International Conference on Social Science & Humanities Trends

International Journal for Research Trends in Social Science & Humanities

Aug 2026

ISSN: 2584-2455

Advanced Preprocessing and Hybrid Neural Architecture for High-Accuracy Image Deepfake Detection

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 widespread adoption of generative artificial intelligence has accelerated the creation of highly realistic deepfake images, creating major concerns regarding digital security, misinformation, identity fraud and media authenticity. Traditional detection approaches often struggle to maintain robustness against evolving synthetic image generation techniques and real-world image transformations. This review paper presents a comprehensive analysis of advanced preprocessing strategies and hybrid neural architectures for high-accuracy image deepfake detection. The study critically examines state-of-the-art frameworks including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), attention-based models and hybrid architectures such as GenConViT, and investigates the role of preprocessing techniques including normalisation, facial alignment, edge enhancement, frequency-domain transformation and feature optimisation. Eleven studies are reviewed and organised into three themes covering deep learning architectures, advanced preprocessing and feature enhancement, and challenges with generalisation. The reviewed evidence is consolidated into comparative tables that map each study to its focus, approach, principal finding and limitation, alongside a comparison of architecture families and a summary of preprocessing techniques. Challenges such as adversarial attacks, cross-dataset generalisation, zero-shot detection limitations and computational complexity are discussed, and seven research gaps are identified and mapped to corresponding future research directions. Directions focusing on explainable AI, lightweight hybrid models, generalised feature learning and multimodal detection are presented to support the development of scalable, interpretable and computationally efficient deepfake detection systems.

Keywords- Deepfake detection; hybrid neural networks; vision transformers; convolutional neural networks; image forensics; feature enhancement; advanced preprocessing; explainable AI; computational efficiency; artificial intelligence.

I. Introduction

Recent advances in artificial intelligence and deep generative models have transformed digital media creation. Generative Adversarial Networks (GANs), diffusion models and encoder-decoder architectures enable the generation of highly realistic synthetic images commonly referred to as deepfakes. Although these technologies have beneficial applications in entertainment, virtual reality, education and healthcare, their misuse has created severe challenges relating to misinformation, cybercrime, digital impersonation and political manipulation.

Deepfake images have become increasingly difficult to distinguish from authentic content as generative modelling improves. The rapid spread of manipulated media across online platforms has raised significant concerns about media integrity and public trust, making automated and reliable detection a critical research area in artificial intelligence and digital forensics.

Traditional detection approaches relied on handcrafted features, statistical analysis and shallow machine learning, but demonstrated limited effectiveness against sophisticated synthetic images produced by modern AI techniques. Researchers have therefore adopted deep learning frameworks capable of automatically extracting complex visual patterns and manipulation artifacts.

CNNs have been extensively used for deepfake detection because of their strong spatial feature extraction capability. More recently, transformer-based architectures and hybrid networks integrating CNNs, Vision Transformers and attention mechanisms have demonstrated superior performance in capturing both local and global contextual features. Hybrid frameworks such as GenConViT further improve accuracy and robustness by combining discriminative feature learning with contextual representation modelling.

Preprocessing and feature enhancement have also become essential components of modern detection systems. Image normalisation, facial alignment, illumination correction, edge enhancement and frequency-domain analysis improve feature representation quality and assist networks in identifying subtle synthetic artifacts. Advanced preprocessing is particularly important for handling compressed, noisy or transformed images encountered in real-world deployment.

Despite substantial progress, current systems continue to face challenges including poor cross-dataset generalisation, vulnerability to adversarial attacks, computational complexity, lack of interpretability and reduced robustness against emerging generative models. There is therefore an increasing need for hybrid architectures that integrate advanced preprocessing, explainable AI and computationally efficient learning strategies.

1. Objectives and Contributions of this Review

This review provides a comprehensive study of advanced preprocessing techniques and hybrid neural architectures for high-accuracy image deepfake detection. It makes four contributions. First, it consolidates eleven recent studies into a structured thematic review supported by comparative tables rather than narrative description alone. Second, it compares the principal architecture families on the dimensions that determine practical deployability, namely accuracy, generalisation, computational cost, transformation resilience and interpretability. Third, it summarises the preprocessing and feature enhancement techniques reported across the corpus and states the purpose each serves. 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. Deep Learning Architectures for Image Deepfake Detection

The advancement of generative artificial intelligence has increased the complexity and realism of deepfake images, creating substantial challenges for existing detection systems. Hopf et al. (2026) presented the NTIRE 2026 Challenge report, emphasising the importance of detection frameworks capable of maintaining performance under diverse real-world image distortions and adversarial manipulations, and reported that models with adaptive feature extraction achieve improved robustness relative to conventional CNN approaches. As a challenge report, this study carries particular evidential weight because it evaluates many competing methods under a single controlled protocol.

Gupta et al. (2026) conducted a taxonomical and systematic review categorising existing approaches into spatial-domain, frequency-domain and hybrid neural network models, and demonstrated that hybrid architectures combining CNNs, Vision Transformers and attention mechanisms outperform standalone models on high-quality synthetic images. Gunasekaran et al. (2026) introduced GenConViT, a hybrid framework integrating CNNs with transformer-based learning for improved feature extraction and classification, achieving higher accuracy and better generalisation across multiple benchmark datasets.

Soundarya and Gururaj (2026) critically reviewed state-of-the-art approaches and identified that transformer-based architectures have gained attention because of their ability to capture global contextual relationships within images. Abdel-Wahab and Alkhatib (2026) emphasised that robust defence frameworks should integrate multiple learning paradigms, including feature fusion and attention-guided architectures, to improve resilience against sophisticated manipulations. These studies collectively indicate that hybrid architectures represent the most promising direction for high-accuracy detection.

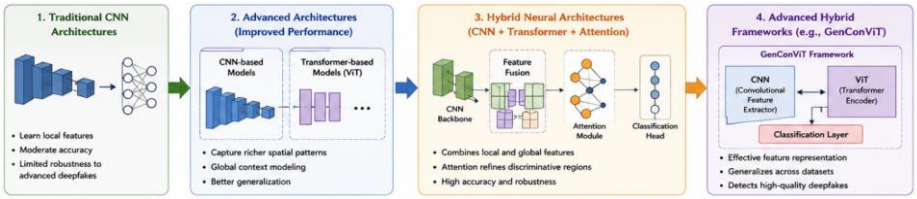

Figure 1: Hybrid deep learning architectures for high-accuracy image deepfake detection

Figure 1 illustrates the evolution and integration of deep learning architectures used for image deepfake detection, presenting traditional CNN models, transformer-based architectures and hybrid frameworks such as GenConViT that combine CNNs with Vision Transformers and attention mechanisms. It highlights how hybrid architectures improve feature extraction, contextual learning, robustness and classification accuracy for detecting sophisticated synthetic images.

2. Advanced Preprocessing and Feature Enhancement Techniques

Preprocessing and feature enhancement play a critical role in the performance and reliability of image-based detection systems. Normalisation, artifact enhancement, edge analysis and frequency-domain transformation improve the discriminative capability of deep models. Hopf et al. (2026) highlighted that robust preprocessing pipelines contribute to improved detection performance under compression artifacts, resizing and image transformations commonly encountered in practice.

Sharma et al. (2026) discussed the importance of facial alignment, noise filtering, illumination correction and texture enhancement in improving feature extraction quality for face deepfake detection, and demonstrated that effective preprocessing enables networks to identify subtle inconsistencies in synthetic facial images. Soundarya and Gururaj (2026) similarly emphasised that advanced preprocessing improves robustness by reducing irrelevant noise and improving feature representation. Yermakov et al. (2026) investigated models capable of generalising across multiple benchmark datasets, and found that preprocessing consistency and feature normalisation significantly influence adaptability and cross-domain performance. This is an important result for the present review because it identifies preprocessing as a determinant of generalisation rather than merely of accuracy, which is a stronger claim than the other preprocessing studies make.

Sar et al. (2026) introduced zero-shot visual deepfake detection approaches that aim to identify manipulated content without prior exposure to specific generation techniques, highlighting the importance of generalised feature extraction and semantic representation learning. Yigitalp Tulga (2026) evaluated web-based detection tools and identified limitations related to preprocessing inconsistency and reduced reliability under varying image quality. Collectively these findings demonstrate that preprocessing and feature enhancement are essential for robustness, scalability and real-world applicability.

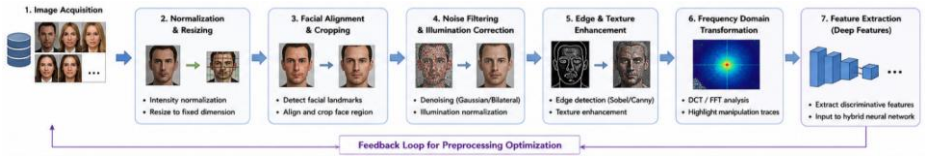

Figure 2: Advanced preprocessing and feature enhancement pipeline for deepfake detection

Figure 2 presents the preprocessing workflow employed in modern image deepfake detection systems, including image acquisition, normalisation, facial alignment,

resizing, noise filtering, illumination correction, edge enhancement, texture analysis and feature extraction. It demonstrates how advanced preprocessing improves feature representation quality, reduces noise interference and enhances the ability of models to identify subtle manipulation artifacts.

3. Challenges, Generalisation and Robustness

Several critical challenges continue to limit the effectiveness of current detection systems. A primary concern is the rapid evolution of generative AI, which continuously produces more realistic images capable of bypassing existing detectors. Raza (2026) argued that many current detectors are optimised for outdated manipulation techniques and may fail to identify emerging synthetic content produced by advanced diffusion and generative models. This study is the most contrarian in the corpus, questioning whether detection research has been calibrated to the threat as it actually developed, and it deserves attention precisely because every other reviewed study assumes the threat model is settled.

Generalisation across datasets and unseen manipulation methods remains a major challenge. Yermakov et al. (2026) proposed generalised detection strategies capable of maintaining consistent performance across benchmarks, and found that many models achieve high accuracy in controlled environments but fail under cross-domain testing because of overfitting and limited adaptability.

Sar et al. (2026) introduced zero-shot visual deepfake detection, in which models attempt to detect manipulated content without training on the specific generation method, representing an important step toward generalised and future-proof systems. Abdel-Wahab and Alkhatib (2026) highlighted the need for adversarially robust frameworks capable of resisting manipulation-aware attacks and media transformations.

The reliability of publicly available detection tools is a further concern. Yigitalp Tulga (2026) reported that many web-based systems exhibit inconsistent performance and limited transparency, reducing their effectiveness in practical cybersecurity and forensic settings. Luong et al. (2026) emphasised that media transformations such as compression, noise injection and format conversion can significantly degrade detection performance, establishing the case for transformation-resilient models. Together, these studies indicate that future research should focus on lightweight hybrid architectures, explainable detection, generalised feature learning and benchmark-driven evaluation protocols.

Figure 3: Challenges and future research directions in robust deepfake detection

Figure 3 illustrates the major challenges and emerging directions in image deepfake detection, highlighting adversarial attacks, dataset generalisation, zero-shot detection challenges, media transformations, computational complexity and reliability limitations of existing tools. It also presents future directions including explainable AI, zero-shot learning, lightweight hybrid architectures, benchmark-driven evaluation and transformation-resilient 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. Deep learning architectures for detection	CNN, transformer and hybrid CNN-ViT architectures, taxonomies and defence frameworks	Hopf et al. (2026); Gupta et al. (2026); Gunasekaran et al. (2026); Soundarya & Gururaj (2026); Abdel-Wahab & Alkhatib (2026)	5
T2. Advanced preprocessing and feature enhancement	Facial alignment, normalisation, texture and edge enhancement, and preprocessing consistency for generalisation	Sharma et al. (2026); Yermakov et al. (2026); Sar et al. (2026); Yigitalp Tulga (2026)	4
T3. Challenges, robustness and transformation resilience	Threat-model calibration and degradation under media transformation	Raza (2026); Luong et al. (2026)	2

Table 2: Comparative summary of the reviewed studies

Study	Research focus	Approach	Principal reported finding	Limitation for robust efficient detection
Hopf et al. (2026)	Robust detection under distortion	NTIRE 2026 challenge benchmark report	Adaptive feature extraction improves robustness over plain CNNs	Challenge conditions may not reflect deployment distributions
Gupta et al. (2026)	Taxonomy of detection approaches	Systematic review of spatial, frequency and hybrid models	Hybrid CNN-ViT-attention models outperform standalone models	Synthesis only; no empirical validation or cost analysis
Gunasekaran et al. (2026)	Hybrid architecture design	GenConViT combining CNN and transformer learning	Higher accuracy and better generalisation across benchmarks	Computational cost of the hybrid not reported

Study	Research focus	Approach	Principal reported finding	Limitation for robust efficient detection
Soundarya & Gururaj (2026)	State-of-the-art review	Critical review of detection approaches	Transformers capture global context effectively	Descriptive; no comparative evaluation on common data
Abdel-Wahab & Alkhatib (2026)	Robust defence frameworks	Review of detection and prevention techniques	Multi-paradigm fusion improves resilience to manipulation	Prescriptive; proposed integration not implemented
Sharma et al. (2026)	Face deepfake detection	Review of preprocessing and feature extraction	Alignment, filtering and texture enhancement improve extraction	Preprocessing effects not isolated or quantified
Yermakov et al. (2026)	Cross-benchmark generalisation	Generalised detection with consistent preprocessing	Preprocessing consistency drives cross-domain adaptability	Generalisation gains bounded by available benchmark diversity
Sar et al. (2026)	Zero-shot detection	Generalised feature and semantic representation learning	Detection without prior exposure to the generation method is feasible	Zero-shot accuracy trails supervised detection on known manipulations
Yigitalp Tulga (2026)	Reliability of public tools	Evaluation of web-based detection services	Public tools show inconsistent performance and low transparency	Evaluates tools, not the underlying architectures
Raza (2026)	Threat-model calibration	Critical analysis of detector development	Detectors are optimised for manipulation types that did not dominate	Position argument; offers critique rather than a detection method
Luong et al. (2026)	Robustness under media transformation	RADAR 2026 challenge on transformed media	Compression, noise and format conversion degrade performance	Audio-domain focus; image transferability untested

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 computational efficiency appears in this paper's own title as a design objective. Second, five of the eleven studies are reviews, taxonomies or position pieces rather than original detection methods, which means the field's synthesis activity currently outweighs its methodological output. Third, Raza (2026) stands apart in questioning whether the threat model itself was correctly specified, a challenge that no other study in the corpus engages with and that has direct implications for how detection performance should be evaluated.

5. Comparison of Architecture Families

Table 3 compares the principal architecture families discussed across the corpus on the five 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.

Table 3: Comparison of neural architecture families for image deepfake detection

Architecture family	Detection accuracy	Cross-dataset generalisation	Computational cost	Transformation resilience	Interpretability
Conventional CNN	Effective on known manipulations	Weak on unseen generators	Moderate	Degrades under compression and resizing	Low, black-box
Frequency-domain model	Strong on generator-specific artifacts	Sensitive to post-processing	Low to moderate	Vulnerable to recompression	Moderate, spectral evidence
Vision Transformer	Strong global context modelling	Improved over plain CNN	High, data-hungry	Better under mild distortion	Low to moderate via attention maps
Hybrid CNN-transformer (e.g. GenConViT)	Highest reported in the corpus	Best reported, still benchmark-sensitive	High unless optimised	Improved through combined local and global cues	Moderate via attention weighting
Zero-shot and generalised framework	Below supervised models on known types	Designed for unseen manipulations	Varies with representation backbone	Depends on semantic feature stability	Moderate, semantic rationale

The table makes the central design tension explicit: the families delivering the best accuracy and transformation resilience are also the most computationally demanding, while the family designed for generalisation trades accuracy on known manipulations to obtain it. Lightweight hybrid design is therefore not a refinement but the principal open problem identified in Section 3.

6. Summary of Reported Preprocessing Techniques

Table 4 summarises the preprocessing and feature enhancement techniques reported across the corpus, together with the purpose each serves in the detection pipeline.

Table 4: Preprocessing and feature enhancement techniques reported in the reviewed literature

Technique	Purpose in the detection pipeline	Reported in
Normalisation and resizing	Standardising input scale and intensity for stable feature learning	Hopf et al. (2026); Yermakov et al. (2026)
Facial alignment	Registering facial geometry so manipulation traces occupy consistent positions	Sharma et al. (2026)
Noise filtering	Suppressing acquisition noise so subtle artifacts remain distinguishable	Sharma et al. (2026); Soundarya & Gururaj (2026)
Illumination correction	Removing lighting variation that masks blending inconsistencies	Sharma et al. (2026)
Edge enhancement	Exposing boundary irregularities introduced by facial compositing	Soundarya & Gururaj (2026)
Texture analysis	Capturing surface statistics that differ between real and synthesised skin	Sharma et al. (2026)
Frequency-domain transformation	Revealing generator-specific spectral signatures not visible spatially	Gupta et al. (2026)
Preprocessing consistency across datasets	Preventing pipeline mismatch from being mistaken for domain shift	Yermakov et al. (2026)
Generalised semantic representation	Supporting detection of manipulation types absent from training	Sar et al. (2026)

III. Research Gap

Although considerable 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 models prioritise detection accuracy without adequately addressing computational efficiency. Advanced transformer-based architectures and deep hybrid networks require substantial computational resources and memory, limiting their applicability in real-time and resource-constrained environments.

Second, cross-dataset generalisation remains a major challenge. Many systems perform effectively on specific benchmark datasets but fail when exposed to unseen generation techniques, image transformations or real-world manipulated content, frequently suffering from overfitting and limited adaptability.

Third, existing preprocessing techniques are often inconsistent and insufficiently optimised. Many studies focus on model architecture while overlooking preprocessing operations such as facial alignment, artifact enhancement, edge analysis and frequency-domain extraction that significantly influence detection performance.

Fourth, detection systems remain vulnerable to adversarial attacks and media transformations. Compression, resizing, noise injection, cropping and format conversion can substantially degrade performance, reducing reliability in practical deployment.

Fifth, most deep learning systems function as black-box models with limited explainability. This restricts user trust and reduces effectiveness in sensitive applications such as digital forensics, legal investigation and cybersecurity operations. Sixth, there is limited research on zero-shot and generalised detection capable of identifying previously unseen manipulation techniques without explicit training.

Finally, benchmark evaluation methodologies lack standardisation, making fair comparison and reproducibility difficult across studies.

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	Poor cross-dataset generalisation	Generalised feature learning evaluated across diverse benchmarks (Sections 5.3 and 5.6)
G3	Inconsistent and unoptimised preprocessing	Adaptive preprocessing pipelines with effects isolated and quantified (Section 5.5)
G4	Vulnerability to adversarial attacks and media transformations	Transformation-resilient models tested under compression, noise and format conversion (Section 5.4)
G5	Limited explainability and interpretability	XAI-integrated detection with region-level explanation for forensic use (Section 5.2)
G6	Limited zero-shot and generalised detection	Zero-shot learning and semantic representation for unseen manipulations (Section 5.3)
G7	Benchmark and evaluation inconsistency	Standardised protocols, datasets and metrics, extended to multimodal detection (Section 5.6)

IV. Conclusion

The increasing sophistication of deepfake generation technologies has created significant challenges for digital media authentication and cybersecurity. This review presented a comprehensive analysis of advanced preprocessing techniques and hybrid neural architectures for high-accuracy image deepfake detection, examining CNN-

based models, Vision Transformers, attention-guided architectures, hybrid frameworks such as GenConViT and advanced preprocessing pipelines.

The review found that hybrid frameworks integrating local and global feature learning outperform conventional standalone models in accuracy, robustness and generalisation, and that preprocessing techniques including normalisation, facial alignment, edge enhancement and frequency-domain analysis materially improve detection performance. Comparing architecture families on a common set of dimensions also made the central design tension explicit: the families delivering the best accuracy and transformation resilience are the most computationally demanding, while the family designed for generalisation trades accuracy on known manipulations to obtain it.

Three further observations emerged from the comparative synthesis. No reviewed study reports computational cost alongside accuracy, so efficiency claims in this field are currently unverifiable. Five of the eleven studies are reviews, taxonomies or position pieces rather than original methods, indicating that synthesis activity currently outweighs methodological output. And Raza (2026) stands alone in questioning whether the threat model itself was correctly specified, a challenge with direct implications for how detection performance should be evaluated.

Despite these advances, challenges remain unresolved including adversarial vulnerability, computational overhead, poor cross-domain generalisation and limited explainability. Future detection systems must combine advanced preprocessing, lightweight hybrid architectures, explainable AI techniques and generalised feature learning to achieve robust, scalable and computationally efficient image deepfake detection.

Future Work

Lightweight Hybrid Neural Architectures

Future research should design computationally efficient hybrid models that maintain high detection accuracy while reducing memory and processing requirements for real-time applications. 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.

Explainable AI-Based Deepfake Detection

Integrating explainable artificial intelligence techniques can improve model transparency, interpretability and trustworthiness in digital forensic and cybersecurity applications, where detection outputs may be used as evidence.

Zero-Shot and Generalised Deepfake Detection

Future systems should emphasise zero-shot learning and generalised feature extraction capable of detecting unseen and evolving generation methods, with cross-domain performance reported as a primary result rather than an ancillary observation.

Transformation-Resilient Detection Frameworks
Research should explore robust models capable of maintaining performance under compression, resizing, cropping, noise injection and adversarial media transformations representative of real distribution channels.

Advanced Preprocessing Optimisation
Further studies should investigate adaptive preprocessing pipelines involving frequency-domain analysis, edge enhancement, texture analysis and artifact amplification, and should isolate the contribution of each stage rather than reporting preprocessing only as a supporting step.

Benchmark-Driven Evaluation and Multimodal Detection
Standardised evaluation frameworks and multimodal detection systems integrating image, video and audio analysis should be developed to enhance scalability and real-world applicability.

References

1. Hopf, B., Timofte, R., Qu, C., Li, J., Wu, F., Lu, D., & Shi, Y. (2026). Robust deepfake detection, NTIRE 2026 challenge: Report. arXiv preprint arXiv:2604.24163.
2. Gupta, P., Narwal, B., & Mohapatra, A. K. (2026). Combating digital deception: A taxonomical and systematic review of deepfake detection approaches. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16(2), e70087.
3. Raza, S. (2026). The deepfakes we missed: We built detectors for a threat that didn't arrive. arXiv preprint arXiv:2605.12075.
4. Yigitalp Tulga, A. (2026). Evaluating web-based deepfake detection tools: Risks, limits, and practical solutions for information security. EDPACS, 71(5), 93-104.
5. Soundarya, B. C., & Gururaj, H. L. (2026). Deepfake detection: Critical review of state-of-the-art approaches and future perspectives. Discover Applied Sciences.
6. Abdel-Wahab, A., & Alkhatib, M. (2026). Toward robust deepfake defense: A review of deepfake detection and prevention techniques in images. Computers, Materials & Continua, 86(2), 1.
7. Sar, A., Roy, S., Choudhury, T., & Abraham, A. (2026). Zero-shot visual deepfake detection: Can AI predict and prevent fake content before it is created? Foundations and Trends in Signal Processing, 19(3), 212-361.
8. Luong, H. T., Liu, X., Kukanov, I., Chai, Z. X., & Lee, K. A. (2026). RADAR Challenge 2026: Robust audio deepfake recognition under media transformations. arXiv preprint arXiv:2605.09568.
9. Sharma, D., Singh, S. P., Jain, S., Jain, A., Kumar, R., & Pratap, P. (2026). Towards face deepfake detection: A review with research challenges and future direction. In Innovative Computing and Communications: Proceedings of ICICC 2025, Volume 8 (pp. 321-327).
10. Yermakov, A., Cech, J., Matas, J., & Fritz, M. (2026). Deepfake detection that generalizes across benchmarks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 773-783).

11. Gunasekaran, R., Rao, P. N., Kumar, G. S., Kambhampati, K., Himabindu, B., & Talluri, T. (2026). Deepfake detection using GenConViT. In Computational Techniques and Smart Manufacturing (pp. 668-676). CRC Press.