International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3238-3242
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0413

Deepfake Detection and Prevention Using Deep Learning

Chaitra alur1, Manthan YK[2] , Madhavan subramaniyan K[3] , Chethan Chandu A[4]
Dr Usman Aijaz (PhD)[5] , Roshan Teja S[6]
[1]Assistant professor, Dept. of CSE, Yenepoya(Deemed To Be University), Bangalore, Karnataka, India.
[5]Head of Department, Dept. of CSE, Yenepoya(Deemed To Be University),Bangalore, Karnataka, India.
[2,3,4,6] UG Scholar, Dept. of CSE, Yenepoya(Deemed To Be University), Bangalore, Karnataka, India.
Email ID: *chaitraalur1998@gmail.com[1], 45782@yenepoya.edu.in[2], 45927@yenepoya.edu.in[3], 45341@yenepoya.edu.in[4], 45392@yenepoya.edu.in[5], usmanaijaz.blr@yenepoya.edu.in[6]*

Abstract
The rapid evolution of artificial intelligence technologies has greatly increased the capabilities of creating synthetic media content, also referred to as deepfakes. These artificially created images, audio recordings, and videos are able to convincingly replicate real people and events in the world. As a result, deepfake technology poses a serious concern when it comes to the authenticity and privacy of information. Even though deepfake technology is useful in the creation of positive content in the fields of entertainment, education, media production, and accessibility, its misuse in the spread of misinformation, identity theft, political manipulation, and defamation is a major concern. The present review paper aims to conduct a comprehensive review of the available research on deepfake generation and detection techniques. It also reviews the state-of-the-art deep learning models used in the creation of synthetic media content and the detection techniques used in the identification of deepfake content. In addition, the review also discusses the prevention and mitigation techniques used in the identification of deepfake content through the use of verification technologies, platforms, regulatory policies, and public awareness campaigns.

Keywords: *Deepfakes, Artificial Intelligence, Deep Learning, Synthetic Media, Deepfake Detection, Digital Forensics, Content Authentication*

1. Introduction
The rapid advancement of artificial intelligence (AI), particularly deep learning, has fundamentally transformed digital media creation. Neural architectures such as Generative Adversarial Networks (GANs) and variational autoencoders now enable machines to synthesize photorealistic faces, replicate voices, and fabricate entire visual narratives that are indistinguishable from authentic recordings. This category of AI-generated synthetic media is widely known as deepfakes a portmanteau of "deep learning" and "fake." While deepfake technology has legitimate applications in film production, medical simulation, and digital accessibility, its misuse poses catastrophic risks. Malicious actors exploit deepfakes for identity fraud, political misinformation, non-consensual explicit content, and large-scale financial scams. In India alone, deepfake-related financial losses in 2025 reached an estimated ₹70,000 crore, with nearly 47% of Indian adults reporting direct or indirect exposure to AI-driven synthetic media attacks. The national threat has escalated to such an extent that entire geographic regions including Maharashtra, Telangana, and Karnataka have emerged as hotspots for organized deepfake crime.

2. Background and Related Work
2.1.Evolution of Deepfake Technology
Deepfake technology made a significant leap in 2014 when researchers introduced Generative Adversarial

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3238-3242
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0413

Networks (GANs). In this system, two artificial intelligence models compete with each other. One model creates fake media while the other tries to detect it. This ongoing competition continually improves the quality of the fakes. Soon after, user-friendly tools like DeepFaceLab and FaceSwap emerged, allowing regular people to swap faces in videos with surprisingly realistic results using just a personal computer. As technology advanced, systems like StyleGAN could create entirely fictional human faces that looked real, making it hard to distinguish between what was genuine and what was fake. This rising threat prompted researchers to develop shared testing datasets such as FaceForensics++, the DFDC corpus, and Celeb-DF, providing the scientific community with a common basis for measuring and improving detection methods.

2.2.Learning-Based Detection Approaches

Since deepfakes aim to mislead the human eye, researchers turned to artificial intelligence to catch what people cannot perceive. Early detection systems trained on large sets of real and fake images learned to identify hidden patterns resulting from the manipulation process. Models like XceptionNet and EfficientNet were particularly effective in this area. Later methods enhanced this by analyzing the entire image rather than just small sections, helping to spot subtle inconsistencies that might otherwise go unnoticed. For videos, detection became even more insightful. Real human faces move in natural, consistent ways. AI systems that study facial motion across multiple frames can easily identify the unnatural stiffness and awkward expressions often found in manipulated videos.

2.3.Prevention and Forensic Authentication

Researchers have also focused on preventing damage caused by deepfakes by finding ways to verify if media is genuine from the moment it is created. One approach involves embedding hidden markers in original photos and videos during creation. These markers break silently if someone tampers with the content, indicating that something has changed. Another method uses digital signatures to link the content creator's identity to the file itself. Any modification will immediately mark the content as untrust worthy. Recently, blockchain technology has been considered a way to store these authenticity records in a permanent, publicly accessible system that no single individual or organization can alter. This offers a reliable and transparent way to trace the true source of any piece of media. Figure 1.

3. Methodology

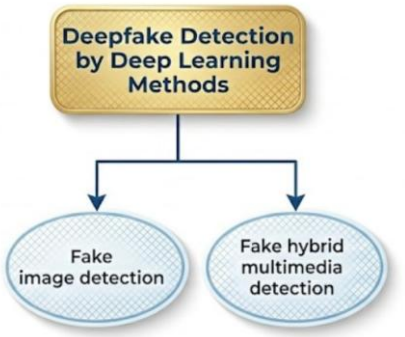

Figure1 Deepfake Detection Methods

3.1.Fake Image Detection: Analyzing Spatial Artifacts

Fake image detection is about finding manipulated or entirely created static images. Even if an AI-generated image looks perfect to our eyes, the algorithms, such as Generative Adversarial Networks or Diffusion Models, often leave tiny digital fingerprints. Deep learning methods in this area usually focus on spotting subtle spatial inconsistencies:

- **Pixel-Level and Frequency Anomalies:** Advanced Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are trained to identify unnatural blending boundaries, unusual noise patterns, or irregular frequencies that don't appear in real photographs.
- **Semantic Inconsistencies**: Models are increasingly made to look for structural

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3238-3242
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0413

problems, like uneven facial features, unnatural light reflections in the eyes, or distorted background textures.

- **Generalizability Challenges:** Current research in this field aims to create "zero-shot" or highly general models. Because generative techniques change quickly, a detection model trained only on GAN-generated faces must be strong enough to recognize new manipulation methods, such as deep-text-to-image diffusion outputs.

3.2. Fake Hybrid Multimedia Detection: Cross-Modal Synchronization

Static images need spatial analysis, but hybrid multimedia, which combines different types of data like video and audio, requires a much more complex, time-based approach. Modern deepfakes often involve changing a person's facial movements to match a synthesized audio track, like lip-syncing, or completely swapping identities within a moving scene. Detecting these advanced forgeries needs multimodal deep learning systems:

- **Temporal Consistency:** Instead of just looking at a single frame, spatial-temporal networks, like 3D CNNs or Recurrent Neural Networks, assess the sequence of frames. They watch for micro-flickering, unnatural motion blur, or inconsistencies in how light interacts with a moving face over time.
- **Audio-Visual Desynchronization:** The best deep learning techniques for hybrid media examine how different data streams relate to each other. For example, models check if the audio waveforms perfectly match the subject's lip movements and micro-expressions. If the visual phonemes don't correspond with the spoken audio, the system marks the media as a hybrid forgery.
- **Biological Signals:** Recent research in this field also uses deep learning to track involuntary biological signals, like artificial changes in heart rate through remote photoplethysmography or unnatural blinking patterns. These are very hard for generative AI to imitate accurately across video and audio streams. Figure 2.

4. Deepfake Pravetion Techniques

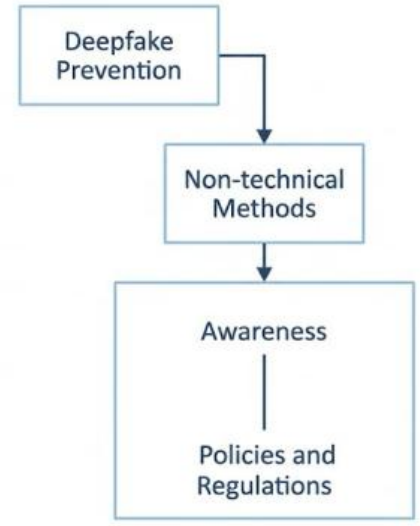

Figure 2 Awareness, Policies, and Regulations: The Societal Defense Framework

A large amount of research shows that relying on technology alone won't stop the spread of deepfakes. A strong defense strategy needs a solid societal framework that includes public awareness, strict policies, and flexible regulations [1].

4.1. Cultivating Ethical Development and Public Awareness

One key weakness in today's tech landscape is the gap between fast algorithm advancements and ethical responsibility. Studies point out a troubling lack of awareness about the ethical issues surrounding artificial intelligence (AI) development, especially concerning user consent, algorithmic bias, and the wider effects of synthetic media [1]. To tackle this, experts suggest establishing strong accountability throughout the AI supply chain, from the initial algorithm developers to service providers, to prevent harmful generative applications [2]. Additionally,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3238-3242
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0413

generative models should adopt "safety by design" principles, including essential safeguards in their architecture before public launch [2]. Besides these basic safeguards, the industry needs to enforce mandatory "red-teaming" during development. By simulating malicious attacks on their generative models, developers can find and fix vulnerabilities before the software is released. Public attitude surveys consistently support this proactive method, showing strong societal demand for localized awareness campaigns, improved digital literacy, and solid identity verification practices to discourage deepfake creation [3].

4.2.Bridging the Legislative and Regulatory Divide

The rapid development of generative AI has significantly outstripped current legal frameworks, putting societies at risk for serious economic, political, and personal harm. Existing criminal laws are often insufficient to tackle the complex and rapidly changing risks posed by deepfake abuse. This demands a variety of legal responses [4]. Regional case studies highlight these serious gaps. For instance, analyses of Indonesia's legal environment show a critical lack of targeted protections for victims of deepfake-related non-consensual pornography and data manipulation. Current electronic information laws do not specifically address the technical details of synthetic media [5]. Similarly, research into India’s legal system finds that while there are general regulations like the Information Technology Act, the inconsistent enforcement fails to adequately regulate or compensate for the damage caused by deepfake content [6]. To support public trust, researchers stress the immediate need for comprehensive, tech-specific regulations [4]. Suggested legislative changes include adjusting broad frameworks, such as the European Union's AI Act, to demand stricter transparency and watermarking rules [7], and revising electoral laws to protect democratic processes. Also, since deepfake distribution crosses borders, a vital next step for global policy is creating international digital treaties to enable cooperation on jurisdictional matters [6].

4.3.Empowering Gatekeepers Through Education and Collaboration

Addressing the deepfake threat needs a united approach from various sectors, giving key information gatekeepers the skills to recognize and intercept altered media. In the media sector, major news organizations must focus on training journalists to spot synthetic misinformation. They should also include advanced media forensics tools in their verification processes to maintain the integrity of digital media [8]. Likewise, the legal sector must change by adding specific AI and deepfake education to formal legal training. This will ensure professionals can handle the unique challenges these technologies present in court [4]. By thoroughly examining algorithmic models, assessing legal risks, and recognizing the legitimate uses of synthetic media, policymakers can create evidence-based guidelines that protect individuals without hindering technological progress [1]. Finally, information gatekeepers should use "pre-bunking" or psychological inoculation techniques. By proactively teaching the public about the tactics used to manipulate stories with deepfakes before a viral campaign starts, media organizations can help build cognitive resilience in society and lessen the overall impact of synthetic deception [3].

Conclusion and Future Directions

The rapid rise of generative AI has completely changed how we trusts digital media. Deepfakes have evolved from simple internet tricks to complex tools that pose real risks to our political systems, economies, and daily lives. This review shows that defending against this wave of synthetic media is no longer just a computer science issue; it requires a joint effort involving both technology and society. The Limits of Catching Fakes. In detecting deepfakes, we have made significant progress from basic forensics to advanced deep learning. We are currently skilled at spotting fake images by identifying tiny errors in pixels or lighting. However,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3238-3242
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0413

the real challenge has shifted to complex media videos where faces, voices, and even biological signals are seamlessly faked together. To detect these, our AI systems must analyze how all these elements interact over time. But there is a problem: our current detection tools are mostly reactive. We find ourselves playing catch-up. Detectors struggle with new manipulation techniques, can be easily misled by small changes, and often need too much computing power to work instantly on regular devices. Moving Toward Proactive Prevention Since we cannot rely solely on detecting fakes after they appear, we must focus on preventing them at the source. On the tech side, this means using tools like digital watermarking and blockchain to securely mark media the moment it is created, proving its authenticity. However, technology alone isn't enough. We need a strong societal defense. This involves holding AI companies responsible for including safety features in their software before it is released. It also requires educating people on how to recognize fakes and establishing international laws so that malicious creators can face prosecution, regardless of where they are located. The Path Forward. The most promising future research direction is bringing these separate solutions together. Imagine a system where the digital "watermark" of a video communicates instantly with an AI detector to confirm its authenticity in real-time. Additionally, we should focus on "Explainable AI." This means that when a computer flags a video as a deepfake, it can clearly explain its reasoning to a human judge or journalist. Ultimately, safeguarding the truth online will not come from a single magic algorithm. It will require ongoing collaboration between the tech industry, lawmakers, and an informed public that knows how to critically evaluate what they see on their screens.

Acknowledgement

The authors gratefully acknowledge the guidance, mentorship, and support of Chaitra Alur and Dr. Usman Aijaz (PhD), Head of the Department, whose insights shaped the direction of this review research paper. The authors also thank Yenepoya (Deemed to be University), Bengaluru Campus, for providing the academic infrastructure and computational resources that made this work possible. Additionally, we acknowledge the assistance provided by AI tools throughout the preparation of this manuscript.

References

[1]. Kaur et al., "A Framework for Ethical AI-Generated Content Governance," Preprints.org, Sept. 2025.
[2]. "Responsible AI framework in the age of deep fakes and false narratives," International Journal of Science and Research Archive, vol. 16, no. 2, pp. 1531-1542, Aug. 2025.
[3]. "Seeing Isn't Believing: Addressing the Societal Impact of Deepfakes in Low-Tech Environments," arXiv preprint arXiv:2508.16618, Aug. 2025.
[4]. "The Era of Artificial Intelligence: Examining Indonesia's Adaptability and Legal Challenges," Law and Humanities Quarterly Reviews, 2024.
[5]. A. V. A. Nasution, Suteki, and A. D. Lumbanraja, "Addressing Deepfake Pornography and the Right to be Forgotten in Indonesia: Legal Challenges in the Era of AI-Driven Sexual Abuse," Universitas Diponegoro, 2025.
[6]. "Deepfake Technology and It's Legal Regulation in India: A Doctrinal and Comparative Study," Vintage Legal, Aug. 2025.
[7]. "Regulating Deep Fakes in the Artificial Intelligence Act," Publisherspanel, 2024.
[8]. "Ethical and Social Implications of Generative AI and Deepfakes," International Journal of Environmental Sciences, 2025.