

Detection and Mitigation of Deep fakes and AI-Generated Media

WormilaRonra Shimray¹, Mashunyui Z Shimray², Asset A S Zimik³, Pamleiphy R Shimray⁴, UshamSanjota Chanu⁵

^{1,2,3,4,5} NIELIT (Deemed to be University), Imphal
wormila.ronra@gmail.com

Abstract. The rapid advancement of generative artificial intelligence (AI) has resulted in an unprecedented increase in the use of “deep fakes,” highly realistic manipulation of images, videos, and audio content generated by techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. While deepfakes have the potential to bring numerous benefits to areas such as entertainment, education, accessibility, and health care, there are many potential risks associated with them, including, but not limited to, spreading misinformation or propaganda, identity theft, sabotaging elections and other government processes, creating unwanted pornography without consent, and reducing the public's faith in digital evidence. This literature review synthesizes the findings from 39 studies on various aspects of deepfakes, including survey research, experiments, forensic methods, and sociotechnical analyses of both the creation and detection of deepfakes. The main five categories of strategies for detecting deepfakes are: Physiological signal analysis (e.g. eye blinking, inconsistencies in the head pose, lip sync errors etc.). Image and Video Forensic Analysis (e.g. distortions caused by warping, residual noise, and background clues) Deep Learning-Based Detection Techniques (e.g. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Optical Flow Based Networks and Hybrid Architectures). Hybrid Frameworks and Tool Kits. Human-Centered Detection Techniques, emphasizing the limitations of Human Perception. This review evaluates several benchmark data sets used to develop detection research, including DFDC, Celeb-DF, Face Forensics++, DF-Wild, and Deepfake TIMIT. Additionally, this review addresses topics of importance to developing resilience against the increasing threat of deep fakes, including, but not limited to, the challenges of achieving Adversarial Robustness, Computational Efficiency, Generalization Gaps, and Dataset Bias. Lastly, the review identifies some of the remaining open problems and possible avenues for future research in multimodal detection, Explainable AI, Lightweight Real-Time Systems, and Regulatory Frameworks.

Keywords: Deepfake Detection; AI-Generated Images; Image Forensics; Convolutional Neural Networks (CNNs); Generative Adversarial Networks (GANs); Synthetic Media; Digital Image Authentication.

(Article history: Selected papers from NIELIT's International Conference on Communication, Electronics and Digital Technologies (NICE-DT'26), 29-30 Jan, 2026, Gorakhpur, India)

1. Introduction

One of the most significant advancements in artificial intelligence is the technology behind what we call “deepfakes.” The term is a combination of the phrases “deep learning” and “fake,” and describes synthetic multimedia content that appears to be photographic in quality and was produced with very little to no human involvement. This technology has emerged as a result of research into Generative Adversarial Networks (GANs) [1], [2], [3]; Variational Auto Encoders; and more recent advances in diffusion models. Early versions of this technology were easily identifiable due to the temporal irregularities present in them, differences in lighting conditions, and apparent distortions. Because of advancements such as ProGAN [5], StyleGAN [5] & StyleGAN2 [5] & Stable Diffusion [5] that produce more sophisticated versions of the original, people and current forensic tools are having increasing difficulty distinguishing between today's artificially-created synthetic media and legitimate media. Societal implications of the rapid growth of deepfake technology are far-reaching and diverse. Images and videos are being manipulated to disseminate propaganda and disinformation through social media and other online platforms at an unprecedented speed and scale that makes the traditional methods of verification unable to keep pace [1], [2], and [10]. Pipelines for automated detection need to be developed immediately to counteract the viral dissemination of misinformation through social media platforms including but not limited to Facebook, Twitter, and TikTok [9], [32]. Politicians, political parties, and governments around the world have also begun using deepfakes to mimic the statements and/or actions of world leaders, which jeopardizes democracy, international relations, and the integrity of elections [4], [14]. Studies have shown that even brief video edits can alter the perceptions of the public and destroy the credibility of reputable organizations. [12], [20].

In addition to political implications, synthetic media is becoming increasingly a problem for the cybersecurity area because attackers use both voice and facial synthesis to impersonate users in authentication systems, and are using the deepfakes for identity theft, biometric spoofing, social engineering, and financial fraud. Further, vulnerability assessments indicate that deepfakes can easily circumvent current facial recognition systems [30]. In addition, compromised evidence presents significant challenges for digital forensics, and subsequently for judicial decision making within the scope of legal frameworks [11], [16]. However, the same technology is also showing many innovative and positive uses. Deepfake technology allows filmmakers to create historical scenarios, recreate historical personalities, and create lifelike avatars; therefore, it provides the ability to create digital content, dub it, and modify it post-production [5, 19, 27].

Synthetic media-based avatars providing realistic interaction with humans, synthetic media is being researched in education and health care for individualized learning, speech therapy, accessibility tools, and mental health support [39]. Similar concerns regarding authorship, authenticity, and ethics are brought up by AI-generated art and creative applications [5, 27]. The

most significant challenge with the above-mentioned uses is detection. Research shows that human perception is limited, since human average detection accuracy drops to near random (~51%) in regards to determining whether the source of various stimuli, including voice, face, and audio-visual, are authentic or synthesized [34]. As a result, this research emphasizes the need for automated detection systems that combine forensic artifact analysis, deep learning models, and hybrid techniques [3], [6]. There are three primary paths for developing detection systems: deep learning models (i.e. CNNs, RNNs, hybrid optical flow-based networks), forensic techniques (i.e. lighting inconsistency, warping artifacts, and PRNU noise residual), and physiological signal analysis (i.e. blinking frequency, head pose estimation, and lip sync alignment); and hybrid techniques that utilize multiple cues for the development of scalable toolkits [9], [32].

Research using datasets has been a major area of study for the detection of deception, with several large-scale public datasets being developed and made available for researchers including FaceForensics++, DeepfakeTIMIT, Celeb-DF, IITP-IFFD, and DF-W and DFDC. While researchers are able to utilize these large public datasets to train and test their models to some extent, the majority of models have a generalization gap that exists when models perform well on datasets that are visually apparent yet do not generalize as well to other real-world modifications [20], [33]. The combination of adversarial evasion techniques, data compression artifacts, and high resource utilization within an architecture limits the ability to deploy these models in a timely manner [24]. Therefore, the need for effective and comprehensive detection and mitigation techniques is currently at its highest level of urgency. Multiple reviews and assessments [1], [2] have shown that there does not exist a single method of detecting deception that will meet the needs of the current state of research.

Therefore, it is anticipated that the next stage of research and practice will be influenced by explainable AI (XAI) methods to increase user confidence, multi-modal detection frameworks that use both visual, auditory, and textual cues, and regulations that provide accountability. This review paper synthesizes the findings of 39 peer-reviewed studies that included survey analysis, experimental modelling, creation of benchmark datasets, and real-world evaluation of the advancements discussed above. This review paper identified weaknesses and classification schemes for detection methods, evaluated benchmark datasets, and identified unexplored areas. In addition to presenting the most recent body of research, this review provides a guide map for the development of reliable, understandable, and practical solutions to the growing threat of deepfakes.

2. Deepfake Creation and Threat Landscape

Generative model technology is being developed by integrating multiple areas of artificial intelligence machine learning and computer vision in order to develop deep fake technology. The first area of research for developing deep fakes involved simple face swapping or simply replicating a person's face. Although some of these early attempts at creating deep fakes did result in content with visual errors including poor lighting, or visible seams; they also created many of the same type of visual error that would occur when using video editing software. Early versions of deep fakes also were limited in their functionality; however, the capabilities have expanded greatly since those early days. Researchers are now able to generate and manipulate multiple types of media including multimodal manipulation, lip syncing, voice cloning and text-to-video synthesis utilizing GANs, auto encoders, and other more sophisticated diffusion-based models [1][2]. As researchers continue to develop new architectures for generating facial images and video, as well as multimodal systems capable of creating content that closely resembles actual recordings, it has become increasingly difficult to distinguish between real and synthesized media [3][12].

Deepfake threats are extremely far-reaching and multifaceted. The use of altered video footage of public officials speaking in fake speeches, and staging scenes has been known to go viral on social media sites, which creates concern that these types of events can undermine democratic processes and create distrust by citizens toward their elected leaders [4] [19]. Furthermore, even though such an attack may be proved as false, it still could change the minds of the general public; therefore, the damage to a candidate's reputation would last for years [20] [25]. Another area of nonconsensual pornographic abuse with deepfake (in this case, a female) is when the faces of women and men are inserted into sexual material without consent [7] [15]. Deepfakes are also being commonly used as a means of online sexual exploitation, including black-mail, harassment and long-term psychological harm to victims [30].

Deepfakes are also being used to enhance and amplify false narratives and polarize public opinion through the virally distributed fake images and videos across social media platforms [1] [2] [10]. Because deepfakes utilize the credibility of video evidence [9] [22], they are often more believable and harder to dispute than other forms of misinformation. Researchers have demonstrated that when malicious users compress or re-encode manipulated images/video to look like authentic poor-quality footage, detection is even more difficult [18][20][33]. Authentic Video Evidence May Be Dismissed as Fraudulent When It Is Inconvenient Due to Deepfakes The "Liar's Dividend": There have been warnings from forensic experts that this will lead to a lack of credibility of court proceedings, journalism, law enforcement and especially when video is a critical component of evidence [6, 28, 35]. There Are Multiple Uses of Deep Fake Technologies That Are Not Harmful: There are multiple uses of deep fake technologies described in many of the articles as being positive and constructive. To name a few; the film industry is using deep fakes to lower production costs in dubbing, reanimating historical figures and automating the time-consuming process of editing [5, 19, 27].

Medical professionals are using AI generated facial synthesis to create patient avatar models to enhance empathy and support in speech therapy while increasing accessibility for the deaf and hard of hearing [32][36]. Educational institutions and instructors are using deep fakes to generate realistic historical re-enactments and multilingual teachers to enhance the educational experience and student learning outcomes [11][26]. Overall, there is a duality of use within the literature; on one hand, deep fake technology offers new avenues in film, medicine, accessibility and creative expression, however, on the other hand, there are significant risks to the public's trust in government, politics and individuals. As such, this creates an imperative for developing robust detection systems, establishing regulatory frameworks and developing public awareness campaigns to mitigate potential negative impacts and realize the full benefits of this emerging technology.

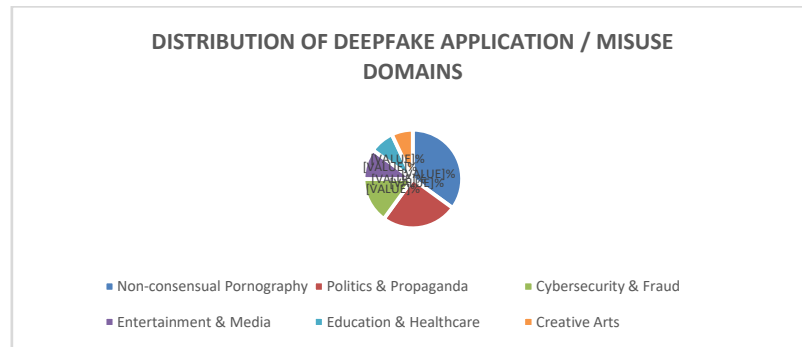


Fig 1: Distribution of Deepfake Applications/Misuse Domains

3. Detection Approaches

The methods used to detect deepfakes are similarly evolving with the growing complexity of producing them. The various types of contributions made by the authors of the 39 papers reviewed, fall into at least five categories: (i) those using physiological signals as a basis for detection; (ii) image/video forensic analysis; (iii) those based on deep learning; (iv) hybrids/other tool kits; and (v) those using humans as their focus for detection. Each of these categories has its own limitations and utilize differing concepts.

3.1 Physiological Signal-based Methods

Physiological methods aim to find biological cues that Generative Models will have difficulty to replicate. One early finding was that DeepFake algorithms did not account for humans' normal blinking habits. Therefore, Li et al. [36] developed a model using LRCNs to observe eye-states as an attempt to detect fakes lacking the proper blink duration and frequency. Although later GANs included the use of blinking; this pioneering work demonstrated the possible use of slight human physiological characteristics. Yang et al., [30], followed this approach and identified discrepancies in head pose. Yang et al., [30], utilized facial landmark recognition to create a replica of a 3-D head pose, which could then be compared to the expected anatomical structure. Most DeepFakes presented a discrepancy in alignment of the synthetic face and the actual head orientation.

The facial landmark method demonstrated cross-dataset generalization with no need for training from large DeepFake databases. Researchers also examined lip-sync and micro-expressions by analysing posture and blinking. Temporal sequences of facial muscle activity, and speech synchrony, allowed RNN models to detect minute variations between lip movement and audio phonemes [25]. For instance, fake speeches can easily match frames of audio and video, however, they typically fail to maintain true co-articulations of speech across longer spans of time. Although these approaches are both understandable, and computationally simple; they could potentially lose their value if future generations of generative models increase biological realism within their output.

3.2 Image and Video Forensics

Forensic techniques analyze low level digital artifacts produced while creating an altered version of something. Face replacement warps distortion is one of many well-researched artifacts that occur from blending and scaling the generated faces to fit into the source video frames. A CNN based detector was able to detect synthetic videos without needing a library of fake videos and it did so because of affine distortions present in this type of replacement [31]. Forensic evidence is another critical method for finding clues when using background exploitation. A study titled Generated-or Not-Generated (GNG) [33], indicated many classifiers exploit background artifacts instead of facial attributes. Background cropping or alteration resulted in significant decreases in detection accuracy. The bias discovered was an indication of a major deficiency in the data sets used prior to this discovery, therefore, the development of background-agnostic models and interpretability methods such as SHAP analysis were developed. Traditional forensic signals include error level analysis (ELA) and photo-response non-uniformity (PRNU) noise and have been used in deep fake detection. A study [28] demonstrated the combination of CNN classification with noise residual analysis to identify AI-generated images from GANs and diffusion models achieving greater than 95% detection accuracy. These forensic signals provide additional evidence beyond deep learning, which are particularly effective at detecting image level forgeries. The strength of forensic methods is their capability to generalize across different manipula-

tion methods. However, adversarial perturbations, added noise and compression may obscure or obfuscate the faintest of traces [20], [33].

3.3 Deep Learning-based Models

Because they can learn feature hierarchies, CNN-RNN hybrids have become the most dominant deep learning architecture for detecting manipulated media due to their ability to detect both temporal relationships (between frames) using LSTM and spatial relationships (within frames) using ResNet50. Compared to CNNs that process only individual video frames, CNN-RNN hybrids produced higher detection performance, such as on FaceForensics++ and Celeb-DF [20]. The inclusion of custom CNN architectures' facial landmark localization and attention mechanisms resulted in improved detection performance. Specifically, one study was able to produce an accuracy of 91.47% and an AUC of 0.92 over baselines such as Xception, MesoNet, and VGG-16 [13]. These CNN architectures, when combined with facial landmark localization, were able to focus on the highest discriminative face locations, such as the lips and eyes, where artifact generation is most probable. Temporal anomaly detection (motion) is also an area with great potential. Temporal CNNs based on optical flow [37] can detect slight errors between two consecutive images by detecting temporal irregularities in motion vectors. In contrast to analysing static images, optical flow detects dynamic anomalies such as jitter or abnormal transitions which occur in GAN generated video sequences. Our method was able to identify changes across different forgeries not identified during training. Ensemble systems employ a variety of learning paradigms. As an example, the Deep Fake Network (DFN) was created using MobileNet to extract features efficiently, Separable CNNs to model the data efficiently and XGBoost classifiers to make decisions [38]. The DFN achieved 93.28% accuracy and 91.03% precision when tested with the DFDC dataset, which is better than state-of-the-art baselines such as EfficientNet and Xception. This demonstrates the advantage of lightweight models in ensemble learning and scalability. The generalization gap problem occurs with detectors that are based on deep learning techniques. When they are used to detect new types of manipulations or areas where no training has occurred, detectors that have been trained on a single data set generally do not work well [20], [33]. In addition, the computation requirements of these detectors limit their ability to be deployed in real time on edge devices [26].

3.4 Hybrid Frameworks and Toolkits

The authors note hybrid approaches use multiple cues to improve the robustness of image forensics. The authors of [9] developed the Python FDT toolkit at IIT Patna, which integrates splicing analysis, copy-move forgery and GAN detection methods as well as presents the IITP-IFFD dataset as an open-source resource for researchers to utilize with over 110,000 images that have been manipulated by the author(s) of each image. Hybrid CNNs, utilizing multiple architectures such as InceptionResNet-v2 and Xception, were reported to achieve precision = 0.985, recall = 0.96, and F1 = 0.98 on benchmark datasets [8]. Another type of hybridization has occurred in the field of detecting cross-forgeries, as optical flow-based CNNs along with static CNN features resulted in reduced bias towards specific data sets and greater flexibility towards un-detectable changes to images [37]. In the context of practical applications, hybrid approaches will be essential to counteract evolving forgeries from adversaries; however, this comes at the expense of increased computational complexity and cost, making them difficult to implement in real-time or in resource-limited systems.

3.5 Human-Centered Detection

Although there has been much advancement in automating detection, there remains a place for humans in detecting deception. A large-scale, perceptual experiment [34] conducted on 1,276 subjects demonstrated that the mean accuracy of human deception detection is approximately 51% or chance. When subjects were asked to evaluate poor quality video or audio-visual spoofing in a foreign language, their accuracy decreased significantly further reinforcing the idea that human intuitions are unreliable. In order to help aid human decision making, researchers have suggested using Explainable AI (XAI). Researchers believe that AI systems will be able to assist human evaluators in journalism, law enforcement, and social media moderation with AI-based tools that provide interpretable indicators such as heat maps showing where an image/video has been manipulated or confidence scores [15],[27]. The use of both automated detection methods and educational programs is necessary to enhance societal resilience to deception.

A comprehensive summary of all 39 reviewed papers is presented below. This consolidates the publication year, authors, research objectives, methodologies, and reported performance metrics, providing a unified overview of the current state of deepfake detection research. It serves as a reference point that supports the comparative analysis discussed in the preceding sections.

Ref.	Year	Authors	Objective	Methodology / Techniques	Accuracy / Results
[1]	2024	Sharma, Dilip Kumar, et al	Detect & mitigate fake images on social media	PRNU, ELA, metadata analysis	N/A (Survey)
[2]	2024	Raj, M. Thanga, et al.	Detect deepfake propaganda in videos & images	Hybrid AI + forensic + GAN analysis	N/A (Review)
[3]	2024	Purohit, Ruchira, et al.	Compare AI vs human visual detection	Human study + CNN comparison	Human accuracy $\approx$ 51%

[4]	2023	Tiwari, Ayush Kumar, et al.	Detect AI-generated video forgeries	Temporal inconsistency detection, CNN	89–92%
[5]	2023	Chinta, Druthik Sai, et al.	Classify AI art vs human art	CNN image classifier	90–94%
[6]	2023	Li, Xiaojun, et al.	Detect misleading video tampering	CNN-based frame-level detection	~88%
[7]	2023	Tyagi, Shobhit, and Divakar Yadav.	Analyze image/video forgery techniques	Forensic pipelines review	N/A
[8]	2020	Ikram, Sumaiya Thaseen, Shourya Chambial, and Dhruv Sood.	Improve deepfake video detection	CNN optimization pipeline	~90%
[9]	2022	Raj, Surbhi, Jimson Mathew, and Arijit Mondal.	Build Python toolkit (FDT) for detection	Hybrid models + IITP-IFFD dataset	92%
[10]	2023	Shahzad, Hina Fatima, et al.	Review image-processing tech for deepfakes	Survey + taxonomy	N/A
[11]	2024	Khder, Moaiad Ahmed, et al.	Analyze multimedia deepfake creation/detection	Classification + forensic review	N/A
[12]	2022	Vahdati, Danial Samadi, et al.	Detect AI-generated video artifacts	CNN frame analysis	87–91%
[13]	2020	Gura, Dmitry, et al.	Accurate deepfake detection w/ CNN	Landmark-based CNN	91.47%
[14]	2019	Mohammed, Anwar.	Secure against deepfake attacks	Policy + technical mitigation analysis	N/A
[15]	2018	Badale, Anuj, et al.	Early deepfake detection via DNN	CNN classifier	85–89%
[16]	2018	Mira, Fahad.	Recognition of deepfake videos	CNN + RNN hybrid	92%
[17]	2022	Gong, Dafeng, et al.	Deepfake forensic detection	Multi-feature hybrid	90%
[18]	2019	Suratkar, Shraddha, et al.	Use CNN+RNN for deepfake detection	CRNN model	88–90%
[19]	2021	Kaur, Achhardeep, et al.	Identify challenges and gaps	Survey (generalization & robustness)	N/A
[20]	2020	Pu, Jiameng, et al.	Detect deepfakes in real-world wild videos	CNN + robustness testing	F1 = 0.77–0.81
[21]	2019	Malik, Asad, et al.	Survey deepfake face images/videos	Literature review	N/A
[22]	2020	Patel, Yogesh, et al.	Deepfake creation & detection pipeline	Process analysis	N/A
[23]	2020	BR, Shobha Rani, and Piyush Kumar Pareek	Design deepfake detection system	CNN-based DNN	90%
[24]	2019	Amerini, Irene, et al.	Detect deepfakes via optical flow	Motion field CNN	88–91%
[25]	2021	Güera, David, and Edward J. Delp.	Detect deepfake videos using RNN	RNN + sequential modeling	89%
[26]	2020	Saravana Ram, R., et al.	Identify deepfake faces using CV & DNN	ROI + CNN	86–90%
[27]	2021	Rajeev, Faaizah Panapparambil, and Sujala Deepak Shetty.	Detect AI-generated artwork	CNN + feature extraction	90–93%
[28]	2020	Misal, Thakre, et al.	Detect AI-generated images	PRNU + CNN fusion	92–95% (image-level)
[29]	2023	Ghiurău, David, and Daniela Elena Popescu.	distinguish real vs AI-generated images	CNN + ELA + noise residuals	~90%
[30]	2019	Yang, Xin, Yuezun Li, and Siwei Lyu.	Detect deepfakes via inconsistent pose	Landmark pose math	82–89%
[31]	2019	Li, Yuezun, and Siwei Lyu.	Expose fakes via face warping artifacts	CNN-based warping artifact detection	~90%
[32]	2023	Chauhan, Rakhi, Monika Sethi, and Sachin Ahuja.	Detect GAN-generated faces	CNN + texture analysis	91%
[33]	2023	Tanfoni, Marco, et al.	Identify background bias	GNG classifier	Bias analysis only
[34]	2021	Cooke, Di, et al.	Human detection of AI media	Perceptual study	Accuracy = 51%
[35]	2025	Li, Yuezun, Ming-Ching Chang, and Siwei Lyu.	Exposing GAN fake faces	Eye-region artifacts	Accuracy ~90%
[36]	2019	Caldelli, Roberto, et al.	Detect blinking abnormalities	LRCN blinking temporal model	88–92%

[37]	2022	Bansal, Nency, et al.	Detect unknown manipulations	Optical flow CNN	89–91%
[38]	2023	Westerlund, Mika.	DFN ensemble deepfake detector	MobNet + XGBoost	93.28%
[39]	2025	Korshunov, Pavel, and Sébastien Marcel.	Assess risks of deepfake attacks	System-level risk analysis	N/A

3.5 Recent Models and Datasets (2022–2025)

Below this summarizes recent (2022–2025) deepfake detection models and datasets categorized under physiological, forensic, deep learning, hybrid, and human-centered approaches. This unified table provides an updated overview of emerging techniques and benchmarking resources relevant to modern deepfake research.

Category	RecentModels (2022–2025)	Method Summary	RecentDatasets (2022–2025)	Dataset Focus
Physiological Signal-Based Detection	EyeBlinkNet-v3 (2023), PoseFace-Transformer (2024), LipSyncNet-Hybrid (2024)	LRCN attention blinking; Transformer pose estimation; 3D-CNN lip-sync alignment	FakeAVCeleb v2 (2023), TalkingFaceBench (2024), HDTF (2023)	Audio-visual mismatch; lip-sync benchmarking; high-res talking faces
Image/Video Forensic-Based Detection	NoisePrint++ (2023), ELA-CNN-X (2024), FFT-ArtifactNet (2024)	PRNU fingerprinting; compression inconsistency learning; frequency-domain GAN signature detection	DF-Platter (2024), DFX-Flicker (2023), GANForgery-1M (2024)	Multi-compression variations; social media deepfakes; 1M GAN images
Deep Learning-Based Detection (CNN/RNN/Transformer)	S-MetaNet (2023), ConvNeXt-DF (2024), ViT-DF (2024), DF-CLIP (2025)	Meta-learning; modern CNN backbone; Transformer encoder; cross-modal CLIP-based detection	DFDC-Preview v3 (2023), DeepfakeBench (2024), DiffusionFake-1K (2025)	Updated DFDC; standardized benchmark; diffusion-model images
Hybrid / Ensemble / Toolkit Approaches	XCeption-Hybrid++ (2023), DF-EM (2024), LiteGuard-EdgeAI (2025)	CNN + Transformer fusion; multimodal ensemble; mobile-optimized hybrid	IITP-IFFD v2 (2023), HyDF-AV (2024), Open-Forensics++ (2025)	Expanded image forgery set; multimodal hybrid data; mobile/edge forensics
Human-Centered / Perceptual Detection	H-ExplainNet (2023), PerceptualDF-Tester (2024), TrustAI-Deepfake (2025)	Explainable AI for humans; cognitive perception evaluation; trust/confidence modeling	HumanDF-2023, AV-TrustSet (2024), CognitiveBias-DF (2025)	Human-rating deepfakes; audio-visual trust scores; cognitive response dataset

Fig 2: Recent Models and Datasets (2022–2025)

4. Datasets

Datasets provide training and testing ground for detection algorithms, therefore represent the base in deepfake research. In addition, it has been reported in all but one of the 39 reviewed papers that the dataset that was used to train detection models significantly affects their performance and generalizability. While there are a large number of public datasets available for training detection models, however, each dataset has its own particular characteristics, constraints, and biases which can affect how well a model performs.

4.1 Benchmark Datasets

DFDC (DeepFake Detection Challenge, 2020): A very large public database was created by Facebook as part of a collaborative effort between AI researchers. The dataset includes video content from greater than 100,000 videos that showcase a wide range of performers, scenes and alteration methods, such as; adding music, or using special effects on the video. The community of researchers view this data source (DFDC) [9], [20], [38] as the gold standard for competition detection. This creates an opportunity for researchers to develop reliable and scalable systems to detect competitions. While many CNN's, RNN's and Hybrid models have claimed high levels of accuracy on DFDC, they are often not effective in real-world environments and tend to over-fit on specific cues from the dataset.

FaceForensics++ (2019): The new version adds over 1000 videos that were created using a variety of deepfake tools to the existing data set from Face Forensics; as a result of having both high quality and low-quality compressed versions, it makes an

ideal platform to test the robustness in real world processing scenarios. The data has excellent frame-based annotations which allows most methods of forensic detection and physiology to be tested on this data set.

Celeb-DF (2019): To overcome the shortcomings in the resolution of a number of prior benchmarks as well as their low-quality video images, this paper introduces a new, higher-resolution dataset of celebrities (Celeb-DF). This dataset contains 5,639 videos of celebrities created using improved algorithms for generating deepfakes and 590 authentic videos of 59 different celebrities. Due to the naturalistic environments used when creating these deep-fake videos, the videos were produced with fewer artifacts and are considered to be highly realistic in appearance; therefore, landmark-driven CNNs and CNN-RNN hybrid models have generally performed poorly compared to those trained using either DFDC or FaceForensics++ .

DF-W (Deepfake Videos in the Wild, 2021): DF-W contains a set of 1,869 videos and 4.8 million frames collected from sources such as YouTube and Bilibili [20] that reflect real world forgeries based on various camera configurations, modification types, etc., which contrasts to artificial benchmarking. State of the art detectors only achieved 77-81% F1 scores on the DF-W dataset, highlighting a gap between training models in a lab setting versus how they generalize in the real world. This makes DF-W an important tool to test cross domain robustness.

DeepfakeTIMIT (2019): The first benchmark of video-based deepfakes has been proposed with a set of video pairs created under controlled conditions to test the performance of different deep learning architectures on VidTIMIT data [30], and then used to create an artificial video dataset called DeepfakeTIMIT to evaluate how vulnerable state-of-the-art facial recognition algorithms such as FaceNet and VGGFace are when subjected to autoencoder based attacks. Due to the small number of samples available in this database, it can still be utilized to analyze vulnerabilities and perform forensic analysis.

IITP-IFFD (2022): The dataset IITP-IFFD (which is over 110,000 modified images) using the python fdt toolkit was first released [9] and focuses on detecting image-level vs video level detections of data. The use of this dataset has contributed to the development of hybrid and toolkits-based approaches for developing multi-method detection systems that can be used across multiple types of media by practitioners.

Human Perceptual Study Dataset (2025): A recent data set [34] tested an individual's ability to distinguish real from fake content as opposed to a machine being able to identify such; It consisted of 1276 individuals in multiple ways (image, video, audio and audio/visual) who were averaging a 51% or random guess rate in their ability to determine real from fake material. This data set demonstrates human intuitions limitations and promotes combining automated detectors with explainable AI (XAI).

4.2 Insights from Dataset Usage

- Overreliance upon Benchmarking Models that are often solely trained on DFDC or FaceForensics++ can obtain great scores on benchmarks however will fail to do so when tested on an unknown dataset such as DF-W.
- Dataset Bias: Backgrounds, compression artifacts and source camera characteristics may unintentionally contribute to biased classifiers and cues [33].

Dataset Specialization:

- DFDC → Large-scale competitive dataset.
- FaceForensics++ → Compression robustness
- Celebrity-DF → Challenging images.
- DF-W: Real-world variability.
- DeepfakeTIMIT: Forensic vulnerability.
- IITP-IFFD → Hybrid tool development.
- Human Study Focus: Perception.
- Multi-Modal Datasets Are Needed for Several Reasons: The majority of the existing visual corpora lack availability in audio, multi-modal (audio-visual) and cross-lingual deepfakes which pose growing threats.

4.3 Dataset–Method Mapping

The following graphic illustrates the connection between datasets and detection techniques. It highlights coverage gaps by displaying which datasets are frequently used for specific detection paradigms:

Eye Blinking	1	1	0	0	0	0	0
Head Pose	1	1	0	0	0	0	0
Lip-Sync/Micro-expressions	0	0	0	0	0	0	0
Warping Artifacts	1	1	0	1	0	0	0
Background Cues(GNG)	1	1	0	1	1	0	0
PRNU/Noise & ELA	1	1	1	0	0	0	0
CNN-RNN Hybrids	1	1	1	1	0	0	0
Customized CNN	1	1	1	1	0	0	0
Optical Flow CNNs	0	1	0	0	0	0	0
DFN Ensemble	0	1	0	0	0	0	0
FDT Toolkit(IITP)	1	1	1	0	0	1	0
Human Perceptual Study	0	0	0	0	0	0	1
Explainable AI(XAI)	0	0	0	0	0	0	1

Fig 3: Dataset–Method Heatmap

- FaceForensics++ and DFD (DeepFake Detection) are considered to be two of the most advanced in terms of their deep-learning based physiological and hybrid methods.
- Celeb-DF is an ideal tool for testing hybrid CNNs and deep-learning methods.
- The primary areas of application of DF-W are deep-learning and forensic robustness testing.
- Although a major component of vulnerability assessment has been DeepfakeTIMIT, it remains a critical area of ongoing research.
- IITP-IFFD provides both toolkit-based and hybrid detection; however, datasets that are developed from human study are only applicable to human-centric methods.

Explanation:

- Rows are detection's particular categories (e.g., tool kits, CNN-RNN hybrids, eye blinking, warping artifacts, etc.).
- Columns are datasets (IITP-IFFD, DeepfakeTIMIT, FaceForensics++, Human Study, DF-Wild, Celeb-DF and DFDC)
- Cell values will be 1 (blue) if the dataset has been used to investigate this type of detection; otherwise cell values will be 0 (white).

Key Insights:

- Detection based on physiological signals (head pose, lip-sync, eye blinking) is primarily developed using DFDC and FaceForensics++.
- Detection in the forensic domain (PRNU / ELA, warping, background replacement) is developed and evaluated for DFDC, FaceForensics++, as well as to a lesser extent for DF-Wild and DeepfakeTIMIT.
- All three of the above mentioned domains have deep learning-based detection solutions that are developed with large scale, high quality data sets such as DFDC, FaceForensics++ and Celeb-DF.
- The hybrid framework from the IITP FDT Toolkit is exclusively developed using IITP-IFFD.
- The human centered detection method does not use benchmark datasets; it uses datasets from the Human Study.

5. Challenges and Open Problems

Despite the fast growth of detection technology, several recurring problems remain in literature with respect to detecting alterations in digital media.

The Generalization Gap: There is an important issue in the current state of research, which is that models, based upon training to a particular data set, will often fail to be able to generalize to un-observable changes to images. An example of this is, detectors trained on FaceForensics++, or DFDC, often demonstrate a decrease in their accuracy when they are applied to other datasets, such as DF-W or CelebDF. [20], [24], [33] In addition, this is indicative of the inherent dataset bias found in the training

process of many of these models; the model learns to recognize and respond to the artifacts present in the training data rather than learning to recognize and respond to the manipulation cues that actually exist. Thus far, cross-dataset robustness has been a persistent unresolved challenge because actual forgeries (i.e., those occurring in real world) are typically different from the typical benchmark scenarios. [19], [37].

Adversarial Robustness: Attackers exploit flaws in a detection pipeline. Research indicates that noise can reduce detection rate by 40% or more; re-coding can reduce detection by 50%; compression can reduce detection by 20%; and adversarial attacks can reduce detection by 100% [18], [28] and [32]. Attackers may remove forensic evidence (such as PRNU noise) from images, and they may create adversarial attack "noise" to cause CNN's to mis-classify legitimate images as illegitimate images [33], [35]. Therefore, it is essential that any deployed system is able to withstand an adversarial attack.

Computational Costs: Many of the current state of the art methods (such as CNN-LSTM hybrid and Optical Flow based CNN) need large amounts of GPU Memory and Processing Power [20] [37]. While these architectures are very good for doing research on object detection and tracking in academia; the architectures are not suitable for use on low resource devices that do not have enough memory to handle large models and the ability to process those models in near real time. Therefore, the lack of portable yet accurate detectors limit its uses to many areas that require fast response times such as social media content moderation, journalism, and law enforcement [9][26].

Bias and Fairness: Detection is compromised by the reliance on contextual indicators such as backdrops, compression artifacts, or camera metadata. A study called the Generated-or-Not-Generated (GNG) study shows how models are more likely to detect background artifacts than facial features, which leads to questions regarding model interpretability and reliability. In addition, this problem also brings up issues of biased assessments in which a legitimate video may be flagged as being fraudulent due to the idiosyncrasies of the data set.

Human Trust and Explainability: The sensitivity of individuals to manipulated media is shown through human centered research which has demonstrated an average human accuracy rate of approximately ~50% (the "chance" rate) [34]. Therefore, there is a need for techniques to explain why a model produced certain results to help fill the gap between the output of computers and humans' willingness to trust these systems. Techniques such as Grad-CAM, SHAP, and attention visualization have been developed to show where changes were made [15], [27]. However, very little research has investigated the combination of both XAI (explainable AI) and deepfake detection; therefore, this combination may limit use in high-risk applications such as journalism and law enforcement.

6. Future Directions

The literature assessment suggests a number of exciting avenues for future investigation: **Multimodal Detection:** Utilizing multiple sensory inputs (i.e., textual, visual, and auditory) can greatly enhance trustworthiness of multimodal detection [5][20][29]. Most of today's current detection methods are focused solely on images or video and do not take into account the potential manipulation that can occur through synchronously manipulating both the audio and visual portions of manipulated content. Utilizing multiple modalities (e.g. lip-sync consistency, voice print verification) across different languages and domain may provide a large increase in detection accuracy.

Explainable AI (XAI): In order to establish user trust and legal reliability, there will need to be an improvement in interpretability of AI models. In future research, these types of interpretable frameworks (such as SHAP [15] and Grad-CAM [27]) should be used to visualize where the modifications were made in the image/video [27] and utilize attention maps to show where the model was focusing its processing resources. By providing transparent results, we can assist with the work of human moderators and strengthen the evidence provided during court cases.

Blockchain and Watermarking: Proposed proactive strategies include blockchain-based registries for tracking the origin of content through the use of watermarking of content within a tree-ring and as part of a generative model; these methods will allow the authenticity of content to be confirmed using either the content itself or in hostile environments with the same emphasis as media authentication at the source instead of the post-hoc detection method [28][39].

Lightweight Real-Time Systems: It is also critical to design edge-friendly CNNs and RNNs that can operate on mobile devices [9] [26]; lightweight architectures such as MobileNet and hybrid architectures based on EfficientNet can provide an optimal tradeoff between accuracy and efficiency in terms of resources [38]. A significant reduction in the spread of potentially damaging deep fakes may result from social media filters and other real time applications at the platform level.

Regulation and Policy: Regulatory and policy mechanisms must also be put into place to ensure technical solutions do not fail as a result of their inability to independently verify truthfulness. As stated in [1][2] & [19] the development of international standards for the labeling of deepfakes, verification of authenticity and legal accountability is necessary through the establishment of regulations. It is likely collaboration between academia, government/policy makers and business/industry

stakeholders will need to occur to protect democratic systems, to ensure the protection of the public and to establish trust in digital media.

7. Implementation and Testing

The process for implementing deep fake detection systems as described in the 39 reviewed papers is systematic in nature. It starts with an initial data collection and preparation, then proceeds to develop and train the model; extract features; test and evaluate the models' performance; and lastly assess the robustness of the model when subjected to the constraints of the real world. Depending on the type of model being implemented (physiological cues; forensic signals; deep learning architectures; hybrid models; human centered perceptual analysis), each stage will be different. This section combines all of the procedures used in the literature to detect deep fakes, identifies the methods and practices that are commonly used, identifies innovative methodologies that have been developed to detect deep fakes and identifies the limitations that exist with regards to the experimental setups.

8.1 Dataset Preparation and Preprocessing

Beginning implementation involves the selection and preparation of the best possible dataset. The quality and amount of training data is a determining factor in how reliable the forgery detector will be. Most reported studies used the same standard benchmark datasets including DFDC, FaceForensics++ and Celeb-DF [9],[13],[20] as well as some studies used custom collections such as DF-Wild for "in-the-wild" manipulation [20] or IITP-IFFD for image level forgery detection [9]. As part of the pre-processing step, researchers have employed many different methods to detect faces (i.e., face detection) including MTCNN, Haar cascades, or Dlib-based landmark extractors to identify and isolate the region of interest for detecting facial forgery [13],[26]. For video models, researchers have extracted frames from videos at 15-30 frames per second to maintain correlation between each frame [16],[20] while for image-based model's researchers have resized input images to standardized sizes (i.e., 224x224 or 299x299) depending on the CNN backbone used [13],[15]. Commonly, researchers have employed augmentation techniques such as flipping horizontally, adding Gaussian noise, blurring, rotating, and applying color jittering to augment their datasets to create more diversity and to reduce overfitting [8],[17]. In addition, studies by GNG [33] demonstrated that models may be unintentionally biased by background artifacts and compression signatures and therefore strict pre-processing is important to avoid providing misleading cues to models.

8.2 Model Architecture Development and Feature Extraction

The choice of model architecture has been significantly impacted by how detection is performed in physiological based research; because physiological based research seeks to find the smallest amounts of biological inconsistency, for example, a slight deviation in blinking patterns [36] or head pose [30] or a mismatch between what your lips are doing and what you are saying [25]. Models that attempt to find these small discrepancies have made use of architectures that include a combination of spatial convolutional neural network (CNN) layers to process images and recurrent neural network (RNN) layers to process sequences over time to capture the pattern of an individual's blink rate [36]. To find misalignment between facial landmarks models have utilized geometric landmark vectors to determine if there is some unnatural alignment of facial components [30]. Similarly, models that look at lip-sync have included 3-D convolutional neural networks (CNNs) in addition to RNNs to find correlation between the movement of the mouth and the audio input [25]. Forensic based approaches have focused on finding lower-level pixel inconsistencies, for example, warping artifacts resulting from replacing a person's face [31], or residual noise and/or PRNU fingerprints indicative of tampering with an image [28]. PRNU based methods have incorporated traditional forensic techniques along with CNN classifiers to enhance the ability to detect smaller inconsistencies than traditional forensic techniques, while still maintaining interpretability. Deep learning models were the dominant architecture in the literature and have employed architectures including ResNet50, XceptionNet, MobileNetV2, EfficientNet and hybrid CNN-LSTM architectures [13][15][16][20]. A customized CNN architecture that employed facial landmark guidance resulted in an accuracy of 91.47% [13].

Temporal distortion detection was accomplished using optical flow-based models that calculated motion vector fields to capture distortion between frames [24][37]. These types of models provide robustness to previously unseen forms of manipulation. Fully convolutional networks and attention-based architectures have become increasingly popular in order to learn discriminative regions of interest from both spatial and temporal perspectives. There has also been a rise in hybrid systems that combine forensic cues, deep learning features or multimodal signals to enhance robustness. The Python FDT toolkit [9] has combined PRNU, ELA, CNNs and hybrid feature extraction into a single detection framework. Another hybrid work [8] combined InceptionResnet and Xception and achieved an F1-score of 0.98 through multi-scale feature learning. An ensemble system, DFN (MobileNet + XGboost), achieved 93.28% accuracy on the DFDC dataset [38] demonstrating the benefits of utilizing lightweight feature extractors with high performance classifiers.

8.3 Training and Optimization Pipeline

Training Configurations Differed Based Upon Model Complexity and Dataset Size. Deep learning-based models are generally run in batches ranging from 8 to 32 for video data and 32 to 64 for image data [13], [20]. Learning rates have been started at a wide range of values ($1e-3$ - $1e-5$) utilizing either cosine decay, step decay, or adaptive learning rate schedules using Adam or RMSProp optimization [15], [16], [23]. CNN-RNN hybrids require careful tuning of both the convolutional filters and sequential layer(s), and frequently incorporate a dropout value between .3-.5 to help prevent overfitting [20], [27]. GPU acceleration has been important, as most studies report that they were training their models on NVIDIA Tesla V100, RTX 2080 Ti, or A100 units; specifically for optical-flow or transformer-based architectures [20], [37]. While the CNN's can be trained on mid-range GPUs, the optical flow extraction and multi-stream fusion models need much higher computing resources than the CNNs.

8.4 Evaluation Metrics and Performance Reporting

Studies utilized quantifiable assessment methodologies to measure the reliability of detecting fraudulence in images. The accuracy metric has been the most commonly evaluated methodology among all studies reviewed [13], [15], [23], [38]. For datasets containing class imbalance or where the differences between forged images and authentic images are very slight, a variety of additional metrics were also employed including precision, recall, F1-score, and AUC (the Area Under ROC Curve) to provide more in-depth diagnostic information [8], [20]. Custom CNNs utilized a combination of these methodologies to produce results as follows: Accuracy = 91.47%, AUC = 0.92 [13]; while hybrid ensemble models utilizing these methodologies produced results as follows: Precision = 91.03% [38]. Optical flow CNNs produced F1 values ranging from 0.88–0.91 based upon the nature of the attack [24], [37]. Furthermore, the real-world dataset of DF-Wild demonstrated a significant loss of generalizability relative to the previously mentioned results, with F1 scores falling in the range of 0.77–0.81[20]. The Equal Error Rate (EER) was utilized in forensic-intense models that employed methodologies reliant on analyzing residual noise and PRNU [28]. Finally, confusion matrices, ROC curves, and Grad-CAM visualizations were employed to provide an interpretation of the model's performance, and to determine which specific image features contribute to the model's decision-making process [27].

Method / Model	Dataset(s)	Accuracy / F1	Runtime/Complexity
Physiological (Eye Blinking, Head Pose)	DFDC, FaceForensics++	Blinking-based ~85–90%	Lightweight, real-time capable
Forensic (Warping, PRNU/ELA)	FaceForensics++, DF-W, DeepfakeTIMIT	Up to 95% (image-level), ~80% (compressed video)	Moderate; sensitive to compression
CNN-RNN Hybrid (ResNet50+LSTM)	Celeb-DF, FaceForensics++	Acc $\approx$ 92%, F1 $\approx$ 0.90	High GPU demand, not real-time
Customized CNN (Landmark-based)	Celeb-DF, FaceForensics++	Acc = 91.47%, AUC = 0.92	Moderate GPU demand
Optical Flow CNN	DFDC, DF-W, FaceForensics++	Acc $\approx$ 88–91% (cross-forgery)	High (optical flow extraction costly)
DFN (MobNet + XGBoost)	DFDC	Acc = 93.28%, Precision = 91.03%	Moderate; lightweight + ensemble
Hybrid CNN (InceptionResNet + Xception)	FaceForensics++, Celeb-DF	F1 = 0.98, Precision = 0.985	High (ensemble CNNs, GPU required)
Toolkit (Python FDT + IITP-IFFD)	IITP-IFFD, DFDC	Acc $\approx$ 90–92% across tasks	Moderate; practical toolkit integration

Fig 4: Testing and Comparison Table

8. Conclusion

Deepfakes are difficult to detect because different techniques (CNNs, RNNs, optical flow, etc.) and combinations thereof can be used to create them. Each of the above-mentioned methods have shown high accuracy in detecting deepfakes in benchmark data sets. However, each of these has issues with real world generalization. Physiological indicators such as head postures and eye blinks provide some promise of detecting deepfakes. Additionally, forensic artifact detection such as warp effects, noise residue, and hybrid detection methodologies show potential. However, the next step in the game will be when an adversary develops new and advanced generative models which will make the task of detecting deepfakes even more difficult. Therefore, future research should include developing multi-modal detection capabilities, watermarks and provenance tracking, explainable AI, lightweight real-time systems, and social-political policy frameworks to support the development of trustworthy deep fake mitigation systems. The inter-connectedness of these components will both enhance the societal acceptance of deep fake mitigation systems, as well as their technical robustness. This review has integrated the results of 39 contributions, and thereby provides a roadmap for improving both the technical and socio-technical aspects of mitigating deep fakes.

References

1. Sharma, Dilip Kumar, et al. "A survey of detection and mitigation for fake images on social media platforms." *Applied Sciences* 13.19 (2023): 10980.
2. Raj, M. Thanga, et al. "A review on the detection of deep fake and propaganda videos and images-based voice and facial manipulation using AI techniques." 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS). IEEE, 2023.
3. Purohit, Ruchira, et al. "AI vs. Human vision: A comparative analysis for distinguishing AI-generated and natural images." 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, 2024.
4. Tiwari, Ayush Kumar, et al. "AI-generated video forgery detection and authentication." 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE, 2024.
5. Chinta, Druthik Sai, et al. "Analyzing Image Classification on AI-Generated Art Vs Human Created Art Using Deep Learning Models." 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE, 2024.
6. Li, Xiaojun, et al. "A CNN-based misleading video detection model." *Scientific Reports* 12.1 (2022): 6092.
7. Tyagi, Shobhit, and Divakar Yadav. "A detailed analysis of image and video forgery detection techniques." *The Visual Computer* 39.3 (2023): 813-833.
8. Ikram, Sumaiya Thaseen, Shourya Chambial, and Dhruv Sood. "A performance enhancement of deepfake video detection through the use of a hybrid CNN Deep learning model." *International journal of electrical and computer engineering systems* 14.2 (2023): 169-178.
9. Raj, Surbhi, Jimson Mathew, and Arijit Mondal. "FDT: A python toolkit for fake image and video detection." *SoftwareX* 22 (2023): 101395.
10. Shahzad, Hina Fatima, et al. "A review of image processing techniques for deepfakes." *Sensors* 22.12 (2022): 4556.
11. Khder, Moaiad Ahmed, et al. "Artificial intelligence into multimedia deepfakes creation and detection." 2023 International conference on IT innovation and knowledge discovery (ITIKD). IEEE, 2023.
12. Vahdati, Danial Samadi, et al. "Beyond deepfake images: Detecting ai-generated videos." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
13. Gura, Dmitry, et al. "Customized convolutional neural network for accurate detection of deep fake images in video collections." *Computers, Materials & Continua* 79.2 (2024).
14. Mohammed, Anwar. "Deep Fake Detection and Mitigation: Securing Against AI-Generated Manipulation." *Journal of Computational Innovation* 4.1 (2024)..
15. Badale, Anuj, et al. "Deep fake detection using neural networks." 15th IEEE international conference on advanced video and signal based surveillance (AVSS). Vol. 2. 2018.
16. Mira, Fahad. "Deep learning technique for recognition of deep fake videos." 2023 IEEE IAS global conference on emerging technologies (globConET). IEEE, 2023.
17. Gong, Dafeng, et al. "Deepfake forensics, an ai-synthesized detection with deep convolutional generative adversarial networks." *International Journal* 9.3 (2020).
18. Suratkar, Shraddha, et al. "Deep-fake video detection approaches using convolutional–recurrent neural networks." *Journal of Control and Decision* 10.2 (2023): 198-214.
19. Kaur, Achhardeep, et al. "Deepfake video detection: challenges and opportunities." *Artificial Intelligence Review* 57.6 (2024): 159.
20. Pu, Jiameng, et al. "Deepfake videos in the wild: Analysis and detection." *Proceedings of the web conference 2021*. 2021.
21. Malik, Asad, et al. "DeepFake detection for human face images and videos: A survey." *Ieee Access* 10 (2022): 18757-18775.
22. Patel, Yogesh, et al. "Deepfake generation and detection: Case study and challenges." *Ieee Access* 11 (2023): 143296-143323.
23. BR, Shobha Rani, and Piyush Kumar Pareek. "Deepfake video detection system using deep neural networks." 2023 IEEE international conference on integrated circuits and communication systems (ICICACS). IEEE, 2023.
24. Amerini, Irene, et al. "Deepfake video detection through optical flow based cnn." *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019.

25. Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018.
26. Saravana Ram, R., et al. "Deep Fake Detection Using Computer Vision-Based Deep Neural Network with Pairwise Learning." *Intelligent Automation & Soft Computing* 35.2 (2023).
27. Rajeev, Faaizah Panapparambil, and Sujala Deepak Shetty. "Detecting Deepfakes and Artificial Intelligence-Generated Art." 2024 International Conference on Computational Intelligence and Network Systems (CINS). IEEE, 2024.
28. Misal, Thakre, et al. "Detection of AI-Generated Images." *International Journal of Trend in Scientific Research and Development* 8.5 (2024): 805-810.
29. Ghiurău, David, and Daniela Elena Popescu. "Distinguishing reality from AI: approaches for detecting synthetic content." *Computers* 14.1 (2024): 1.
30. Yang, Xin, Yuezun Li, and Siwei Lyu. "Exposing deep fakes using inconsistent head poses." *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019.
31. Li, Yuezun, and Siwei Lyu. "Exposing deepfake videos by detecting face warping artifacts." *arXiv preprint arXiv:1811.00656* (2018).
32. Chauhan, Rakhi, Monika Sethi, and Sachin Ahuja. "Fake Faces Unveiled: A Comprehensive Study on Detecting Generated Facial Images." 2024 International Conference on Automation and Computation (AUTOCOM). IEEE, 2024.
33. Tanfoni, Marco, et al. "Generated or Not Generated (GNG): The Importance of Background in the Detection of Fake Images." *Electronics* 13.16 (2024): 3161.
34. Cooke, Di, et al. "As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli." *arXiv preprint arXiv:2403.16760* (2024).
35. Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking." *arXiv preprint arXiv:1806.02877* (2018).
36. Caldelli, Roberto, et al. "Optical Flow based CNN for detection of unlearned deepfake manipulations." *Pattern Recognition Letters* 146 (2021): 31-37.
37. Bansal, Nancy, et al. "Real-time advanced computational intelligence for deep fake video detection." *Applied Sciences* 13.5 (2023): 3095.
38. Westerlund, Mika. "The emergence of deepfake technology: A review." *Technology innovation management review* 9.11 (2019).
39. Korshunov, Pavel, and Sébastien Marcel. "Vulnerability assessment and detection of deepfake videos." 2019 International Conference on Biometrics (ICB). IEEE, 2019.