsistencies. Huang et al. (2025) proposed a multimodal analysis framework capable of localising manipulated regions and generating explainable outputs, showing that benchmark-driven evaluation combined with localisation improves interpretability and reliability.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Kaur et al. (2025) introduced a fusion-based framework using spectral features for audio deepfake detection. Although focused on audio, their work highlights the value of multimodal feature fusion and spectral representation learning, which may also benefit image-based systems. Lin et al. (2025) proposed a quantum-trained CNN for deepfake audio detection, indicating growing interest in computationally efficient and optimised architectures for synthetic media analysis.

Pham et al. (2025) provided a comprehensive survey of deepfake speech detection, emphasising the importance of benchmark datasets, feature engineering and adaptive learning frameworks. Collectively this theme indicates that benchmark-driven feature analysis and multimodal representation learning play a crucial role in scalable and generalised detection. It should be noted, however, that three of the four studies in this theme are drawn from the audio and speech domain, so their contribution to image-based detection is analogical rather than direct.

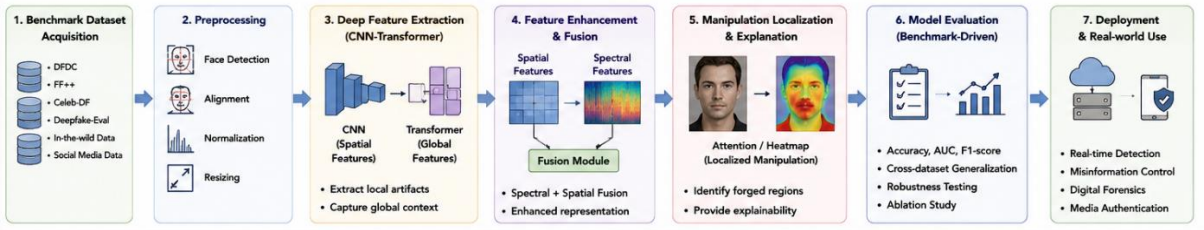

Figure 2: Benchmark-driven feature analysis and deepfake detection pipeline

Figure 2 presents the benchmark-driven detection workflow used for feature analysis and performance evaluation, including benchmark dataset acquisition, preprocessing, deep feature extraction, spectral and spatial feature fusion, multimodal representation learning, localisation of manipulated regions and model evaluation. It highlights the role of benchmark datasets and enhanced feature analysis in improving generalisation, interpretability and reliability.

3. Challenges, Cybersecurity Threats and Deployment Concerns

Several technical and societal challenges continue to affect the effectiveness of current systems. A major concern is the increasing sophistication of generation technologies, which evolve continuously to bypass existing detectors. Stephen (2025) discussed the role of artificial intelligence in investigating and preventing cybercrime and highlighted deepfake technology as a major threat to digital security and online trust.

Pedersen et al. (2025) examined deepfake-driven social engineering attacks in corporate environments and emphasised that manipulated visual and audio content can be exploited for identity fraud, phishing and misinformation campaigns, stressing the need for robust and explainable systems capable of operating in real-world cybersecurity settings. This study is valuable because it specifies a concrete deployment context with its own operational constraints, which most of the corpus leaves unstated.

Rosca et al. (2025) explored deepfake detection at the text level and demonstrated that multimodal manipulation involving images, text and speech creates new challenges for current frameworks. Huang et al. (2025) highlighted the importance of explainable AI for improving transparency and enabling localisation of manipulated regions.

Computational complexity and model efficiency present a further challenge. While transformer-based architectures achieve high accuracy, they require significant computational resources, limiting real-time deployment; Singh et al. (2025) noted that future frameworks should focus on lightweight architectures capable of maintaining performance under resource constraints. Benchmark inconsistency, lack of standardised protocols and poor cross-domain generalisation also remain unresolved, with Gamage et

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

al. (2025) reporting that many models perform well on specific datasets but fail to generalise to unseen manipulations and real-world scenarios.

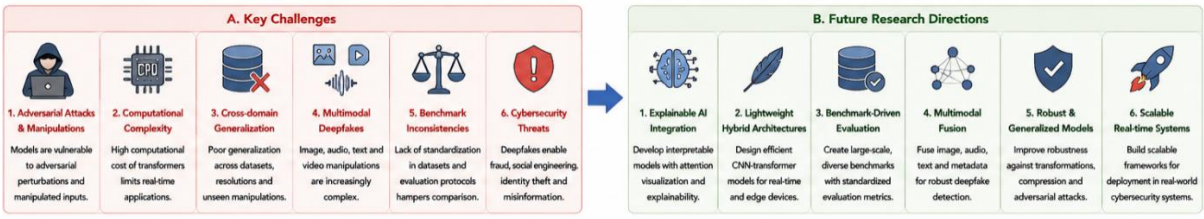

Figure 3: Challenges and future research directions in hybrid CNN-transformer deepfake detection frameworks

Figure 3 illustrates the major challenges and emerging directions in efficient hybrid CNN-transformer detection systems, highlighting adversarial attacks, computational complexity, cross-domain generalisation, multimodal deepfakes, benchmark inconsistency and cybersecurity threats. It also presents future directions including explainable AI integration, lightweight transformer architectures, benchmark-driven evaluation, multimodal fusion and scalable real-time detection frameworks.

4. Consolidated Summary of the Reviewed Literature

Table 1 organises the reviewed studies by theme, and Table 2 compares each study on research focus, approach, principal finding and the limitation it leaves unresolved. The limitation column reflects the assessment of this review rather than statements made by the original authors.

Table 1: Thematic organisation of the reviewed literature

Theme	Analytical focus	Studies	Count
T1. Hybrid CNN-transformer architectures	CNN feature extraction, spatio-temporal hybrids, transformer contextual learning and localisation	Soundarya et al. (2025); Zafar et al. (2025); Alsolai et al. (2025); Huang et al. (2025); Singh et al. (2025)	5
T2. Benchmark-driven feature analysis	Comparative benchmarking, spectral feature fusion and optimised architectures	Gamage et al. (2025); Kaur et al. (2025); Lin et al. (2025); Pham et al. (2025)	4
T3. Challenges and cybersecurity threats	Cybercrime prevention, social engineering in corporate settings and text-level manipulation	Stephen (2025); Pedersen et al. (2025); Rosca et al. (2025)	3

Table 2: Comparative summary of the reviewed studies

Study	Research focus	Approach	Principal reported finding	Limitation for efficient hybrid detection
Soundarya et al. (2025)	Image deepfake detection	CNN with deep feature extraction	CNNs capture spatial inconsistency and texture artifacts	Local features only; no global contextual modelling
Zafar et al. (2025)	Hybrid spatio-temporal detection	CNN spatial extraction with sequential learning	Hybrid design outperforms standalone CNN models	Temporal modelling requires video input
Alsolai et al. (2025)	Robust image detection	Guardian-AI hybrid learning model	Optimised extraction improves accuracy on realistic manipulations	Computational cost of the hybrid not reported

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Study	Research focus	Approach	Principal reported finding	Limitation for efficient hybrid detection
Huang et al. (2025)	Detection, localisation and explanation	SIDA large multimodal transformer model	Combined detection, localisation and explanation on social media images	Large multimodal model; heavy for real-time deployment
Singh et al. (2025)	AI algorithms for detection	Review of detection advances	Hybrid CNN-transformer models generalise and scale better	Synthesis only; no empirical comparison on common data
Gamage et al. (2025)	Model assessment for misinformation	Comparative benchmark study	Benchmark-driven evaluation enables fair comparison	Reports generalisation failure without proposing a remedy
Kaur et al. (2025)	Audio deepfake detection	Spectral feature fusion model	Spectral fusion improves audio detection performance	Audio domain; image transferability analogical only
Lin et al. (2025)	Efficient audio detection	Quantum-trained CNN	Optimised training reduces resource requirements	Audio domain; requires specialised training infrastructure
Pham et al. (2025)	Speech deepfake detection	Comprehensive survey with critical analysis	Benchmarks, feature engineering and adaptive learning are decisive	Speech domain; survey rather than method
Stephen (2025)	AI in cybercrime prevention	Investigative analysis	Deepfakes are a major threat to digital security and trust	Non-technical; no detection method or evaluation
Pedersen et al. (2025)	Social engineering in corporate settings	Threat, detection and defence analysis	Manipulated media enables fraud, phishing and misinformation	Threat-focused; detection performance not measured
Rosca et al. (2025)	Text-level deepfake detection	Text-level detection paradigm	Multimodal manipulation across text and media raises new challenges	Text domain; visual applicability indirect

Three observations follow from Table 2. First, the corpus is architecture-rich but cost-blind: no reviewed study reports computational cost, inference latency or parameter count alongside accuracy, even though efficiency is the leading term in this paper's title. Second, only four of the twelve studies operate directly on still images, which is a material constraint on how far the reviewed evidence supports conclusions about image detection and is examined in Table 4. Third, Huang et al. (2025) is the only study that addresses detection, localisation and explanation within a single system and evaluates on social media content, making it the closest existing analogue to the framework this review advocates.

5. Component-Level Comparison of CNN, Transformer and Hybrid Designs

Table 3 compares the three architectural approaches discussed across the corpus on the dimensions that determine practical deployability. The comparison is qualitative and reflects the relative positions reported in the reviewed studies rather than measured benchmark values, which the corpus does not report on a common basis.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 3: Comparison of CNN, transformer and hybrid CNN-transformer designs

Design	Local feature extraction	Global context modelling	Computational cost	Cross-domain generalisation	Interpretability
CNN	Strong on texture and spatial artifacts	Limited; short receptive range	Moderate	Weak on unseen generators	Low, black-box
Transformer	Weaker without convolutional front end	Strong via self-attention	High, data and memory intensive	Improved global consistency modelling	Moderate via attention maps
Hybrid CNN-transformer	Retained through convolutional stages	Added through attention stages	High unless optimised	Best reported in the corpus	Moderate, improved by fusion weighting
Hybrid with localisation and XAI	Retained	Retained	Highest; explanation adds overhead	Evaluated on social media content	High, region-level explanation

The table makes the design rationale explicit: hybridisation is motivated by complementary weaknesses rather than by accuracy alone, since CNNs and transformers fail in different places. It also makes the cost of that combination visible, which is why lightweight hybrid design rather than hybrid design as such is the open problem identified in Section 3.

6. Modality Coverage of the Reviewed Corpus

Because several reviewed studies address audio, speech or text rather than still images, Table 4 states the modality of each study and assesses how directly its findings transfer to image-based detection. This matters for a review whose conclusions concern the image domain specifically.

Table 4: Modality coverage and transferability to image-based detection

Study	Primary modality	Transferability to image-based deepfake detection
Soundarya et al. (2025)	Image	Direct; CNN feature extraction on still images
Alsolai et al. (2025)	Image	Direct; image detection model
Huang et al. (2025)	Image (multimodal model)	Direct; social media image detection, localisation and explanation
Gamage et al. (2025)	Image and video	Direct; benchmark comparison of detection models
Zafar et al. (2025)	Video (spatio-temporal)	Partial; spatial stream transfers, temporal stream requires sequences
Singh et al. (2025)	Cross-modal review	Partial; architectural conclusions transfer, empirical basis mixed
Kaur et al. (2025)	Audio	Analogical; spectral fusion principle only
Lin et al. (2025)	Audio	Analogical; efficiency-oriented training principle only
Pham et al. (2025)	Speech	Analogical; benchmarking and feature engineering principles
Rosca et al. (2025)	Text	Indirect; motivates multimodal framing
Pedersen et al. (2025)	Multimodal threat analysis	Contextual; establishes deployment requirements, not methods
Stephen (2025)	General cybersecurity	Contextual; motivation only

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Four studies transfer directly, two partially, three by analogy and three provide context or motivation only. The evidential base for image-specific claims is therefore narrower than a count of twelve implies, and future work in this area would benefit from a corpus weighted more heavily toward the image domain.

III. RESEARCH GAP

Although significant advances have been achieved in image deepfake detection, seven research gaps remain unresolved. They are stated below and mapped to corresponding future research directions in Table 5.

First, many current systems focus on improving classification accuracy without adequately addressing computational efficiency. Transformer-based architectures require high computational resources and memory, limiting applicability in real-time and resource-constrained environments.

Second, CNN-based models focus mainly on local spatial feature extraction and frequently fail to capture global contextual dependencies. While transformer frameworks improve contextual learning, many systems still lack effective feature fusion mechanisms capable of combining local and global representations efficiently.

Third, cross-domain generalisation remains poor. Many models achieve high performance on specific benchmark datasets but fail on unseen manipulations, social media images, compressed data or real-world content, largely because of overfitting and insufficient benchmark diversity.

Fourth, benchmark inconsistency and the lack of standardised evaluation protocols remain major challenges. Studies use different datasets, preprocessing strategies and performance metrics, making fair comparison and reproducibility difficult.

Fifth, detection systems remain vulnerable to adversarial attacks, image transformations and multimodal manipulations. Compression, resizing, cropping, noise injection and manipulation-aware perturbations can substantially reduce performance and reliability.

Sixth, many frameworks operate as black-box systems with limited explainability and interpretability, reducing user trust and limiting applicability in digital forensics and cybersecurity domains.

Finally, limited research integrates multimodal fusion, explainable AI and lightweight transformer architectures within unified hybrid frameworks. Future research should therefore develop efficient hybrid CNN-transformer models integrating benchmark-driven evaluation, explainable AI, multimodal feature fusion and computational optimisation.

Table 5: Research gaps mapped to corresponding future research directions

No.	Identified research gap	Corresponding future research direction
G1	Accuracy prioritised over computational efficiency	Lightweight hybrid architectures reporting latency, memory and parameter count alongside accuracy (Section 5.1)
G2	Ineffective fusion of local and global representations	Feature fusion designs evaluated for the contribution of each stream (Sections 5.1 and 5.6)
G3	Poor cross-domain generalisation	Generalised feature learning and cross-domain adaptation on real-world content (Section 5.6)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

No.	Identified research gap	Corresponding future research direction
G4	Benchmark and evaluation inconsistency	Standardised datasets, metrics and protocols for reproducible comparison (Section 5.2)
G5	Vulnerability to adversarial attacks and transformations	Transformation-resilient training under compression, resizing and perturbation (Section 5.5)
G6	Limited explainability and interpretability	Explainable AI with manipulation localisation for forensic use (Section 5.3)
G7	Multimodal fusion, XAI and lightweight design not integrated	Unified multimodal frameworks combining image, audio, video and text (Sections 5.4 and 5.1)

IV. CONCLUSION

Deepfake technology has emerged as a significant threat to digital media authenticity, cybersecurity and public trust because of the increasing realism of AI-generated synthetic images. This review presented a comprehensive analysis of efficient hybrid CNN-transformer models for benchmark-driven image deepfake detection, examining CNN-based architectures, transformer-based contextual learning, feature fusion strategies, benchmark-driven evaluation methods and explainable AI techniques.

The review found that hybrid CNN-transformer architectures provide superior performance by combining local spatial feature extraction with global contextual learning, and that benchmark-driven feature analysis, multimodal learning, explainable AI and manipulation localisation further improve robustness, interpretability and scalability. Comparing the designs component by component also clarified the rationale: hybridisation is motivated by complementary failure modes rather than by accuracy alone, since CNNs and transformers fall short in different places.

Three further observations emerged from the comparative synthesis. No reviewed study reports computational cost alongside accuracy, so efficiency claims in this field remain unverifiable despite efficiency being the leading term in this review's own framing. Only four of the twelve studies operate directly on still images, so the evidential base for image-specific claims is narrower than the corpus size implies. And Huang et al. (2025) stands out as the only study addressing detection, localisation and explanation within one system evaluated on circulated social media content.

Despite these advances, challenges remain unresolved including computational complexity, adversarial vulnerability, poor cross-domain generalisation, benchmark inconsistency and limited real-time deployment capability. Future systems should prioritise lightweight hybrid architectures, generalised feature learning, benchmark standardisation and multimodal fusion.

Future Work

Lightweight Hybrid CNN-Transformer Architectures

Future research should develop computationally efficient hybrid architectures suitable for real-time and edge-device deployment without compromising detection accuracy. Given that no study in this corpus reports inference cost, future work should report latency, memory footprint and parameter count alongside accuracy as standard practice.

Benchmark-Driven Standardised Evaluation

Standardised benchmark datasets, evaluation metrics and testing protocols should be developed to improve reproducibility, scalability and fair comparison, with evaluation on circulated real-world content treated as a primary result rather than a supplementary check.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Explainable AI and Manipulation Localisation

Integrating explainable AI and manipulation localisation can improve transparency, interpretability and forensic reliability. Localisation is particularly valuable because it provides evidence at region level rather than a single authenticity score.

Multimodal Deepfake Detection Frameworks

Future systems should integrate image, audio, video and textual analysis within unified multimodal frameworks to improve robustness against complex synthetic manipulations, drawing on the spectral and semantic representation principles established in the audio and text literature.

Adversarially Robust Detection Systems

Research should explore adversarial defence mechanisms and transformation-resilient learning strategies capable of maintaining performance under compression, resizing, cropping and perturbation attacks representative of real distribution channels.

Generalised Feature Learning and Cross-Domain Adaptation

Future studies should focus on adaptive feature fusion and generalised learning approaches capable of detecting unseen generation techniques across diverse real-world datasets, with the contribution of the local and global streams reported separately rather than only in aggregate.

REFERENCES

1. Soundarya, B. C., Gururaj, H. L., & Naveen Kumar, C. M. (2025). A framework for deepfake detection using convolutional neural network and deep features. Procedia Computer Science, 258, 3640-3648.
2. Lin, C.-H. A., Liu, C.-Y., Chen, S. Y.-C., & Chen, K.-C. (2025). Quantum-trained convolutional neural network for deepfake audio detection. In IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) (pp. 1-5).
3. Stephen, G. (2025). Investigation and prevention of cybercrimes using artificial intelligence.
4. Rosca, C.-M., Stancu, A., & Iovanovici, E. M. (2025). The new paradigm of deepfake detection at the text level. Applied Sciences, 15(5), 2560.
5. Kaur, N., Dixit, A., & Kingra, S. (2025). A deep learning fusion model leveraging spectral features for audio deepfake detection. International Journal of Advanced Networking and Applications, 16(5), 6551-6560.
6. Zafar, F., Khan, T. A., Akbar, S., Ubaid, M. T., Javaid, S., & Kadir, K. (2025). A hybrid deep learning framework for deepfake detection using temporal and spatial features. IEEE Access.
7. Singh, L. H., Charanarur, P., & Chaudhary, N. K. (2025). Advancements in detecting deepfakes: AI algorithms and future prospects - a review. Discover Internet of Things, 5(1), 1-30.
8. Gamage, C. L., Isoaho, J., & Mohammad, T. (2025). Assessing deepfake detection models: A comparative study for misinformation detection and prevention.
9. Alsolai, H., Mahmood, K., Alshuhail, A., Ben Miled, A., Alqahtani, M., Alshareef, A., Alallah, F. S., & Alghamdi, B. M. (2025). Guardian-AI: A novel deep learning based deepfake detection model in images. Alexandria Engineering Journal, 126, 507-514.
10. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., & Cheng, G. (2025). SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 28831-28841).
11. Pedersen, K. T., Pepke, L., Staermose, T., Papaioannou, M., Choudhary, G., & Dragoni, N. (2025). Deepfake-driven social engineering: Threats, detection techniques, and defensive strategies in corporate environments. Journal of Cybersecurity and Privacy, 5(2), 18.
12. Pham, L., Lam, P., Tran, D., Tang, H., Nguyen, T., Schindler, A., Skopik, F., Polonsky, A., & Vu, H. C. (2025). A comprehensive survey with critical analysis for deepfake speech detection. Computer Science Review, 57, 100757.