IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

Deepfake Generation and Detection Using Machine Learning: A Comparative Review of Techniques, Datasets and Benchmarks

Mr. Bansod Yash Santosh[1], Mr. S. N. Shinde[2], Prof. J. R. Mahajan[3]
[1] Student, [2]Guide, [3]Co-Guide
Department of Artificial Intelligence and Data Science
Adsul Technical Campus, Chas, Ahilyanagar, India

Abstract : *Deepfakes are highly realistic synthetic images, videos and audio produced with deep learning, and they have moved from a research curiosity to a serious threat to privacy, elections and financial security. This paper presents a comparative review of deepfake generation and detection, synthesised from three recent studies: a critical review of state-of-the-art detection approaches (Soundarya and Gururaj, 2026), a benchmark and survey of generation and detection (Pei et al., 2026), and the literature they cover. We organise the field into generation (GANs, VAEs, diffusion models, NeRF), detection (feature-based, deep-learning, frequency, temporal, multimodal and audio methods), benchmark datasets, and reported performance. We find that detectors reach over 98% accuracy on in-dataset tests such as FaceForensics++, but cross-dataset AUC often falls to the 65–85% range on harder sets such as DFDC. Generalisation to unseen forgeries, robustness to compression, lack of a unified evaluation protocol and the rapid rise of diffusion-based generation remain the central open problems.*

Keywords: Deepfake, Deepfake Detection, Machine Learning, Deep Learning, GAN, Diffusion Models, Face Forgery, Audio Deepfake.

I. INTRODUCTION

The term "deepfake" combines "deep learning" and "fake" and refers to photorealistic media created or altered by deep learning algorithms. Facial images carry rich personal information such as identity, age, gender and emotion, which makes the face the main target of manipulation [1]. Positive uses exist in filmmaking, virtual reality, entertainment and digital humans, but the technology is also used for non-consensual explicit content, harassment, political disinformation and voice-cloning fraud [1], [2].

Generation has progressed from graphics-based pipelines to Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs) and, more recently, diffusion models that produce content which is very hard to tell from real footage [2]. Detection has evolved in parallel, from handcrafted features to deep networks and multimodal systems. Because generators keep improving, detectors must keep evolving as well.

The objectives of this paper are:

1. To categorise the main deepfake generation tasks and underlying models.
2. To review machine-learning-based detection techniques for image, video and audio.
3. To summarise benchmark datasets and evaluation metrics.
4. To compare reported performance and identify open challenges and future directions.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40883

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

II. BACKGROUND AND RELATED WORK

A Types of Facial Manipulation

Facial manipulation is commonly divided into four groups: entire face synthesis, identity swap, attribute manipulation, and expression swap (face reenactment) [1]. Pei et al. organise generation into four mainstream tasks: face swapping, face reenactment, talking face generation and facial attribute editing, with forgery detection treated as the counter-task [2].

B. Unified Problem Formulation

Generation can be written as a controlled content-generation problem,

$$Io = \varphi G(It, C),$$

where It is the target image, C is condition information (image, audio, text) and φG is the generation network. Detection is the corresponding classification problem, So = φD(Io), where So is a fake score for the content Io [2].

C. Existing Surveys

Soundarya and Gururaj provide a critical review of detection approaches covering image, video and audio, together with benchmark datasets and real-world cases [1]. Pei et al. extend the scope to include diffusion-based generation, related sub-fields (head swapping, face super-resolution, makeup transfer and others), and a benchmark of representative methods [2]. The present paper consolidates both views into a single comparative treatment.

III. DEEPFAKE GENERATION TECHNOLOGIES

A. Generative Frameworks

VAEs encode data into a latent distribution and decode it back, trained by maximising the evidence lower bound while minimising reconstruction error. GANs use two networks, a generator and a discriminator, in a minimax game:

$$\min G \max D\ V(D, G) = Ex[\log D(x)] + Ez[\log(1 - D(G(z)))]$$

Important GAN variants include CGAN, Pix2Pix, CycleGAN, ProGAN, StyleGAN and StyleGAN2, which have been widely used to synthesise realistic faces [1], [2]. Diffusion models treat generation as gradual denoising of random noise. DDPM popularised the approach, latent diffusion models improved efficiency, and later work extended it to video generation [2]. Neural Radiance Fields (NeRF) add 3D consistency and are increasingly combined with generators for multi-view-consistent faces [2].

B. Face Swapping

Early face swapping used traditional graphics (key-point matching and 3D morphable models, 3DMM). GAN-based methods then improved generalisation, resolution (for example MegaFS at the million-pixel level), geometry preservation (HifiFace, FlowFace) and the trade-off between identity replacement and attribute retention (BlendFace, StyleSwap). Diffusion-based approaches such as DiffSwap and DiffFace recast swapping as conditional inpainting [2].

C. Face Reenactment

Reenactment transfers motion from a driving image or video to a target while keeping the target's identity. Methods are grouped into 3DMM-based, landmark-matching, feature-decoupling (for example StyleGAN latent-space approaches) and self-supervised designs, with recent HiDe-NeRF and diffusion-based DiffusionAct [2].

D. Talking Face Generation

Audio- or text-driven methods (Wav2Lip, MakeItTalk, SadTalker) were extended with multimodal conditioning, diffusion models (DAE-Talker, EMO, VASA-1) and 3D models such as AD-NeRF and SyncTalk. Remaining weaknesses are fine control of emotion intensity and coupling between head pose and expression [2].

E. Facial Attribute Editing and Audio Synthesis

Attribute editing changes age, expression, hair or accessories using latent-space disentanglement, text guidance and diffusion. For audio, WaveNet, Tacotron and related text-to-speech systems can clone a target voice, and consumer tools make voice-based fraud easy [1]. Table 1 lists common creation tools reported in [1].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

Table 1. Representative deepfake creation tools [1].

Category	Representative tools / models
Face swap / editing	FaceApp, ZAO, DeepFaceLab, FSGAN, FaceSwapGAN
Image synthesis	StyleGAN, StyleGAN2, StarGAN, DiscoFaceGAN
Voice synthesis	Respeecher, Resemble AI, Descript, SV2TTS, Lyrebird
Avatar / video	Synthesia

IV. DEEPFAKE DETECTION TECHNIQUES

Deepfake detection is usually posed as binary classification (real vs fake), based on artifacts in geometry, lighting, texture, colour and expression [1]. Reviewed methods fall into the following groups.

A. Feature-Based Methods

These exploit specific, interpretable cues. DeepVision uses eye-blink count, duration and interval, with the Eye Aspect Ratio (EAR) computed from landmark distances, reaching 87.5% accuracy [1]. Matern et al. use eye-colour and iris inconsistencies (AUC 85.1%). Nguyen and Derakhshani use eyebrow matching (AUC 0.984 on UADFV but 0.712 on Celeb-DF). Physiological methods such as FakeCatcher, DeepFakesON-Phys and DeepRhythm use heart-rate and skin-colour (photoplethysmography) signals, with DeepFakesON-Phys reporting 98%. Yang et al. use inconsistent 3D head pose (84.3% on UADFV) [1].

B. Deep-Learning-Based Methods (Image and Video)

CNN backbones such as XceptionNet, VGG and ResNet, recurrent models (CNN + LSTM), attention-based models, Vision Transformers and hybrid CNN–ViT models dominate the literature. Examples include CNN + LSTM temporal pipelines (97.2% on HOHA [Güera and Delp]; 97.6% [Hashmi et al.]), a Convolutional Vision Transformer (91.5% on DFDC), AMTENnet for manipulation traces (95.17%) and frequency CNNs [1].

C. Space-Domain Methods

These look for image-level inconsistencies: texture (Gram-Net), blending boundaries (Face X-ray), face-versus-context discrepancy (Nirkin et al.), self-blended images (SBIs), gradient artifacts (LGrad) and noise traces (NoiseDF) [2].

D. Time-Domain Methods

These target abnormal physiology (blinking, head pose, gaze) and inter-frame inconsistency, including FTCN, LipForensics, M2TR, graph-based relation learning (MRL) and style-latent flow analysis. Multimodal methods such as AVoiD-DF, POI-Forensics and RealForensics check audio–visual consistency [2].

E. Frequency-Domain Methods

F3-Net, HFI-Net, FreqNet, and more recent WMamba and WaveDIF analyse frequency components, since forgery traces concentrate in mid and high frequencies and are robust to some visual improvements in generators [2].

F. Data-Driven Methods

These learn generator fingerprints or self-consistency (pairwise self-consistency learning, implicit identity-driven detection, hierarchical fine-grained localisation such as HiFi-Net) to improve generalisation [2].

G. Audio Deepfake Detection

Audio detectors use MFCC, LFCC, log-power and mel-spectrogram features with classical or deep models. Reported results include a Quadratic SVM (97.56% accuracy), TCN on Fake-or-Real (92%), a CNN on FoR-combined (97.23%), SENet-based models on ASVspoof 2019, an ELTP–LFCC–DBiLSTM countermeasure and a HuBERT-based model (97% precision). The FakeAVCeleb study found that single unimodal classifiers were weak, with a best accuracy of 76%, supporting fusion of audio and visual detectors [1].

V. BENCHMARK DATASETS AND METRICS

Table 2. Widely used deepfake detection datasets [1], [2]. Counts differ slightly between sources, and real/fake counts for the same dataset are not always reported consistently.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

Dataset	Content (as reported)	Notes
UADFV	49 real + 49 fake videos	Early set focused on eye blinking
Deepfake-TIMIT	640 fake videos (64×64, 128×128)	Derived from VidTIMIT
FaceForensics++	1,000 real, 5 forgery methods	Most common benchmark; HQ and LQ versions
Dataset	Content (as reported)	Notes
Celeb-DF (v2)	590 real, 5,639 fake	Improved quality, fewer artifacts
DFDC / DFDC-P	Preview: 1,131 real, 4,119 fake	Paid actors; many perturbations
DeeperForensics-1.0	60,000 videos, 17.6M frames	Large, diverse perturbations
FakeAVCeleb	~20K audio-video clips	Audio-visual multimodal
KoDF	237,942 videos	Korean subjects, adds Asian-face diversity
ForgeryNIR	50,000+ identities	Near-infrared modality

Metrics. Generation is evaluated with PSNR, SSIM, LPIPS, FID, KID, identity retrieval (IDRet), expression and pose error, landmark distance (LMD) and lip-sync scores (LSE-C, LSE-D). Detection is evaluated with AUC, accuracy, Equal Error Rate (EER), average precision and F1-score [2]. Note that [1] identifies accuracy as the most commonly used detection metric, while [2] relies mainly on AUC and ACC with explicit cross-dataset testing.

VI. COMPARATIVE PERFORMANCE ANALYSIS

All numbers below are taken from the original publications as summarised in [1] and [2]. Training data, splits and preprocessing differ between papers, so the tables indicate trends and are not a strictly controlled comparison.

A. In-Dataset Performance on FaceForensics++

Table 3. Self-dataset results on FF++ (HQ = mild compression, LQ = heavy compression) [2].

Method	LQ ACC (%)	LQ AUC (%)	HQ ACC (%)	HQ AUC (%)
F3-Net	93.02	95.80	98.95	99.30
Zhao et al.	88.69	90.40	97.60	99.29
FDFL	89.00	92.40	96.69	99.30
LipForensics	94.20	98.10	98.80	99.70
RECCE	91.03	95.02	97.06	99.32
Guo et al.	92.76	96.85	99.24	99.75

B. Cross-Dataset Generalisation (AUC %)

Table 4. Cross-dataset AUC [2].

Method	DFDC	Celeb-DF	Celeb-DFv2	DeeperForensics
Face X-ray	80.92	80.58	–	–
Zhao et al.	67.52	98.30	90.03	99.41
LipForensics	73.50	–	82.40	97.60
RECCE	69.06	68.71	–	–
SBIs (FF++)	72.42	–	93.18	–
WMamba	82.97	96.29	–	–

C. Observations

1. In-dataset accuracy on mildly compressed FF++ is near saturation (about 97–99%), whereas heavy compression lowers accuracy by roughly 3–10 points.
2. Cross-dataset performance is much lower and varies widely. DFDC is consistently the hardest, with AUCs mostly between 67% and 83%.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

3. Methods that target generalisable artifacts (blending boundaries, self-consistency, frequency cues) tend to transfer better than methods that fit one dataset
4. Biological-signal and eye-based features give interpretable cues but their performance is inconsistent across datasets (for example eyebrow matching: 0.984 vs 0.712 AUC) and they can be defeated as generators learn to model blinking and gaze.
5. For audio, high accuracy (92–97.6%) is reported on Fake-or-Real, but unimodal detectors perform poorly on audio-visual data, which supports multimodal fusion [1].
6. In generation, recent diffusion and 3D-aware methods improve identity fidelity and temporal consistency, which in turn makes detection harder [2].

VII. CHALLENGES

Generalisation: Detectors trained on known forgeries fail on unseen generators and in-the-wild content [1], [2].

Robustness: Compression, resizing and adversarial perturbations reduce accuracy and can be used to evade detectors [2].

Evaluation inconsistency: There is no unified protocol. Training sets, preprocessing and metrics differ, so published numbers are hard to compare [2]. Dataset statistics are also reported inconsistently across sources.

Temporal coherence, illumination, eye and lip realism, identity leakage: Open weaknesses in generation that are also useful detection cues, until they are fixed [1].

Emerging generators: Diffusion-based and fully AI-generated content lacks the localised blending artifacts many detectors rely on [2].

Real-time and lightweight deployment: Large models limit practical use, though platforms and banks already deploy detection and liveness checks [2].

Ethics and law: Privacy, consent and misuse are being addressed through regulation such as the EU AI Act and China's deep synthesis provisions [2].

VIII. FUTURE SCOPE

1. Multimodal detection combining visual, audio, text and physiological evidence with fine-grained temporal modelling.
2. Self-supervised and weakly supervised learning to reduce annotation needs and adapt to new manipulation types.
3. Generalisable training, such as synthetic-forgery augmentation, domain-invariant representation and frequency-domain learning.
4. Standardised benchmarks and protocols that fix training sets, splits and metrics, and cover diffusion-generated content.
5. Provenance and transparency mechanisms (watermarking, signed metadata) alongside detection.
6. Lightweight, real-time detectors suitable for edge and platform deployment.
7. Robustness to adversarial attacks and compression.

IX. CONCLUSION

This paper reviewed deepfake generation and detection by consolidating three recent studies. Generation has advanced from GANs to diffusion and 3D-aware models, producing content that is increasingly indistinguishable from real media. Machine-learning detectors, especially deep and multimodal ones, achieve very high accuracy under controlled in-dataset conditions, but generalisation across datasets, robustness to compression and the lack of a common evaluation protocol remain the main obstacles. Future work should emphasise generalisable, multimodal and lightweight detectors, standardised benchmarking, and provenance-based safeguards to reduce harm from synthetic media.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

REFERENCES

[1] B. C. Soundarya and H. L. Gururaj, "Deepfake detection: critical review of state-of-the-art approaches and future perspectives," Discover Applied Sciences, vol. 8, art. 201, 2026. doi: 10.1007/s42452-025-08174-9.
[2] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, and D. Tao, "Deepfake generation and detection: a benchmark and survey," arXiv:2403.17881v5 [cs.CV], Mar. 2026.
[3] I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, 2014.
[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF CVPR, 2019, pp. 4401–4410.
[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. NeurIPS, 2020.
[6] A. Rössler et al., "FaceForensics++: learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, 2019, pp. 1–11.
[7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: a large-scale challenging dataset for deepfake forensics," in Proc. IEEE/CVF CVPR, 2020, pp. 3207–3216.
[8] B. Dolhansky et al., "The deepfake detection challenge (DFDC) preview dataset," arXiv:1910.08854, 2019.
[9] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: exposing AI-generated fake face videos by detecting eye blinking," arXiv:1806.02877, 2018.
[10] L. Li et al., "Face X-ray for more general face forgery detection," in Proc. IEEE/CVF CVPR, 2020, pp. 5001–5010.
[11] Y. Qian et al., "Thinking in frequency: face forgery detection by mining frequency-aware clues," in Proc. ECCV, 2020.
[12] A. Haliassos et al., "Lips don't lie: a generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF CVPR, 2021.
[13] R. Tolosana et al., "Deepfakes and beyond: a survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.
[14] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in Proc. 1st Workshop on Synthetic Multimedia, 2021, pp. 7–15.