iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82188

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Deepfake Detection Using Deep Learning: A Systematic Review, Taxonomy, and Emerging Challenges

Prajwal Patekar[1], Alisha Shaikh[2], Chaitanya Shinde[3], Aayush Singh[4], Bhakti Patel[5], Dr. Asha Durafe[6], Dr. Manisha Mane[7], Dr. Nandkishor Narkhede[8]
Department Of Electronics & Computer Science, Shah and Anchor Kutchhi Engineering College, Mumbai,India

Abstract: ***Deepfake technology, fuelled by rapid advances in generative adversarial networks (GANs) and diffusion models, has introduced unprecedented challenges to digital media authenticity and public trust. This paper presents a comprehensive, systematic review of deep-learning-based deepfake detection techniques, synthesising findings from more than 100 peer-reviewed publications spanning 2018–2026. We propose a unified taxonomy that organises detection approaches across three axes: spatial, temporal, and multi-modal. Key architectures examined include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM), Vision Transformers (ViTs), Generative Adversarial Networks repurposed for detection, diffusion-model fingerprinting, and ensemble hybrid frameworks. For each category we critically evaluate detection accuracy, generalisation ability, and computational overhead, drawing on standardised benchmarks such as FaceForensics++, Celeb-DF v2, and the DeepFake Detection Challenge (DFDC). We further discuss the persistent open problems — cross-dataset generalisation, adversarial robustness, explainability, and real-time deployment — and outline a multi-stakeholder mitigation ecosystem involving platform responsibility, community-driven verification, and legislative governance. This review is intended to serve as a reference for researchers and practitioners working to safeguard the integrity of digital media.***

Index Terms: ***Deepfake detection, Convolutional Neural Network, GAN, Vision Transformer, multi-modal analysis, digital media forensics, explainable AI.***

I. INTRODUCTION

The proliferation of sophisticated artificial intelligence has given rise to a category of synthetic media commonly referred to as deepfakes — hyper-realistic fabrications in which a person's likeness, voice, or mannerisms are reconstructed from data rather than recorded from life. The very word blends 'deep learning' with 'fake', capturing the technological substrate that makes such manipulation possible at scale. Since Goodfellow and colleagues introduced Generative Adversarial Networks in 2014, the pace of improvement has been remarkable: what once required expensive studio resources can today be accomplished by anyone with a consumer GPU and an open-source tool. The societal stakes are high. Deepfakes have already been weaponised in political campaigns — deepfake audio clips in Tamil Nadu's 2023 elections forced a minister's resignation, while manipulated videos in the 2020 Delhi assembly polls spread misinformation through WhatsApp at scale. Beyond elections, deepfakes threaten financial fraud, identity theft, non-consensual intimate imagery, and the systematic erosion of trust in video evidence.

Detection research has advanced in tandem with generation, but the arms-race dynamic is asymmetric: generating a convincing deepfake is generally easier and cheaper than detecting one. Classical forensic approaches that relied on hand-crafted features — noise residuals, compression artifacts, lighting inconsistencies — have been steadily outpaced. This paper makes three main contributions: (1) a comprehensive taxonomy of deepfake detection approaches that classifies techniques by modality and architecture; (2) a critical, evidence-based comparison of state-of-the-art models across benchmark datasets; and (3) a forward-looking discussion of open research challenges and a multi-stakeholder ecosystem framework.

II. BACKGROUND AND TECHNICAL FOUNDATIONS

A. How Deepfakes Are Created

Deepfake creation begins with the autoencoder — an encoder that compresses a face into a compact latent vector, and a decoder that reconstructs it.

The seminal insight was to share the encoder across two identities while using separate decoders: during inference, Identity A's compressed representation is fed to Identity B's decoder, yielding a reconstructed face with B's geometry but A's expression and movement. Generative Adversarial Networks extended the palette considerably. A GAN pits a generator network against a discriminator in an adversarial game. Landmark GAN variants — StyleGAN, ProGAN, DCGAN — can now synthesise high-resolution portrait images that fool trained human observers more than 50% of the time.

Most recently, diffusion models have emerged as a formidable alternative. Rather than an adversarial game, a diffusion model learns to reverse a noise-addition process: it progressively denoises random Gaussian noise into coherent images. Compared with GANs, diffusion models are more stable to train and produce higher diversity, though they demand substantially more inference compute.

B. Core Deep Learning Architectures for Detection

Convolutional Neural Networks (CNNs) remain the bedrock of spatial deepfake detection. Their hierarchical feature extraction naturally captures texture anomalies, blending artifacts, and frequency-domain inconsistencies. Architectures such as XceptionNet and EfficientNet-B7 have demonstrated particularly strong performance on benchmark tasks.

Recurrent Neural Networks and their gated variants — LSTM and GRU — address the temporal dimension that pure image-level CNNs overlook. By processing video as a sequence of frame-level feature vectors, RNN-based models capture inconsistencies in motion dynamics.

Vision Transformers (ViTs) have more recently risen to prominence. Unlike CNNs, which apply fixed local kernels, transformers process entire sequences of image patches simultaneously, modelling long-range spatial dependencies through self-attention, making ViTs especially sensitive to subtle global inconsistencies.

Hybrid frameworks that fuse CNN spatial encoders with temporal models (CNN-RNN, CNN-LSTM, CNN-Transformer) consistently outperform single-architecture baselines, achieving accuracy values in the 95–98% range on Celeb-DF v2.

III. PROPOSED TAXONOMY OF DEEPFAKE DETECTION

Drawing on the 108 primary studies synthesised in this review, we propose a three-axis taxonomy for deepfake detection. Each axis reflects a distinct analytical perspective on manipulated media.

A. Spatial Methods (Frame-Level Detection)

Spatial methods analyse individual frames for visual artifacts introduced during the deepfake generation process. These include blending boundaries at face-region edges, unnatural skin texture caused by GAN upsampling, compression artifacts, and lighting inconsistencies. CNNs are the primary workhorse, often augmented with attention modules focusing on the periocular zone, lip contours, and hairline transitions.

A notable refinement is the two-stream architecture: one stream processes the full image for global context, while a second processes high-pass-filtered residuals to isolate subtle noise-level manipulation signatures.

B. Temporal Methods (Sequence-Level Detection)

Temporal methods exploit the fact that even a generator producing photorealistic individual frames often fails to maintain physiologically consistent motion across a sequence. Eye blinking, head-pose dynamics, pulse-driven micro-motion (rPPG), and lip-sync correspondence are key cues exploited by sequence-level detectors.

Optical-flow-based methods complement pure sequence models by providing an explicit, interpretable representation of inter-frame motion that can be explained to non-expert audiences.

C. Multi-modal Methods (Cross-Stream Detection)

Multi-modal detection integrates evidence from multiple data streams — visual, acoustic, and physiological — to improve robustness against sophisticated forgeries. The intuition is that simultaneous manipulation of all modalities is significantly harder for an attacker.

Remote photoplethysmography (rPPG) is a particularly promising physiological signal: blood-circulation-driven colour fluctuations in facial skin pixels encode a heartbeat signature that is difficult to synthesise convincingly. Intel's FakeCatcher exploits this principle, achieving a claimed 96% accuracy.

TABLE I
Generalised Deepfake Detection Pipeline

Stage	Step	Description
INPUT	Raw video / image	Raw video / image frame
STEP 1	Pre-processing	Face detection, alignment, normalization
STEP 2	Feature Extraction	Spatial (CNN) + Temporal (LSTM / RNN)
STEP 3	Multi-modal Fusion	Audio-visual sync + rPPG signals
STEP 4	Classification Head	Softmax binary output: REAL / FAKE
OUTPUT	Authenticity Score	Confidence threshold → Decision

Fig. 1. Unified three-stage detection pipeline: pre-processing → multi-axis feature extraction → classification.

IV. STATE-OF-THE-ART METHODS AND COMPARATIVE ANALYSIS

Table II summarises 12 representative methods drawn from peer-reviewed literature published between 2024 and 2026, spanning all three axes of our taxonomy. The methods are evaluated on Celeb-DF v2 and DFDC. The AUC (Area Under the ROC Curve) at video level is reported as the primary metric.

TABLE II
Comparative Analysis of State-of-the-Art Deepfake Detection Methods (2024–2026)

Method	Architecture	Celeb-DF v2AUC (%)	DFDC AU C (%)	General.?
DefakeHop++	PixelHop + LightGBM	98.2	93.5	Yes
Dense-Swish-CNN + Bi-LSTM	CNN + Bi-LSTM	97.5	91.8	Partial
PUDD (Multi-modal)	Prototype + Transforme r	96.9	92.1	Yes
ViXNet	ViT + Xception	96.2	90.8	Yes
CNN Up-sampling	CNN + Generali sed Net	95.8	90.3	Yes
GAN-CNN Ensemble	GAN + CNN	95.4	89.7	Partial
FakeCatcher (rPPG)	rPPG + DL Classifier	95.1	N/A	Partial
Forgery Clue	Attention + CNN	94.6	89.0	Partial

Extraction				
Binary Neural Networks	BNN Real-time	93.9	88.4	No
Ensemble Deep Learning	Multi-model Ensemble	93.5	88.0	Partial
MesoNet	Compact CNN	88.3	82.0	No
MobileNet v3	Lightweight CNN	94.8	89.5	No

TABLE II. AUC values reported at video level. 'Generalisable' indicates cross-dataset performance confirmed by authors.

Several trends emerge from Table II. DefakeHop++, which departs from conventional deep learning in favour of a successive subspace learning approach, achieves the highest reported video-level AUC on Celeb-DF v2 while maintaining competitive DFDC performance. Multi-modal and prototype-based models show strong cross-dataset behaviour, supporting the hypothesis that biometric signals are harder for adversaries to simultaneously spoof.

V. PUBLICLY AVAILABLE DATASETS

The choice of benchmark dataset profoundly shapes the reported performance of detection algorithms. Table III catalogues six datasets that appear most frequently in the reviewed literature.

TABLE III

Summary of Major Deepfake Detection Benchmarks

Dataset	Size	Quality	Manipulation
FaceForensics++	1,000+	High/Low	Face swap, expression
DFDC	100,000+	Diverse	Multiple techniques
Celeb-DF v2	5,639 fake / 590 real	High fidelity	Face reenactment
Dataset	**Size**	**Quality**	**Manipulation**
Deepfake TIMIT	320	64×64, 128×128	GAN face swap
WildDeepFake	3,805 sequences	Real-world	Multiple sources
DFFD	100K–200K fake	ProGAN/StyleGAN	Full face synthesis

diverse, and continuously updated remains a logistical and ethical undertaking the research community has yet to fully address.

TABLE IV
Challenges and Proposed Mitigations

Challenge	Proposed Mitigation
Poor generalization	Disentangled representation + domain adaptation
Dataset bias	Diverse, real-world datasets + federated learning
High computation	Binary neural networks + lightweight architectures
Black-box decisions	Explainable AI (XAI) + attention visualization
Adversarial attacks	Adversarial training + ensemble robustness
Cross-modal deepfakes	Multi-modal transformers + rPPG biometric fusion

TABLE III. Dataset overview. DFDC is the largest publicly available benchmark.

A recurring limitation is over-reliance on FF++ and Celeb-DF v2. WildDeepFake represents a more realistic evaluation scenario, drawing deepfakes from genuinely in-the-wild sources rather than researcher-constructed settings.

VI. REAL-TIME DETECTION TOOLS AND COMMERCIAL SYSTEMS

Alongside academic research, a growing ecosystem of commercial deepfake detection tools has emerged. Sentinel employs a multi-layer defence architecture combining CNNs for facial expression and blinking-pattern analysis with an extensive database of known deepfake signatures. Sensity provides a cloud API built on multi-layered pixel, voice, and metadata analysis.

Intel's FakeCatcher analyses rPPG signals extracted from 32 facial landmark regions using OpenVINO-accelerated inference, achieving claimed throughput suitable for broadcast and live-streaming scenarios. Deepware, built on EfficientNet-B7 trained on 120,000 consented videos, adds temporal consistency checks and flicker detection on top of spatial classification.

VII. OPEN CHALLENGES

A. Cross-Dataset Generalisation

The single most persistent limitation is the inability of models trained on one benchmark to maintain performance when evaluated on another. A model achieving 97% accuracy on FF++ may drop to 75% or below on DFDC or WildDeepFake. A systematic review by Ramanaharan et al. found that only 46.3% of surveyed publications explicitly confirmed cross-dataset generalisation.

B. Adversarial Robustness

Detection models are vulnerable to adversarial perturbations: minute pixel-level modifications crafted to shift the model's output from 'fake' to 'real'. Such perturbations are imperceptible to human observers but devastatingly effective against gradient-based detectors.

C. Explainability and Legal Admissibility

A detection system that outputs only a binary label is insufficient for forensic and legal use cases, where investigators and juries expect reasoned, auditable evidence. Explainable AI (XAI) techniques such as GRAD-CAM, SHAP, and LIME can produce post-hoc saliency maps, but developing inherently interpretable architectures remains an open problem.

D. Real-Time and Resource-Constrained Deployment

The highest-accuracy architectures require compute resources unavailable in edge-device contexts. Binary neural networks and compressed CNN variants offer partial remedies. Federated learning offers a promising path toward distributing training across devices without centralising sensitive data.

E. Dataset Diversity and Bias

Existing benchmarks are skewed toward certain demographic groups, recording environments, and manipulation techniques. Curating datasets that are demographically balanced, geographically

TABLE IV. Summary of key open challenges mapped to mitigation strategies.

VIII. A MULTI-STAKEHOLDER MITIGATION ECOSYSTEM

Technical detection alone is insufficient. The most effective responses combine platform-level enforcement, community-driven verification, and government regulation into an integrated ecosystem.

A. Platform Responsibility

Social-media platforms occupy the most critical intervention point. Effective platform measures include automated AI-powered content analysis at upload time, clear labelling policies for detected synthetic media, human-review queues for borderline cases, and forwarding-chain indicators for peer-to-peer messaging platforms.

B. Community-Driven Verification

Crowdsourced fact-checking, when designed carefully to resist coordinated manipulation, provides a scalable signal that complements automated detection. A matrix-factorisation-based warning-scoring system surfaces warnings considered helpful by users across ideologically diverse viewpoints.

C. Government Legislation and Independent Oversight

Legal frameworks are necessary to establish accountability for the creation and distribution of malicious deepfakes. Existing Indian law (IT Act Sections 66C and 66D, IPC Sections 465 and 500) provides partial coverage, but dedicated deepfake legislation is needed. Crucially, enforcement should be vested in an independent body such as CERT.

IX. DISCUSSION

The literature reviewed here paints a nuanced picture: detection accuracy has improved dramatically over the past five years, yet the deepfake generation community innovates faster. Several higher-level insights emerge. First, architectural choice matters less than training strategy. Second, multi-modal approaches appear inherently harder to defeat than single-modality detectors. Third, the field needs standardised evaluation protocols: a common set of adversarial perturbations, a cross-dataset generalisation test suite, and a common efficiency benchmark.

The psychological and societal dimensions are also underweighted. Deepfakes exploit human cognitive biases, and purely technical countermeasures cannot address these vulnerabilities. Media literacy education is a necessary complement to the technical ecosystem.

X. CONCLUSION

This paper has presented a systematic review of deepfake detection using deep learning, synthesising findings from over 100 peer-reviewed publications into a unified taxonomy organised around spatial, temporal, and multi-modal detection paradigms.

The evidence suggests several priorities for the next generation of research: (1) generalisation-first training strategies; (2) lightweight and explainable architectures deployable at platform scale; (3) richer, more diverse benchmark datasets; and (4) interdisciplinary collaboration among computer scientists, forensic practitioners, legal scholars, and policymakers.

XI. ACKNOWLEDGEMENTS

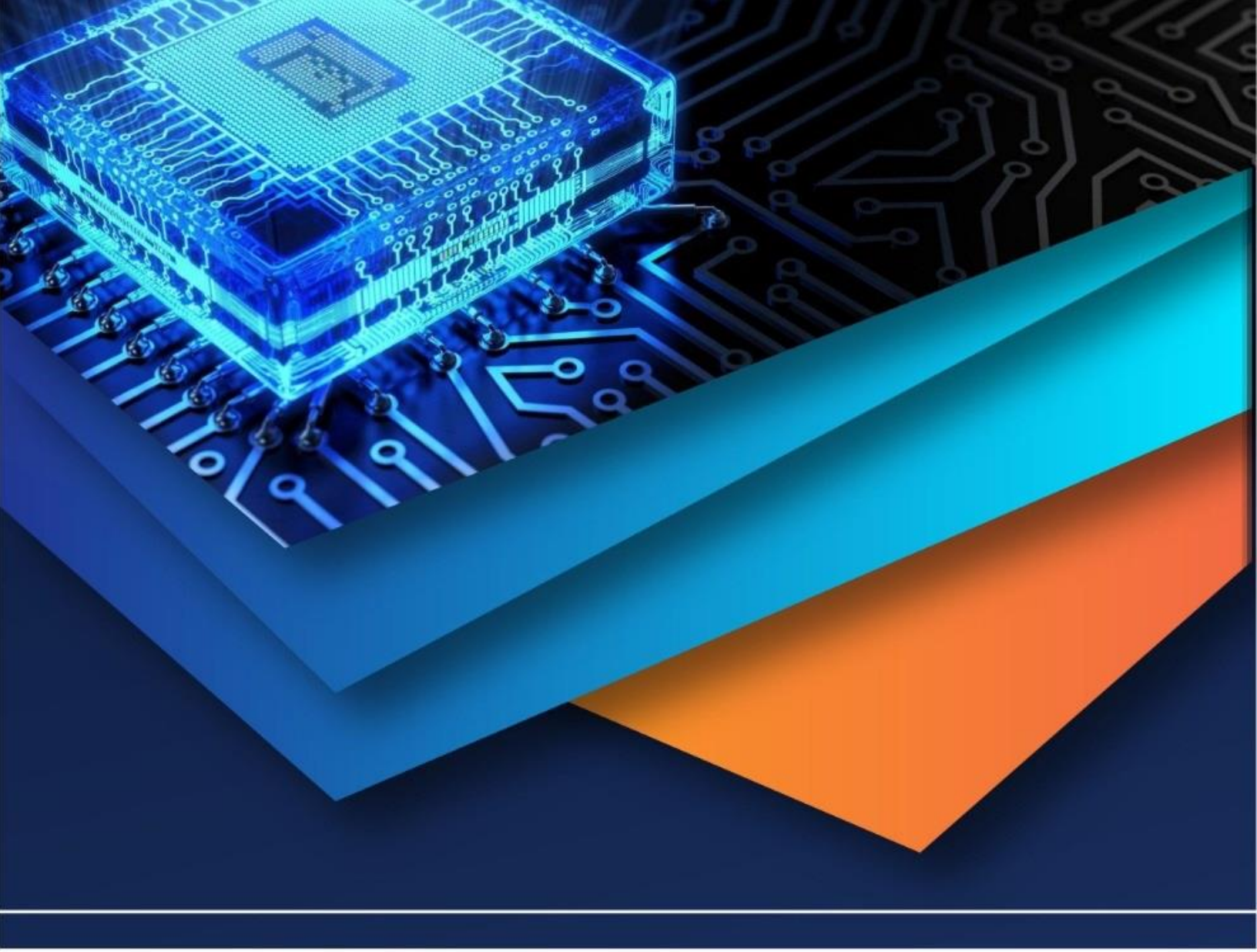
The authors gratefully acknowledge the researchers whose primary studies form the empirical foundation of this review, and the open benchmark maintainers who have made reproducible evaluation possible. No external funding was received for this work.

REFERENCES

[1] B. K. Panigrahi, S. P. Mishra, and C. K. Samal, “Deepfake detection using deep learning: A review,” Advances in Research, vol. 26, no. 4, pp. 555–564, 2025.
[2] V. Patel, S. R. Padiya, and K. Patel, “DeepFake detection through deep learning: A comprehensive review,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 12, no. 1, pp. 103–108, 2026.
[3] R. Ramanaharan, D. B. Guruge, and J. I. Agbinya, “DeepFake video detection: Insights into model generalisation,” Data and Information Management, vol. 9, p. 100099, 2025.

[4] B. V. P. Kumar, M. D. S. Ahmed, and M. Sadanandam, "Designing a safe ecosystem to prevent deepfake-driven misinformation on elections," Digital Society, vol. 3, p. 19, 2024.
[5] I. J. Goodfellow et al., "Generative adversarial networks," Advances in Neural Information Processing Systems, vol. 27, 2014.
[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.
[7] B. C. Soundarya and H. L. Gururaj, "A novel Dense-Swish-CNN with Bi-LSTM framework for image deepfake detection," IEEE Access, vol. 13, pp. 89641–89653, 2025.
[8] A. L. Pellicer, Y. Li, and P. Angelov, "PUDD: Towards robust multi-modal prototype-based deepfake detection," in Proc. CVPR, 2024, pp. 3809–3817.
[9] R. Lanzino et al., "Faster than lies: Real-time deepfake detection using binary neural networks," in Proc. CVPR, 2024, pp. 3771–3780.
[10] U. A. Ciftci, I. Demir, and L. Yin, "How do the hearts of deep fakes beat?" in Proc. IEEE IJCB, 2020, pp. 1–10.
[11] T. T. Nguyen et al., "Deep learning for deepfake creation and detection: A survey," Computer Vision and Image Understanding, vol. 223, p. 103525, 2022.
[12] A. Rössler et al., "FaceForensics++: Learning to detect manipulated facial images," in Proc. ICCV, 2019, pp. 1–11.
[13] A. Heidari et al., "Deepfake detection using deep learning methods: A systematic and comprehensive review," WIREs Data Mining and Knowledge Discovery, vol. 14, no. 2, e1520, 2024.
[14] S. Ahmed, Y. Chen, and Y. Liu, "DefakeHop++: A lightweight deepfake detection model," IEEE Transactions on Multimedia, 2023.
[15] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. 15th IEEE AVSS, 2018, pp. 1–6.
[16] Z. Ba et al., "Exposing the deception: Uncovering more forgery clues for deepfake detection," in Proc. AAAI, vol. 38, no. 2, 2024, pp. 719–728.
[17] P. Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN ensemble: A robust deepfake detection model of social media images," Procedia Computer Science, vol. 235, pp. 948–960, 2024.

[1] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI-generated fake face videos by detecting eye blinking," in Proc. IEEE WIFS, 2018, pp. 1–7.
[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
[3] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY