Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 29 (2026), No. 10-C, pp. 4765–4773
DOI : 10.47974/JDMSC-3090

Deepfake threats in digital security : Detection and prevention mechanisms in networked environments

Vaibhav Banga [@]
Nokia Corporation
Bengaluru 560045
Karnataka
India

Netra Jain [†]
Department of Computer Science
Apaji Institute of Mathematics & Applied Computer Technology (AIM & ACT)
Banasthali Vidyapith
Banasthali 304022
Rajasthan
India

Amit Hirawat [§]
Faculty of Computer Science & Engineering
Poornima University
Jaipur 303905
Rajasthan
India

Rekha Jain *
Arpana Sinhal [‡]
Department of Computer Applications
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

[@] *E-mail:* vaibhav.banga@nokia.com
[†] *E-mail:* jainetra9@gmail.com
[§] *E-mail:* amit.hirawat@poornima.edu.in
* *E-mail:* rekha.jain@jaipur.manipal.edu (Corresponding Author)
[‡] *E-mail:* arpana.sinhal@jaipur.manipal.edu

© Taru Publications

Abstract

Deepfake Technology is developed by both computer vision research and the deep learning that has evolved as a spin-off from the advances made in this area. Deepfakes are classified as artificial digital media (including video, photography, and audio) and are created with a high degree of accuracy to replicate the appearance, voice, and actions of human beings. However, deepfakes also can pose a very real threat to an individual's or organization's privacy, data protection, and authenticity of digital content. In addition to having beneficial uses, in the last few years there have been many cases of using deepfakes for malicious reasons including, but not limited to, misrepresentation and misinformation, identity fraud, social engineering, and discrediting individuals and/or companies.

Subject Classification: *68M18, 68M12.*

Keywords: *Deepfake, Cybersecurity, Generative adversarial networks (GANs), Transformer models, Network security.*

1. Introduction

Generating fakes through advanced technologies, such as Deepfake or synthetic media, is referred to as generating synthetic media with machine learning methods. Deepfake technology has changed the way we communicate digitally, as well as created much concern regarding how much of our online content is real. The objective for these technologies was initially to be beneficial (in terms of helping with both entertainment and research) [1]. Researchers have spent a significant amount of resources studying the impacts of deepfakes on society, developing ways to detect deepfakes and creating ways to prevent the dissemination of deepfakes. This paper identifies methods that can detect deepfakes within a networked system, evaluates existing methods for preventing the use of deepfakes and identifies potential directions for future research. The discussion begins with the introduction of deepfakes in section 1, followed by literature review mentioned in section 2 and then detailing about Deepfake technologies in section 3. Section 4 describes various threats related to Deepfake. Detection and prevention mechanisms are defined in section 5 and 6 respectively.

2. Literature Review

Deepfakes have rapidly developed with the growth of artificial intelligence (AI) and deep learning to change the manners in which digital media exists. With the help of Generative Adversarial Networks (GANs) it is now possible to produce fake videos and images that replicate the accuracy of the real [1]. Many researchers are also designing various detection and prevention methods including varied methods from traditional image based ways of classifying to more complex transformer based methods and verification methods that are supported by blockchain.

Early investigations into deepfake detection focused on detecting visual inconsistencies in altered content. A notable development by Rossler et al. [2] was the development of the FaceForensics++ dataset to allow for the use of convolutional neural networks (CNN) to detect pixel level inconsistencies in fake images of people. Following this, Dolhansky et al. [3] created a benchmark dataset called the DeepFake Detection Challenge (DFDC), which was the first large-scale dataset that allowed researchers to evaluate detection models using a common testing dataset. Afchar et al. [4] developed a new model called MesoNet, which is a compact CNN that has been optimized to detect very small compression artifacts in altered video images. In terms of accuracy for specific datasets; the CNN-based models did not consistently exhibit an ability to generalize across all datasets or real-life formats containing very high amounts of lossy compression.

Li and Lyu [5] shifted focus from CNN-Based Models towards utilizing bio-signals, such as eye blink patterns, to differentiate 'REAL' video content from 'MANIPULATED' video content. Table 1 describes the work done in domain of Deepfake.

Table 1
Work done by researchers in the domain of Deep Fake

Approach	Key Researchers/Year	Technique	Advantages	Limitations
CNN-Based Detection	Rössler et al. (2019); Afchar et al. (2020)	Pixel-level forgery detection using CNNs [2, 4]	High accuracy on benchmark datasets	Poor cross-dataset performance
Biological Signal Analysis	Li & Lyu (2020); Intel (2022)	Eye blinking, PPG-based detection [5]	Uses natural human signals, harder to spoof	Sensitive to video resolution
Audio-Visual Fusion	Haliassos et al. (2022); Wang et al. (2023)	Synchronization between audio and video cues [6]	Improved multimodal robustness	High computational cost
Blockchain Provenance	Soudy et al. (2024)	Media hashing and cryptographic metadata [7]	Ensures authenticity and traceability	Implementation complexity
Transformer-Based Models	Kingma and Welling (2014)	Global attention and self-supervised features [8]	Strong generalization and scalability	Requires large-scale datasets

3. Deepfake Technology

The manipulations in media such as images, video or audio can be highly realistic, making it difficult for humans or conventional detection systems to distinguish real content from fake. To effectively detect and prevent fraudulent use of technology, you must first understand the technological foundation that underlies it. This need for knowledge is especially relevant in all types of networked environments including social media, enterprise networks and streaming services.

3.1 *Generative Adversarial Networks (GANs)*

Generative Adversarial Networks (GANs) are the predominant framework being used for deepfakes today. A GAN consists of two separate but related networks [1]:

1. Generator G – Produces synthetic media by learning the statistical patterns of real data.
2. Discriminator D – Evaluates whether the generated media is real or fake.

The objective function Eq. (1), of a GAN is a minimax game,

$$\min_G \max_D \; V(D,G) = E_{x\sim p_{\mathrm{dt}}(x)}[\log D\,(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right] \tag{1}$$

Where:

$\boldsymbol{x}$**:** A real data sample drawn from the real data distribution.

$\boldsymbol{p_{data}(x)}$: The true (unknown) probability distribution of real-world data.

$\boldsymbol{z}$: A random noise vector sampled from a prior distribution.

$\boldsymbol{p_z(z)}$: The prior distribution of the noise vector (commonly Gaussian $\mathcal{N}(0,1)$or Uniform).

$\boldsymbol{G(z)}$**:** The fake/generated sample produced by the generator from noise z.

$\boldsymbol{D(x)}$**:** The discriminator's output — the probability that input $x$is real (value between 0 and 1).

Here Discriminator D tries to maximize the probability of correctly identifying real vs. fake data and Generator G tries to minimize log(1 - D(G(z))), i.e., fool the discriminator. An important part of the GAN training process includes alternating between optimizing (improving) the Discriminator (D) and the Generator (G), leading ultimately to the creation of increasingly realistic, synthetic outputs. GAN architectures consist of a generator (G) that creates synthetic images and a discriminator (D) that evaluates whether a particular image is authentic or original (i.e. comes from the training data) or synthetic (i.e. generated by G). The D is

continually improved until its ability to distinguish between artificially (synthetic) and genuine (authentic) images is compromised. The GAN architecture has shown significant promise in enhancing the quality and realism of face swaps, lip syncing, and/or synthesizing audio and is used within multiple types of social networking, internet video and messaging applications.

3.2 *Variational Autoencoders (VAEs)*

Another technique utilized in the creation of deep fakes is the Variational Autoencoder (VAE). VAEs have an approach that contrasts with the Generative Adversarial Network (GAN); they encode the input into a continuous latent space and then, through the process of decodes the encoded input, reconstructs the original media. VAEs can produce very smooth transitions between different types of facial expressions and voices. The majority of the time VAEs will not be able to create as realistic images as GANs; however, they require much less data and processing capability than GANs, which is why they are a good candidate for use by devices with limited processing power on a network [8]. The use of a VAE is to learn the conditional distribution and thus to learn the distribution on the underlying latent variable. The marginal likelihood Eq. (2) of data x is:

$$p_\theta(x) = \int p_\theta(x \mid z)\ p_\theta(z)\, dz \tag{2}$$

Where
x: Input data (e.g., a real face image)
z: Latent variable (compressed representation)
$\boldsymbol{p_\theta(x \mid z)}$**:** Decoder distribution (reconstructs x from z)
$\boldsymbol{p_\theta(z)}$ **:** Prior distribution

4. Threats of Deepfakes

The threats created by deepfakes are not limited to individual deception but have led to large-scale cybersecurity threats that have resulted in the degradation of trust and information integrity in digital networks. The main types of threats created by deepfakes are:

4.1 *Social and Political Threats*

The misuse of deepfake technology represents one potential avenue for manipulation of public opinion and the destabilization of democratic institutions. For example, artificial video representations of political leaders or public figures can be rapidly disseminated through social media networks providing fertile ground for the proliferation of disinformation, heightened social tension, and reduced trust in government institutions [1].

4.2 *Corporate and Financial Threats*

BEC (Business Email Compromise), along with Executive Impersonation attempts, are becoming more vulnerable to the use of Deep

Fake (audio/video) technology. The perpetrator can use a voice clone of a senior official's voice to deceive an employee into conducting unauthorized transactions or requesting other payment to a non-authorized account.

4.3 *Privacy and Reputation Risks*

People can be a target for non-consensual content creation, including fake explicit videos and impersonation. Once they are posted online on a social media platform, these deepfakes can cause irreparable damage to your name and invade your privacy! It is easy to distribute these types of materials quickly and virally using networked systems, therefore, it will make it even more difficult to remove or recover them.

5. Detection Mechanisms in Networked Systems

Detection of non-consensual deepfakes in a networked environment has unique obstacles. Video and picture files are frequently compressed, streamed, or resized, so any traditional method for detection will be less successful when used on compressed, streamed or resized files. As a result of these issues, researchers have developed numerous technical solutions for this type of detection:

5.1 *Image-Based Detection*

Various video frame analysis methodologies utilize convolutional neural networks (CNNs) (i.e., MesoNet, XceptionNet) to analyze pixel-level anomalies. High accuracy is retained with well-controlled datasets for these methodologies but there is a degradation in performance with compressed or streamed networked video data [2].

5.2 *Temporal and Motion Analysis*

In comparison to CNN methodologies, other models including 3D CNNs, long short-term memory networks (LSTM), and temporal transformers can identify inconsistencies in facial motions with respect to time caused by frame drops or latency in networked streams. These models are successful in identifying subtle movements in video captured with these network constraints.

5.3 *Physiological Signal-Based Detection*

Detection of placement errors with respect to facial movements can occur by utilizing biological indicators (e.g., an eye blink, blood flow) because they are resilient against rudimentary pixel manipulation [5]. For instance, Intel's FakeCatcher utilizes subtle signals from photoplethysmography (PPG) as an indicator of a genuine person's state on a video call.

5.4 *Audio-Visual and Multimodal Detection*

Lip-sync analysis and audio-visual correlation are useful for detecting the lack of correlation between speech and facial movements. Detecting the lack of correlation between audio and facial movements across multiple networked platforms is further enhanced by utilizing transformers (e.g., AV-HuBERT) and multimodal fusion techniques during video conferencing or live streaming when both audio and video data are present [6].

5.5 *Transformer and Self-Supervised Approaches*

Using Vision Transformers (ViT), Swin-Transfomers, and Convolutional Vision Transformers (CvT) models provide attention to global features which can improve their generalization. Training models with self-supervised methodologies allows them to adapt to unseen deepfakes without utilizing labeled datasets [7]. Specific advantages of using these methodologies in networked environments include resilience to compression and format variations and ability to scale for cloud-based deployment.

6. Prevention for Networked Digital Security

Preventing deepfake creation and the spread of deepfake content within a networked digital environment incorporates addressing both technical as well as policy-based challenges and will help to foster trust between platforms.

6.1 *Digital Watermarking and Fingerprinting*

Digital watermarking and digital fingerprinting are techniques designed to allow networked digital platforms to:

- Embed imperceptible metadata or patterns into the media at the point of creation.
- Create a unique identifier for each piece of media that can be used to trace its location through social media and streaming services.
- Verify whether a video or image has been tampered with or not after it has been compressed or reformatted in any way.
- Provide evidence as to the authenticity of the source of the video or image so that malicious deepfakes are less likely to spread undetected.

6.2 *Blockchain-Based Authentication*

Blockchain technology is a decentralized and immutably recorded ledger that can be used to verify the authenticity of media files [9]. Through this process, social media platforms or enterprise networks will be able to verify that the video or image has not been altered in any way since its creation, and the process itself has built-in resistance to tampering. It will

enable the rapid verification of the media content produced by large volumes of networked digital channels and resources [10].

6.3 *Network Security Integration*

Network firewalls and intrusions detection systems can also include AI algorithms to identify suspicious media before they become widely disseminated across the network. An organization's enterprise network policy will establish that all sensitive media must be verified prior to distribution throughout the organization.

7. Conclusion

Deepfake technology continues to pose significant threats related to digital security in today's networked environments. Deepfakes are generated using advanced technologies such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) which produce synthetic media with such realism that it can fool both human and automated detection methods. With deepfakes being rapidly disseminated via social media platforms, enterprise communications platforms, and streaming services, the potential for damage resulting from these manipulated media is substantial including increased risks of misinformation, financial fraud, identity theft, and risks to national security.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 2672–2680 (2014).

[2] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1–11 (2019), doi: 10.1109/ICCV.2019.00009.

[3] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397* (2020).

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, pp. 1–7 (2020), doi: 10.1109/WIFS49906.2020.9360895.

[5] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 46–52 (2020), doi: 10.1109/CVPRW50498.2020.00016.

[6] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5039–5049 (2022), doi: 10.1109/CVPR52688.2022.00500.

[7] A. H. Soudy, O. Sayed, H. Tag-Elser, M. A. Abdelrahman, and M. A. Moustafa, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Comput. Appl.*, vol. 36, pp. 19759–19775 (2024), doi: 10.1007/s00521-024-10181-7.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, pp. 1–14 (2014).

[9] R. Jain, B. Batra, B. Kanwer, L. Raja, and V. Bhatnagar, "Transforming network security through zero trust architecture: Principles, challenges, and future directions," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 2-B, pp. 1097–1106 (2026), doi: 10.47974/JDMSC-2649.

[10] S. Sharma, M. Sharma, V. Bhatnagar, and R. Jain, "Utilizing discrete structure and applied algebra to model and analysis network resilience against cyber attack," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 2-B, pp. 885–893 (2026), doi: 10.47974/JDMSC-2539.

Received April, 2026