Deepfake Image Detection: Methods, Datasets, Evaluation Metrics, and Research Challenges: A Comprehensive Survey

Sudha P. Patil
Swami Vivekanand Mahavidyalaya
Udgir, Maharashtra, India

Sudhir B. Jagtap
Swami Vivekanand Mahavidyalaya
Udgir, Maharashtra, India

ABSTRACT
Deepfake technology has rapidly evolved with advances in deep learning and generative models. To generate highly convincing synthetic media many advanced artificial intelligence (AI) technologies including deep learning, generative AI, auto encoders, and diffusion models are used. While AI tools and technologies can revolutionize virtual reality and digital content production, they also represent serious risks in terms of creating false information or media, identity theft, or media manipulation therefore, developing deepfake detection methods has become an important field of study within the fields of computer vision and multimedia forensics. This paper presents a comprehensive review of deepfake detection research from 2018 to 2025. The methods discussed in this work reference spatial domain and frequency domain techniques, biological signals, transformer-based models, and hybrids of the mentioned detection techniques. A wide range of benchmark datasets, detection techniques, and evaluation metrics are provided.

Keywords
Deepfake Detection, Generative Adversarial Networks (GANs), Multimedia Forensics, Computer Vision, Image Forgery Detection, Transformer Models, Frequency-Domain Analysis, Digital Media Security.

1. INTRODUCTION
Currently, the advancements made within both deep learning and generative modeling methods have transformed the process of creating digital media[13]. One of the more significant advancements in Artificial intelligence is the development of deepfake technology, which allows for the creation manipulation of images, audio and video that can nearly match the realism found in real-world images or media content. Deepfake media is chiefly created using generative adversarial networks (GANs), autoencoders, and diffusive architecture [5], [16]. These novel technologies provide constructive applications in the areas of entertainment and virtual reality, they also create significant issues regarding determining the credibility and truthfulness of digital media [24].

The availability of deep fake-generating tools has allowed malicious actors to make fake contents such as images or videos that have many harmful purposes, such as spreading false information, creating false political narratives, stealing identities, executing cyber fraud, or conducting social engineering attacks [24]. The deception and misinformation from these synthetic media will confuse and mislead users, thus composing false narratives on social media platforms [24].
For example, deepfake images might be used to impersonate an individual, produce phony evidence, and disseminate misleading information via social media. The exponential growth and development of synthetic media have brought forth a tremendous amount of concern in the areas of digital forensics and cybersecurity, journalism, and law enforcement. As a result, establishing effective techniques for detecting deepfake content has emerged as a vital research challenge.

Traditional image forensics techniques rely on detecting compression artifacts and sensor noise inconsistencies; however, modern deepfake generation methods produce highly realistic images with minimal artifacts, reducing their effectiveness [8].To address these challenges, deep learning-based detection methods using CNNs, attention mechanisms, and transformer-based models have been extensively explored [10], [19].

Available detection methods are categorized into different categories.The spatial methods analyze pixel level inconsistencies that are introduced in generation process using deep neural networks (DNNs). Frequency Domain Techniques uses the spectral artifact that appears in frequency spectra of generated images. In detection of manipulated contents, Biological signal-based methods analyze physiological signals such as eye blinking, head movements, or facial expressions. Recently the transformer based models used to capture relationships within images and providing improved detection performance. Now many studies have proposed hybrid detection methods that integrates spatial, frequency, and temporal features to enhance their robustness and generalization.

Another important aspect of deepfake detection research is the existence of benchmark datasets used to train and evaluate detection algorithms. Several publicly accessible datasets have been developed to support research in this area, including FaceForensics++, Celeb-DF, the DeepFake Detection Challenge (DFDC), DeeperForensics, and WildDeepfake. These datasets contain a diverse range of altered images and videos generated using different methods and compression levels. Although these datasets have facilitated significant advancement in deepfake detection research. In particular, models that are trained on one dataset often fail to generalize well to unseen datasets because of differences in generation techniques, compression artifacts, and data distributions.

The main contributions of this study are summarized as follows:

1. **Extensive Literature Review:** In recent years many deepfake papers are published. We reviewed and analyzed deepfake detection studies published between 2018 and 2025, exploring approaches both classical and recent deep learning-methods .

2. **Deep Fake Generation Techniques and Comprehensive Taxonomy:** This section present the techniques used in deep fake generation and a systematic classification of deepfake detection methods based on the fundamental detection principles and feature representations.

3. **Dataset and Evaluation Analysis:** It explores widely used benchmark datasets and evaluation metrics for deepfake detection.

4. **Research Challenges and Future Directions:** The key research challenges in deepfake detection and directions for developing robust and generalizable detection systems is discussed.

The structure of this paper is as follows. Section 2 conducts extensive review of existing literature. Section 3 offers an overview of deepfake generation techniques. Section 4 introduces a taxonomy of techniques of deepfake detection. Section 5 reviews the benchmark datasets and evaluation metrics that are widely used. Section 6 addresses the research challenges and potential future directions. Finally, Section 8 gives the conclusion of the study.

2. EXTENSIVE LITERATURE REVIEW

CNN-based architectures dominate the field [4], [8], while hybrid and attention-based methods demonstrate improved performance and robustness[10, 35].Table 1 summarizes different deep fake detection approaches.

Table 1. summarizes different deepfake detection approaches across image, video, and multi-modal domains.

Sr. No	Year	Authors	Paper Title	Modality	Method Used	Dataset	Metric
1	2018	Afchar et al.	MesoNet: a Compact Facial Video Forgery Detection Network	Video / Facial images	CNN	DeepFake datasets, FaceSwap-style datasets	Accuracy
2	2019	Li & Lyu	Exposing DeepFake Videos by Detecting Face Warping Artifacts	Video	CNN	FaceForensics ++	Accuracy
3	2019	Rossler et al.	FaceForensics++: Learning to Detect Manipulated Facial Images	Image/ Video	CNN (XceptionNet)	FF++	Accuracy
4	2020	Zhuang et al.	Deepfake Image Detection Based on Pairwise Learning	Image	CNN	CelebA	Accuracy
5	2020	Guarnera et al.	DeepFake Detection by Analyzing Convolutional Traces	Image	CNN	Forensics	Accuracy
6	2020	Pan et al.	Deepfake Detection through Deep Learning	Image	CNN	Multiple	Accuracy
7	2020	Jose A. Lopez et al.	Audio Deepfake Detection Generalization	Audio	RNN	ASVspoof	EER
8	2020	Shruti Agarwal et al.	Detecting Deep-Fake Videos from Appearance and Behavior	Video	CNN	WLDR, FF, DFDC,DFD	Accuracy
9	2021	Zhao et al.	Multi-attentional Deepfake Detection	Video	Attention CNN	DFDC	Accuracy
10	2021	Liu et al.	Lightweight 3D CNN for Deepfake Detection	Video	3D CNN	FF++	Accuracy
11	2021	Zhou et al.	CNN + Vision Transformer for Deepfake Detection	Video	Hybrid	DFDC	Accuracy
12	2021	Zhou et al.	Joint Audio-Visual Deepfake Detection	Multi	CNN+RNN	FakeAVCeleb	Accuracy
13	2021	Groh et al.	Crowd + Machine Deepfake Detection	Multi	Hybrid	Multiple	Accuracy
14	2022	Malik et al.	DeepFake Detection Survey	Multi	Survey	Multiple	-
15	2021	Deressa Wodajo et al.	CNN + Vision Transformer for Deepfake Detection	Video	Hybrid	DFDC	Accuracy
16	2021	Zhou et al.	Joint Audio-Visual Deepfake Detection	Multi	CNN+RNN	FakeAVCeleb	Accuracy
17	2021	Groh et al.	Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds	Multi	Hybrid	Multiple	Accuracy
18	2022	Ali Raza et al.	A Novel Deep Learning Approach for Deepfake Image Detection	Image	hybrid of VGG16 and CNN	Kaggle Deepfake Dataset	Accuracy
19	2023	Patel et al.	Improved Dense CNN	Image	Dense	Celeb-DF	Accuracy

Sr. No	Year	Authors	Paper Title	Modality	Method Used	Dataset	Metric
			Detection		CNN		
20	2023	Wang et al.	Convolutional Pooling Transformer for Deepfake Detection	Video	Hybrid	DFDC	Accuracy
21	2024	Yin Ni et al.	A Deepfake Detection Algorithm Based on Fourier Transform of Biological Signal	Video	rPPG + FFT + MVHM + Spatial Attention + CNN	FaceForensics ++, Celeb-DF, DFDC Preview	Accuracy
22	2024	Bar Cavia et al.	Real-Time Deepfake Detection in the Real-World	Image	LaDeDa, Patch-based ResNet50, Tiny-LaDeDa	WildRF	mAP WildRF
23	2024	Ahmed A. Hasanaath et al.	FSBI: Deepfakes Detection with Frequency Enhanced Self-Blended Images	Image	SBI + DWT + CNN	FF++, Celeb-DF	Cross-dataset Accuracy/ AUC
24	2025	Fazeel Zafar et al.	A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features	Video	MTCNN + EfficientNet-B0 + TempCNN + FPN	FFIW-10K	Accuracy
25	2025	Wasim Ahmad et al.	CapST Leveraging Capsule Networks and Temporal Attention for Accurate model attribution in deepfake videos	Video	Modified VGG19 + Capsule Network + Spatial-Temporal Attention	DFDM	Accuracy
26	2025	Shobana Gorintla et al.	Deepfake Image Detection and Classification–Vision Transformers	Image	ViT, Transfer Learning, Self-Attention, ImageNet Pretrained Model,	Kaggle Deepfake-and-Real-Images Dataset	Accuracy

3. DEEPFAKE GENERATION TECHNIQUES AND TAXONOMY OF DEEPFAKE DETECTION METHODS

Deepfake technology relies on advanced generative models that can synthesize highly realistic images and videos. These models learn complex data distributions from large datasets and generate new samples that closely resemble the real-world data. The rapid advancement of deep learning techniques has significantly improved the quality and realism of synthetic media, making deepfake detection increasingly difficult. Understanding the underlying generation techniques is essential for developing effective detection strategies. In general, deepfake generation approaches can be categorized into Generative Adversarial Network (GAN)-based methods, autoencoder-based face-swapping techniques, and diffusion-based generative models and transformer based architectures. These methods focuses on creating, swapping or manipulating facial expressions identity in images and videos. Face Swap technique replaces a human face , Face Reenactment mimics the movement or pose, Talking Face Generation synchronize audio or text mouth movement and Face Editing modifies the attribute such as hair color or glasses.

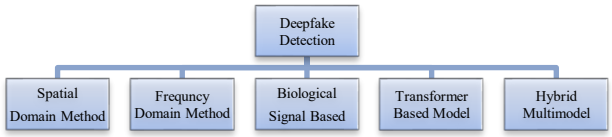

Figure 1. Taxonomy of Deepfake Detection Methods

Deepfake detection methods are categorized into spatial-domain, frequency-domain, transformer-based, and hybrid multimodal approaches.

3.1 Spatial-Domain Methods

Spatial-domain methods use CNN architectures such as XceptionNet, ResNet, and EfficientNet to detect visual inconsistencies in images **[4], [8]**. These methods learn visual inconsistencies, including unnatural textures, blending artifacts, and irregular lighting conditions. Although they

achieve high accuracy on known datasets, their generalization to unseen deepfake techniques remains limited.

3.2 Frequency-Domain Methods

Frequency-domain approaches analyze spectral artifacts using transformations such as FFT and DCT, which help identify GAN-generated inconsistencies [21].

3.3 Biological Signal-Based Methods

Biological signal-based methods analyze physiological patterns such as eye blinking and facial movements for detecting manipulated videos **[3]**. They are effective for video deepfakes but have limited applicability to static images because temporal information is required.

3.4 Transformer-Based Methods

Transformer-based models such as Vision Transformers (ViT) capture global dependencies and provide improved detection performance compared to traditional CNNs **[19]**.

3.5 Hybrid Methods

Hybrid approaches combine spatial and frequency features to improve robustness and generalization, although they increase computational complexity [35].

4. BENCHMARK DATASETS AND EVALUATION METRICS

Several benchmark datasets such as FaceForensics++, Celeb-DF, DFDC, DeeperForensics, and WildDeepfake are widely used for training and evaluating deepfake detection models [4], [6], [7], [18].

Although these datasets support research progress, models trained on one dataset often fail to generalize well to others due to variations in data distributions and manipulation techniques [6].This section reviews widely used deepfake datasets and discusses the common evaluation metrics used to measure the performance of deepfake detection models.

Table 2 Benchmark Datasets for Deepfake Detection

Dataset	Year	Type	Size
FaceForensics++	2019	Video	1.8M frames
Celeb-DF	2020	Video	5,639 videos
DFDC	2020	Video	100,000 videos
DeeperForensics	2020	Video	60,000 videos
Wild Deepfake	2020	Image	3,805 images

Evaluation Metrics

To evaluate the performance of deepfake detection models, researchers commonly use several classification metrics. These metrics measure the ability of the detection algorithms to correctly identify manipulated images.

The most widely used evaluation metrics include accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Accuracy

Accuracy measures the proportion of correctly classified samples to the total number of samples. It is defined as:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively.

Although accuracy is widely used, it may not be sufficient when dealing with imbalanced datasets.

Precision

Precision measures the proportion of predicted manipulated images that are manipulated.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

A high precision value indicates that the detection model produces fewer false positives.

Recall

Recall measures the proportion of manipulated images that are correctly detected.

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

A high recall indicates that the model can detect most of the manipulated samples.

F1 Score

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of the model performance.

$$\text{F1 Score} = \frac{2 \text{ X (Precision X Recall)}}{(\text{Precision} + \text{Recall})}$$

ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) measures the ability of a classifier to distinguish between real and manipulated images across different decision thresholds. The ROC-AUC is widely used to evaluate deepfake detection models because it provides a threshold-independent measure of performance.

5. RESEARCH CHALLENGES IN DEEPFAKE DETECTION

Despite significant progress, deepfake detection remains challenging due to generalization issues, adversarial attacks, and increasingly realistic synthetic media. The rapid advancement of generative models has made it increasingly difficult to distinguish between manipulated media and authentic content. As deepfake generation technologies continue to evolve, detection algorithms must adapt to new manipulation techniques and increasingly realistic and synthetic images.

6. CONCLUSION

Deepfake image generation has progressed a lot due to the development of GANs, autoencoders, and diffusion models. This progress has created serious challenges for digital media authenticity. This survey looked at recent deep learning-based detection techniques, including spatial, frequency, transformer, and hybrid approaches. CNN-based methods capture visual artifacts well, while frequency-domain techniques improve resistance to complex manipulations. Hybrid models that combine different feature representations have shown better detection performance across benchmark datasets. The survey also examined commonly used datasets and evaluation criteria . New methods based on transformers, explainable AI, and self-supervised learning show promise for improving robustness and understanding. Future research should focus on creating lightweight, generalized, and explainable detection systems that can manage unseen manipulations. Standardized datasets and clear benchmarking protocols are important for fair performance assessment. Overall, hybrid deep learning methods that combine spatial, frequency, and attention-based

features represent a promising way to build reliable and scalable deepfake image detection systems.

7. REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network,", Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018.

[2] H. H. Nguyen, J. Yamagishi and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," Proc. IEEE ICASSP, 2019.

[3] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," Proc. IEEE CVPR Workshops, 2019.

[4] J. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," Proc. IEEE ICCV, 2019.

[5] Lakshmanan Nataraj et al., "Detecting GAN-Generated Fake Images Using Co-occurrence Matrices", arXiv:1903.06836v2 [cs.CV] 3 Oct 2019

[6] B. Dolhansky et al., "The Deepfake Detection Challenge Dataset," arXiv:2006.07397, 2020.

[7] Y. Li et al., "Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Detection," Proc. IEEE CVPR, 2020.

[8] B. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," Proc. IEEE CVPR Workshops, 2020

[9] D. Pan, L. Sun, R. Wang, and X. Zhang, "Deepfake Detection through Deep Learning," IEEE, 2020.

[10] H. Zhao et al., "Multi-Attentional Deepfake Detection," Proc. IEEE CVPR, 2021.

[11] X. Zhou et al., "Joint Audio-Visual Deepfake Detection," Proc. IEEE ICCV, 2021.

[12] Z. Liu et al., "A Lightweight 3D Convolutional Neural Network for Deepfake Detection," 2021.

[13] Malik et al., "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18790, 2022.

[14] Raza et al., "A Novel Deep Learning Approach for Deepfake Image Detection," 2022,.

[15] Zhengjie Deng et al.,"Deepfake Detection Method Based on Face Edge Bands," IEEE, 2022.

[16] Zeyang Sha et al.,"DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Models," 2023.

[17] Yogesh Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection", 2023,

[18] Bojia et al., "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection", Proc. ACM Multimedia, 2024.

[19] D Lamichhane, "Advanced Detection of AI-Generated Images Using Vision Transformers",IEEE Xplore 2024.

[20] Pradeepan P et al., "Detection of Deepfake Medical Images Based on Spatial and Frequency Domain Analysis", IEEE, 2024.

[21] Ga San Jhun, Pyo Min Hong, Keun Lee, et al., "A New Wave of Texture Feature Enhancing Deepfake Detection via Image Waveform," IEEE Xplore 2024,

[22] Weinan Zhang et al.,"Fake Image Detection Based on Attention-Based Feature Aggregation", 2024.

[23] H. Kaur et al., "Deepfake Video Detection: Challenges and Opportunities," IEEE Access, 2024.

[24] Alessandro Gnutti et al., "Learned Image Compression for Deepfake Detection," 2025

[25] Sahar Husseinil et al.,"RAW Data: A Key Component for Effective Deepfake Detection," 2025

[26] Yikun Ji et al. ,"Towards explainable fake image detection with multi-modal large language models",

[27] R. Singh et al., "Detecting the Undetectable: Deep Learning Model for Fake Images," IEEE Conference, 2024.

[28] A. Kumar et al., "Deepfake Image Detection Using Deep Learning," IEEE Conference, 2025.

[29] A. Devi et al., "DeepGuardNet: A CNN-Based Deepfake Detection Model,"Procedia Computer Science, 2025.

[30] S. Srinivasan et al., "Deepfake Image and Video Detection Using Deep Learning Algorithms," IEEE, 2025,

[31] Ambika et al., "A Novel Framework for Deepfake Image Detection," 2025.

[32] Yash Jugade et al., "Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy," 2025.

[33] J. Vijaya et al., "Generation And Detection of Deepfakes using Generative Adversarial Networks (GANs) and Affine Transformation".

[34] Chih- Chung Hsu,Yi-Xiu Zhuang and Chia-Yen Lee, "Pairwise Learning for Deepfake Detection", MDPI 2020.

[35] Deressa Wodajo et al., "Deepfake Video Detection Using Convolutional Vision Transformer", 2021.

[36] Matthew Groh et al.,"Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds," 2021.

[37] Shruti Agarwal et al., "Detecting Deep-Fake Videos from Appearance and Behavior,"International Workshop on Information Forensics and Security(WIFS) , 2020.

[38] Wasim Ahmad, Yan-Tsung Peng, Yuan-Hao Chang, Gaddisa Olani Ganfure, Sarwar Khan, "CapST: Leveraging Capsule Networks and Temporal Attention for Accurate Model Attribution in Deep-fake Videos",ACM Trans. Multimedia Comput. Commun. Appl, 2025, arXiv:2311.03782.

[39] Gorintla, S., Tammina, V. K., Sankarabathina, C., Pasupuleti, H. K., and Yarajarla, R. V. N. S., "Deepfake Image Detection and Classification–Vision Transformers," Computer Research and Development, vol. 25, no. 6, pp. 757–766, 2025.

[40] Bar Cavia et al., "Real-Time Deepfake Detection in the Real-World,"arXiv:2406.09398v1F.

[41] Zafar et al., "A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial

Features," IEEE Access, vol. 13, 2025. Y. Ni, W. Zeng, P. Xia, G. S. Yang, and R. Tan, "A Deepfake Detection Algorithm Based on Fourier Transform of Biological Signal," Computers, Materials & Continua, vol. 79, no. 3, 2024.

[42] A. A. Hasanaath et al., "FSBI: Deepfakes Detection with Frequency Enhanced Self-Blended Images," arXiv:2406.08625, 2024.

[43] Y. Ni, W. Zeng, P. Xia, G. S. Yang, and R. Tan, "A Deepfake Detection Algorithm Based on Fourier Transform of Biological Signal," Computers, Materials & Continua, vol. 79, no. 3, 2024.