



MULTIMEDIA DEEPFAKE THREATS AND DETECTION STRATEGIES BASED ON LIMITED DATA: AN EXHAUSTIVE SURVEY

Nour Thani Mustafa¹, Omar Ibrahim Al-Saif^{*2}, Wasan Shakir Awad³

¹Computer Technique Engineering Department, Northern Technical University, Mosul-Iraq.

²Polytechnique College- Mosul, Northern Technical University, Mosul-Iraq.

³Ahlia University, Manama, Bahrain.

¹<https://orcid.org/0009-0002-9390-6743>, ²<https://orcid.org/0000-0003-2832-7868>, ³<https://orcid.org/0000-0001-7152-3480>

Email: nour.thani@ntu.edu.iq, Omar.alsaif@ntu.edu.iq, wawad@ahlia.edu.bh

ARTICLE INFO

Article History

Received: January 29, 2026

Reviewed: March 30, 2026

Accepted: April 30, 2026

Published: May 29, 2026

Keywords:

Deepfakes;
Multimedia Security;
Generative Adversarial Networks;
Autoencoders;
Deepfake Detection;
Cybersecurity;
Misinformation;
Identity Theft;
Blockchain Authentication;
Multimodal Analysis.

ABSTRACT

The high-speed development of multimedia deepfake technology has fundamentally altered the digital creativity and, at the same time, increased cyber and social threats. The current state of deepfake generation is informed by recent artificial intelligence methods, such as Generative Adversarial Networks (GANs), which learn to create realistically visual and audio content via adversarial learning; autoencoder-based models, which allow manipulating identity and expression compared to images; and diffusion models, which are trained to produce high-fidelity media by refining noise to data. Even though these practices aid in innovative uses in entertainment, availability, and customized media, they also help to perpetrate malpractices like political and social manipulation, privacy invasion, identity theft, misinformation, and cybersecurity assaults. Due to the recent popularity of deepfakes, there has been rapid growth in the design of a wide range of media detection models for video and audio-based deepfakes. In this study, literature on state-of-the-art models for deep-fake detection has been compiled and summarized, which includes CNN-based Spatio-Temporal models, CNN and LSTM hybrid models, Capsule Networks, and Sharp Multiple Instance Learning models (S-MIL), which were tested on popular deep-fake databases like FaceForensics++, Face2Face, DFDC, and ASV-Spoof. Results from the study indicate a high level of efficacy for the models in crossDataset environments, with accuracy between 84.6% and 100%, precision up to 98%, and AUC reaching 1.0. Limits that still exist for deep-fake detection models relate to high computational complexity, sensitivity to attacks, and the lack of generalizability for real-time processing.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Manipulation of digital video especially facial manipulation has been the issue of concern in recent years. To a large extent, this attention increase can be attributed to the presence of deepfakes, as such models are based on deep learning and depend on such models to modify or create visual images. Deepfake methods normally utilize the idea of auto-encoders or the generative adversarial networks (GANs) to apply the facial attributes of a source person to a target video. These models can produce highly realistic manipulated videos because they can work with large training dataset [1]. After the emergence of a 2017 Reddit post where a Reddit user uploaded an anonymous video in which he edited explicit videos with deepfake software, the academic volatility of deepfake detection increased exponentially. Several methods of detection have been proposed thereafter. There are the methods, which apply recurrent neural networks to detecting temporal inconsistencies between video frames and some, which apply convolutional neural networks to detect visual artifacts in manipulated faces [2]. The ease of use of deepfake generation software like Deep-Face-Lab, Fake-App, and Faceswap has led to the increased applications of facial manipulation technology with many being developed based on GAN or auto-encoder-decoder architecture [3].

GAN architecture is composed of two key elements, a generator network which creates fake content under the influence of random inputs with the help of a training dataset, and a discriminator network which tries to distinguish between authentic pictures and right-created ones. The two types of models are trained together in an adversarial model that continually enhances the performance of both models [4]. By contrast, systems based on auto-encoders implement deepfakes by encoding latent features of facial images with the help of an encoder and decoding faces with the help of a decoder. Individuals with each pair of encoders-decoders trained on different people, but the parameters of the encoder shared, one can reconstruct the face of another person with the latent features of another individual, and swap the face between the source and target face [5]. The bulk and heterogeneous data on actual and fake videos are needed to create trustworthy deepfake detection anthropometric systems, and the problem is generally stated as a dichotomous classification problem, identifying whether a video is authentic or not. A number of benchmark datasets have thus been provided. Table 1 indicates that there are 7 datasets that are popular in deepfake studies [6].

Table 1: The top seven datasets used in deep fakes.

Reference	Dataset Name	Subjects	No. of real videos	No. of fake videos
[7]	FaceForensics++ (FF++)	977	1000	5000
[8]	DeepFake-TIMIT	64	960	640
[9]	WildDeepfake	100	707	-
[10]	Deepfake Detection Challenge (DFDC)	66	5244	5244
[11]	Celeb-DeepFake (Celeb-DF)	72	6229	5639
[12]	DeeperForensics-1.0	100	60000	1000
[13]	Google/Jigsaw DeepFake Detection	3000	—	3000

Source: Authors, (2026).

Figure (1) provides a comparative analysis regarding the usage of deepfake datasets in terms of real and forged video numbers. This diagram not only provides information about the size of the available datasets but also about their balance.

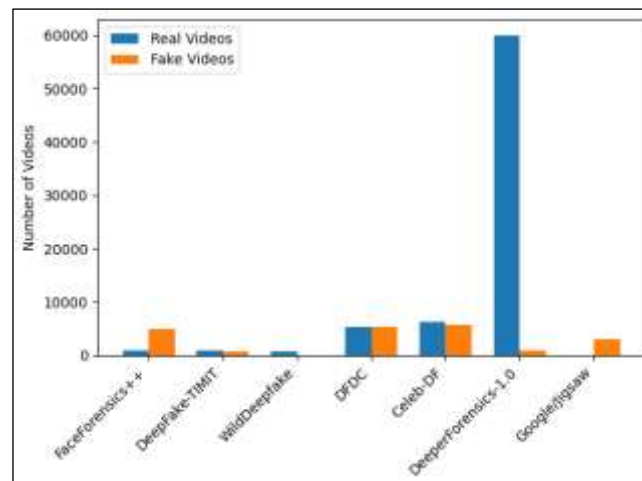


Figure 1: Comparison of Real and Fake Videos Across Deepfake Datasets.

Source: Authors, (2026).

The fact that the data acquisition and generation processes differ essentially in the widely used deepfake datasets explains the imbalance between real and fake videos. Human subjects are required to provide real videos under rigorous privacy, consent, and legal protocols, and they need to be carefully edited in order to provide diversity in identity, pose, lighting, and expression. By contrast, artificial videos can be produced in large quantities via algorithmic methods by engaging various manipulation methods to a small collection of original videos so that a single genuine sample yields tens or even hundreds of fake variants. This is why FaceForensics++ datasets have significantly more fake samples than real. On the other hand, datasets such as DFDC are designed to have balanced class distributions to allow fair benchmarking whereas others (e.g., WildDeepfake and Google/Jigsaw) are mostly built on actual or manipulated online content, which leads to the lack of or uneven counts of classes. Such discrepancies represent the original aim of the dataset that was to optimize the diversity of the manipulation, realism, or evaluation fairness, and have a direct impact on detector training, generalization, and bias.

Consequently, researchers need to focus on the issue of class imbalance when choosing datasets carefully and employ suitable methods, including re-weighting, resampling, or cross-dataset validation, to guarantee a high level of deepfake detection. Despite the significant development of detection technologies, many challenges still exist. With the advancement in the methods of Deepfake generation, videos produced are becoming realistic, and hence it is challenging to train on old detection algorithms to detect other methods of manipulation [14]. Deepfakes development has quickly transformed into an important issue in cybersecurity, offering the possibility of making manipulated multimedia look quite realistic and capable of deceiving humans and automated mechanisms. They become increasingly available, hence increasingly dangerous, including identity fraud, political misinformation, financial fraud, impersonation attacks, and massive social engineering. While methods of generating deepfakes are rapidly developing compared to detection tools, organizations have to grapple with growing dilemmas of trust and veracity, as well as safeguarding of the digital content [15].

This paper presents a survey of the environment of the multimedia deepfake threat and countermeasures in place against it. It discourses on modern-day generation technologies, model detection, challenges presented by adversarial, and shortcomings of existing methods. The survey also summarizes the research gaps still remaining as barriers to deploying the results in the real world, such as a lack of sufficient data, problems of model generalization, and the never-ending arms race between the attackers and defenders. The key contributions of the work are that it presents: i) a systematic taxonomy of the different types of deepfakes and attack scenarios; ii) a review of state-of-the-art detection and prevention techniques that leverage both audio and video representations of deepfakes; iii) the evaluation metrics, datasets, and benchmarks; and iv) an open research gap and open challenges towards enabling resilient and scalable deepfake defense systems.

II. OVERVIEW OF MULTIMEDIA DEEPAKE TECHNOLOGIES

A broad assortment of AI-powered techniques multi-media deepfake technologies allow the generation, manipulation, and creation of artificial amounts of content in pictorial, audio, video, and multisensory formats. All these applications are based on deep learning models, such as autoencoders, GANs, diffusion models, and cross-modal transformers, all of which are expected to learn complex patterns using large datasets and reproduce them with impressive realism. To conceive such approaches, the concerned terms must be appropriately defined first. Deepfakes are typically images and videos which have been created or edited using deep neural networks. It should be stated that synthetic media can be used to mean completely artificial media, manipulated media forcefully refers to the editing of media, face-swapping suggests that one person has their face replaced by another person, and voice cloning can be used to define the imitation of speech features using small speech samples. This combination of definitions portrays a seamless conceptual framework [16]. The procedures of deepfakes generation differ in terms of the extent of sophistication and the quality of the product. Initial systems did not use any type of autoencoders, but expressed the reconstruction of faces in compressed but meaningful feature representations; by sharing a common encoder and training them to solve different identities, autoencoders could allow the provision of a face-swapping example early on. Sturdier methods were developed under the name Generative Adversarial Networks, which is a combination of a generator and a discriminator, competing with each other to generate more realistic work [17]. GANs offered significant advances to realism, which in turn allowed lip-syncing, facial reenactment and realistic portraits on a big scale.

Other approaches have been domain on the basis of face-swap and voice-swap operation under geometrical alignment, landmark tracking and a audio embedding model to transfer identify or vocal style between the target and the source. Recent-generation deepfakes utilize diffusion models and cross-modal learning to produce video based on text-based descriptions or create synchronized-audio-videos, greatly increasing the potential of creative tools and the potential harm possible [1]. The forms of deepfakes can be extremely varied: there are also images: still photographs that have been altered or generated to appear as actual individuals that may be the source of impersonation fraud or internet alters egos. Video deepfakes are methods that interfere with frame sequences to imitate facial expressions, behavior or full-body movement-all of which is able to imitate extremely believable impersonation in the real world. The tone, accent and rhythm of a speaker are reproduced in audio deepfakes in order to reproduce speech that is almost similar to real audio records of the victim and commit fraud in the financial, political and social environment. Multimodal deepfakes include audio, video, text and in some cases, physiological signals such as gesture or affect to build entirely synchronized synthetic personalities capable of interacting in real-time [6]. These include applications in filmmaking, advertising, accessibility improvement to the pioneer applications in personalized digital assistants. Moreover, the risks are severe: non-consensual creation of content, political misinformation, identified theft, impersonation attack, sabotage in the reputation, loss of trust among people. The two-purpose of deepfake technologies predisposes that, to design detection systems, regulation protocols, and authentication protocols, it is always important to comprehend the technical basis of such technologies, as well as its application and misuse scenario [18].

III. CYCLE OF DEEPAKE ANALYSIS AND COUNTERMEASURES

The increased sophistication of the deepfake generation-related strategies has proven that a systematic plan is needed that would guide the process of collecting, analyzing, and refining the distorted multimedia content that would be examined during a forensic analysis. Deepfakes can alter facial expressions, flags, and even entire scenes and pass as extremely genuine, and current technologies of conceiving are not sufficient. To address these challenges, scholars have proposed systematic analysis cycles that comprised data management, multimedia evaluation and improved feature mining. The flexibility and efficiency of the deepfake detection algorithm would be ensured by such a process since the manipulation algorithms will continue to evolve. Figure 2 shows the simplified version of this cycle and shows how the content is transferred between the raw data storage and refined fiefs extraction [19].

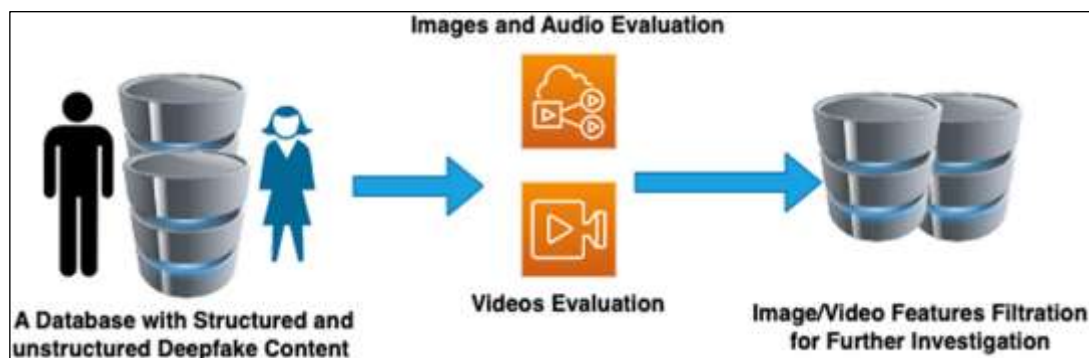


Figure 2: Cycle of Deepfake Analysis and Countermeasures.

Source: Authors, (2026).

The process starts with the establishment of a complete database that includes both structured and unstructured material based on deepfake. Structured content can consist of curated datasets ready to be used in training and benchmarking, including labels, metadata and standardized formats. Such datasets assist machine learning models to acquire credible trends related to manipulated content. Unstructured content, in turn, includes raw videos, images, and audio recordings that have been captured in uncontrolled settings, e.g. social media platforms or user submissions. Such kind of a material is necessary to mirror the state of the real world where compression artifacts, variable lighting, as well as unequal recording quality complicate identifying deepfakes. The mix of structured and unstructured sources provides the opportunity to have a strong detection in different situations because the analysis cycle should be robust [20]. After compiling the data, the second important stage is concerned with multimedia evaluation that implies examination of images, audio, and videos. Image based assessment is the type of assessment that studies spatial patterns which are difficult to recreate with generative models. They consist of anomalies in skin texture, shadows of incorrect proportions, blurred boundaries at the face, and minor irregularities between the image noise patterns. Audio testing, however, focuses on unnatural phonemic shift, abnormal breathing rhythms, disproportionate tone of emotion, or calculating errors normally found in synthetic voices [21]. Video analysis is a synthesis of spatial and temporal analysis in that the motion dynamics, transitions of facial expressions, frequency of blinking of the eyes, and accuracy of lip-sync and head movement patterns are analyzed. More complex deep learning models in the form of convolutional neural networks, recurrent neural networks, and attention-based models are used during this stage of evaluation to detect fine details indicating tampering [22].

After the assessment of the multimedia, the system moves to the feature filtration which is the refinement of extracted information which is displayed to show the best indicator of manipulation. This step is important as deepfake datasets are often redundant, with inconsistent quality, and with an excessive amount of noise. As a benefit, effective filtration lowers the computational complexity and at the same time maintains discriminative features that are required to obtain accurate classification. Examples of these features would be micro-expressions, which happen too fast or too homogenous, odd photometric characteristics, non-uniform physiological cues, including inconsistent patterns in the pulsation of skin-tone, and statistical anomalies that are detectable on the basis of machine learning models. The dimensionality reduction, feature normalization and anomaly scoring are the techniques that are often used to improve the quality and reliability of the filtered features [21]. The final refined features are then submitted to forensic models or human analysts so that they can be authenticated. Ratings on whether the content is authentic or altered may be made by an automated detection system relying on only the learned patterns, but borderline cases may contain additional contextual information reviewed by a human. Notably, the process is recursive: the results of the outcomes of the detection can be processed with the help of the feedback and change the database, modify feature filters, or refine evaluation methods. By doing this, the automatic loop of improvement guarantees that this detection system would be resistant against new implementations of deepfake generation, which are becoming lifelike and harder to perceive [23].

IV. SECURITY PROTOCOLS IN MULTIMEDIA ENABLED DEEPPKES

With the growing trend and proliferation of deepfakes enabled by multimedia, it has become an utmost requirement to ensure the security of the content. To address the concerns arising from deepfakes, different techniques are being opted for by researchers and experts in the domain. This includes everything from common cryptography techniques to artificial intelligence-driven techniques. Figure (3) showcases the most prominent security measures to combat the use of multimedia-assisted deepfakes. The central figure of the flow illustrates the concept of digital content security, whereas the other components represent the eight most prominent measures: robust, fragile digital watermarking, Blockchain Distributed Ledger Technology (BDLT), the detection of deepfakes employing deep learning and CNN models, audio-visual synchronization, hashing (SHA-256), validation of metadata, and multimodal analysis.



Figure 3: Security Protocols for Countering Multimedia-Enabled Deepfakes.
Source: Authors, (2026).

Fast propagation of deepfakes enriched with multimedia has incited serious problems to be solved by developing robust mechanisms for maintaining the integrity and authenticity of digital content. Contemporary security frameworks place a greater reliance on various combinations of multimodal analysis, blockchain-based distributed ledger systems, deep learning algorithms, and cryptographic watermarking for proposing countermeasures against manipulations related to deepfakes.

Digital watermarking has made critical contributions among these by embedding invisible and tamper-resistant identifiers in images, audio, and video files. Metadata-driven or cryptographically generated identifiers assure authenticity, traceability, and protection from violations of intellectual property rights [24]. Fragile watermarking is intended for very minute changes in sensitive media, while robust watermarks remain resistant against operations such as resizing, compression, or format conversion [25]. BDLT also enhances content authentication by storing cryptographic messages of original multimedia files on an immutable decentralized registry. Each change of the content generates a hash that is different compared to the others; hence, any illegitimate interference of the content will be detected immediately. This gives more transparency and trust that ultimately adds value to blockchain-enabled authentication protocols in most issues, including journalism to digital rights management to legal forensics. Detection approaches based on deep learning would provide another important defense level, that is, to analyze visual and audio information that can point to tampered materials. These are examples of deep learning architectures such as convolutional neural networks, audio-video synchronization networks that may be used to identify characteristic abnormalities of deepfakes, such as artificial facial expression and hypersynchronous lip motion or unmatched audio-video, to aid in tracking such abuse in real-time on both social media sites and live streaming platforms [26]. Added protection is through hash and metadata check mechanisms. The possibility to compare with the hashes of the newly uploaded content is offered by a cryptographic hash algorithm, e.g. SHA-256 that generates unique digital fingerprints of every multimedia content in the secured ledgers to detect unacceptable alterations in the same 19, 20.

The metadata verification is used in addition to this process to verify the authenticity of the media assets and the fact that they have not been manipulated and on top, inserting details, including, but not limited to, timestamps, geolocation tags, and device specifications. Combined all these multilayered security features comprise an integrated defense mechanism to address the expanding threats that are associated with deepfakes that can be facilitated through multimedia [27]. Table 2 mainly concentrates on the topic of protective and provenance-based security measures that are aimed to prevent, authenticate, or trace multimedia content in the magnitude of deepfake threats. Detection-based techniques though is also included but this is only deemed to work well when protective measures have failed or when content which has been manipulated has already entered the market. The technologies listed above, although BdlT is somewhat lower-priority than the other technologies, Blockchain Distributed Ledger Technology (BDLT) is considered immutable, decentralized, and verifiable content provenance, as a threat source, instead of possibly detecting artifacts in retrospect. Digital watermarking, in its turn, despite the historic significance, is slowly becoming defined as a legacy method because of the susceptibility to drastic changes and the latest generative algorithms; thus, only strong watermarking is preserved to ensure no redundancy. Techniques on detection like deep learning, audio-visual synchronization, and multimodal analysis are still maintained to reflect reactive security layers which are complementary protection mechanisms in a multilayered security architecture.

Table 2: Security Protocols for Countering Multimedia-Enabled Deepfakes.

Security Protocol / Technology	Description	Primary Purpose	Strengths	Limitations
Digital Watermarking (Robust) [28]	Embeds identifiers resistant to compression, resizing, and re-encoding.	Ensures media integrity during distribution.	Survives most transformations; ideal for copyright protection.	May slightly reduce quality if not optimized.
Digital Watermarking (Fragile) [29]	Embeds highly sensitive markers that break upon tampering.	Detects even minimal manipulation.	Excellent for authenticity verification of sensitive content.	Not suitable for content that undergoes compression.
Blockchain Distributed Ledger Technology (BDLT) [30]	Stores cryptographic hashes of original media on decentralized immutable ledgers.	Content authentication and anti-tampering validation.	Tamper-proof, transparent, decentralized, trustworthy.	High computational cost; scalability issues.
Deep Learning-Based Detection (CNNs) [31]	Uses convolutional neural networks to detect spatial artifacts in images/videos.	Automated detection of manipulated frames.	High accuracy; effective for visual defects.	Requires large labeled datasets.
Audio-Visual Synchronization Models [31]	Examines coherence between speech, lip movement, and facial expressions.	Detects mismatches typical of deepfake videos.	Strong at capturing temporal inconsistencies.	Performance drops with noisy or low-quality videos.
Hashing (SHA-256) [32]	Generates unique digital signatures for multimedia content.	Identifies unauthorized modifications.	Fast, secure, widely supported.	Ineffective alone against sophisticated deepfake reconstruction.
Metadata Verification [33]	Checks embedded data such as timestamps, source device, and location.	Supports authenticity and forensic validation.	Complements hashing; helps identify forgery.	Metadata can be spoofed if not cryptographically protected.
Multimodal Analysis [34]	Combines visual, audio, text, and behavioral cues.	Provides more accurate deepfake detection.	Reduces false positives; robust across content types.	Requires high computational resources.

Source: Authors, (2026).

V. THREAT LANDSCAPE OF MULTIMEDIA DEEPPAKES

Deepfakes are becoming more advanced, as accessible, multimedia, which use actors to create hyper-realistic synthetic content, which is used to influence perception, disrupt trust, and abide by vulnerabilities. Deepfakes can also be deployed in politics to affect an election, slander an opponent or even the trigger social turmoil by portraying events or claiming statements that do not occur. At the individual level, deepfakes are a threat to privacy by promoting identity theft, revenge porn, or unsanctioned impersonations. Deepfakes have the potential to form a great instrument within the misinformation, disinformation campaigns, which are actively spreading fake stories across social media and online platforms at an unprecedented pace [35]. Within the framework of cybersecurity, deepfakes allow complex fraudulent activities, phishing sites and the masquerading as known personalities. Such threats are also coupled with legal, ethical, and policy gaps as they cannot keep up with the high rate of synthetic media evolution, and the victim has no resources to complain to them, and enforcing and creating ethics are not an easy task [36]. Table (3) shows Threat Landscape of Multimedia Deepfakes.

Table 3: Threat Landscape of Multimedia Deepfakes.

Threat Category	Description	Methods / Examples	Potential Impact	Mitigation Challenges
Political and Social Manipulation [37]	Use of deepfakes to influence political outcomes, sway public opinion, or destabilize societies	Fake speeches of politicians, manipulated protests, synthetic news coverage	Erosion of trust in democratic institutions, social unrest, polarization	Detecting deepfakes in real-time, regulating content without infringing free speech
Privacy Violations and Identity Theft [38]	Unauthorized use of personal images, videos, or voices	Deepfake pornography, impersonation in calls or videos, social media misuse	Reputational damage, financial loss, emotional distress	Victims often unaware, difficulty in tracking perpetrators, lack of clear legal recourse
Misinformation and Disinformation Campaigns [39]	Spreading false information via realistic media	Fake celebrity endorsements, fabricated events, misleading news clips	Public confusion, manipulation of beliefs, harm to public health or safety	Identifying synthetic content among massive online data, countering viral spread
Cybersecurity and Fraud Threats [40]	Exploiting deepfakes for scams or unauthorized access	Voice cloning for banking fraud, AI-generated phishing videos, bypassing biometric authentication	Financial theft, corporate espionage, security breaches	Continuous advancement in deepfake technology, insufficient AI-based defense mechanisms
Legal, Ethical, and Policy Challenges [41]	Legal frameworks and ethical standards lag behind technological advances	Limited legislation on synthetic media, cross-border jurisdiction issues	Regulatory gaps, ethical dilemmas, difficulties in enforcement	Harmonizing laws internationally, balancing innovation with protection, establishing accountability

Source: Authors, (2026).

VI. COMPARATIVE ANALYSIS OF DEEPPAKE RESEARCHES

By [42] investigated deepfake (DF) detection based on faces2face video datasets and a mixture of DF-specific methods. They had the desire to exploit the distinct artifacts in videos manipulated by Face2Face, and the training models to identify the minor inconsistencies in face movements and pattern of expressions. The model had a very high detection accuracy of 95%-98 with frame-level feature extraction and analysis of Temporal sequences. The research revealed the utility of specificity of detection strategy to particular process of reenacting DF especially facial reenactment. Nevertheless, the approach had the disadvantage of overfitting to the features of Face2Face dataset used, which restricts the potential to use this approach to other deepfake generation mechanisms. Although the high accuracy is a result of the capability of the dataset-dependent detection models, this diversity of the training data is limited and this limits its applicability in real-world situations, where DFs may be produced by an even larger number of different algorithms that evolve over time. The paper has discussed the need to have variety of the datasets and cross dataset testing in future research studies in order to increase the strength of the research. In turn [43] proposed a deep neural network with hierarchical learning layers (DNN-HLL) for audio deepfake detection using the ASPspoo dataset. The model achieved an equal error rate (ERR) of 0.045, which indicates good performance in discriminating real from imitated audio.

The hierarchical structure allows the model to capture low-level and high-level acoustic patterns, which is essential in detecting slight speech manipulation. However, the study was limited by the scarce diversity of the dataset, which would diminish the capacity of the model to generalize to other types of audio data or unknown attacks. Further, real-world audio includes some environmental variation and background noise that was not well represented in ASPspoo. The research finds the potential of DNN-HLL for audio deepfake detection but demands larger, more diverse datasets and cross-dataset testing to attain robustness in practice. According to [44] examined the problem of deepfake detection by experimenting the GAN-based video generation videos on the FaceForensics++ data. They employed convolutional neural networks (CNNs) to acquire spatiotemporal feature of video frames to allow detecting subtle artifacts of manipulation that are otherwise inconspicuous to human participants. A striking precision of 98 was attained in the experiment, and this implicates the power of CNNs in identifying pixel-level inconsistency that is created during the process of synthesizing GAN-based. Also, pre-processing was done to standardize the input frames and further featured extraction was done to boost the overall detection performance. GAN-based detection models are very precise and can be attacked by adversarial examples, where minor changes can potentially fool the network. The authors proposed to enhance the robustness with the help of strong learning techniques and adversarial learning. The method is effective to find a few of the GAN-manipulated examples, however, generalization to unseen GAN models and new manipulation strategies is still to be addressed and as such an adaptive and ever-changing system of detection is required.

By [45] created a deep learning algorithm that was used to detect deepfake videos through the use of artifacts created as a result of face warping during the generation step. CNN-based technique detects minimal distortions that are caused during warping of synthetic faces and matching with original frames, which is generally undetectable by human eye. Data tests on the model have been done on various DF video data with accuracy of between 84.6 to 99.4, which varies with the origin of the video. The paper points out that since the quality of data sets differs depending on their generation processes, performance is data set specific, which poses a problem of high variance. The method indicates that CNNs can be successfully trained to ensure that they record artifacts due to deepfake synthesis. Nevertheless, it can be overcome by video employing the methods of new generation, or by post-processing, which is why there is a necessity of constant change in the detection algorithms. The article emphasizes the significance of strong feature evolution in deepfake detection of large sample sizes. In [46] suggested a system of deepfake detection and models of other kinds of attacks by use of capsule network that would be used to detect such attacks like computer-generated video and print-and-play attack using printed material or recorded media. The model exploits the specific ability of capsule network to study hierarchical spatial relationships, and detect discrepancy of facial structures, resolving the problem of inverse graphics. The approach was tested on the Capsule-Forensics dataset having an accuracy between 89.1 and 100%. Training also used random noise in the course of work. Although the quality of the videos that the model handled is close to excellent, low-resolution or highly compressed videos deteriorate the performance of the model.

The article illustrates the suitability of the capsule networks to generalized computer vision tasks and the prospects of generalized deepfake detection models. But scalability and real-time processing is still not yet achieved because of computational complexity suggesting further optimization to be used in real-time stream of videos or large data sizes. According to [47] addressed the issue of partially manipulated faces in deepfake videos using a Sharp Multiple Instance Learning (S-MIL) framework. Contrary to standard MIL, which induces instance embedding's to map indirectly into bag-level predictions, S-MIL induces a direct mapping and circumvents gradient vanishing as well as enhances training efficiency. The model treats face in a video as instances and the video as a bag and picks up intra-frame and inter-frame inconsistency. It was evaluated on celebDF, DFDC, and Face Forensics++ datasets and achieved AUCs of 0.9944 on celebDF, 1.0 on Face Forensics++, and 0.8511 on DFDC. While S-MIL works well at detecting half-altered faces and generalizes well between datasets, its higher complexity increases computational requirements. The approach can be further applied to single-frame DF image detection with a generic framework for various DF detection tasks. The research prioritizes instance-level modeling in an effort to detect localized manipulation that could escape classical video-level classifiers. By [48] proposed a hybrid CNN-LSTM model for deepfake video detection. CNNs were used to extract frame-level spatial information, which was subsequently passed through LSTM layers to capture temporal relations among sequences. The model was trained and evaluated on a number of datasets including Face Forensics, DFDC, and celebDF, where the overall accuracy was 92%.

The approach performs well in differentiating genuine and forged video content according to spatial and temporal discrepancies but requires immense computational ability, which might limit its application where there are not enough resources. The experiment demonstrates the effectiveness of a combination of convolutional and recurrent models in the detection of DF, which provides more accurate results than CNN-only simple models. Future work suggested enhancing computational cost-effectiveness and using the framework to multimodal data, i.e., incorporating audio cues, to further reinforce detection robustness. In [49] developed a 3D CNN model that can detect deepfakes by extracting spatiotemporal information from adjacent video frames. The network was trained using the FaceForensics and VidTIMIT datasets and correctly detected 94.5% to 99.4% of deepfakes. The 3D convolutional layers allow the model to analyze temporal dynamics together with spatial details, which improves the ability to detect subtle video manipulations. Even though highly effective, the approach is vulnerable to adversarial noise, which can lead to severe degradation of performance. The research demonstrates the benefit of temporal sequence modeling as well as frame-level analysis in robust detection of deepfakes. However, computational cost of 3D CNNs may limit real-time usage. The authors underscore the importance of ongoing model revision based on latest DF generation strategies and adversarial attacks, as well as the potential to use light-weight architectures for deployment.

Table 4: Related Works Comparison.

Ref	Modality	Dataset	Model / Approach	Performance	Strengths	Limitations
Bonettini et al. [42]	Video	Face2Face	DF-specific methods, frame-level feature extraction + temporal analysis	Accuracy: 95–98%	Captures subtle facial movement artifacts; high detection on Face2Face	Overfitting to dataset; limited generalization to other DF types
Frank et al. [43]	Audio	ASVspoof	DNN-Hierarchical Learning Layers (DNN-HLL)	ERR: 0.045	Captures low- and high-level acoustic patterns; strong discrimination of real vs fake	Dataset diversity limited; real-world noise not well represented
Salvi et al. [44]	Video	FaceForensics+ +	CNN for spatiotemporal features	Precision: 98%	Detects subtle pixel-level GAN artifacts; pre-processing boosts performance	Vulnerable to adversarial attacks; limited generalization to unseen GANs
Jung et al. [45]	Video	Multiple DF datasets	CNN-based face warping artifact detection	Accuracy: 84.6–99.4%	Detects minimal distortions; effective across multiple datasets	Dataset-specific performance; may fail on newer DF methods or post-processed videos

Guera et al. [46]	Video + attacks	Capsule-Forensics	Capsule Network	Accuracy: 89.1–100%	Captures hierarchical facial structures; generalizable to other attacks	Computationally heavy; low-res/compressed videos reduce performance; not real-time
Zhang et al. [47]	Video	celebDF, DFDC, FaceForensics+ +	Sharp Multiple Instance Learning (S-MIL)	AUC: 0.9944–1.0	Detects partially manipulated faces; good cross-dataset generalization	High computational complexity
Yuezun et al. [48]	Video	FaceForensics, DFDC, celebDF	CNN-LSTM hybrid	Accuracy: 92%	Combines spatial and temporal features; more accurate than CNN-only	High computational requirements; may limit real-time deployment
Nguyen et al. [49]	Video	FaceForensics, VidTIMIT	3D CNN	Accuracy: 94.5–99.4%	Extracts spatiotemporal info; robust detection of subtle manipulations	Vulnerable to adversarial noise; computationally intensive; real-time limited

Source: Authors, (2026).

Figure (4) highlights the major gaps and limitations identified in recent studies on the detection of deepfakes. It shows the level of occurrence of various limitations like performance on specific datasets, computational complexity, susceptibility to adversarial attacks, and the need for real-time processing. This graphical representation draws attention to the areas of major limitations in detecting deepfakes.

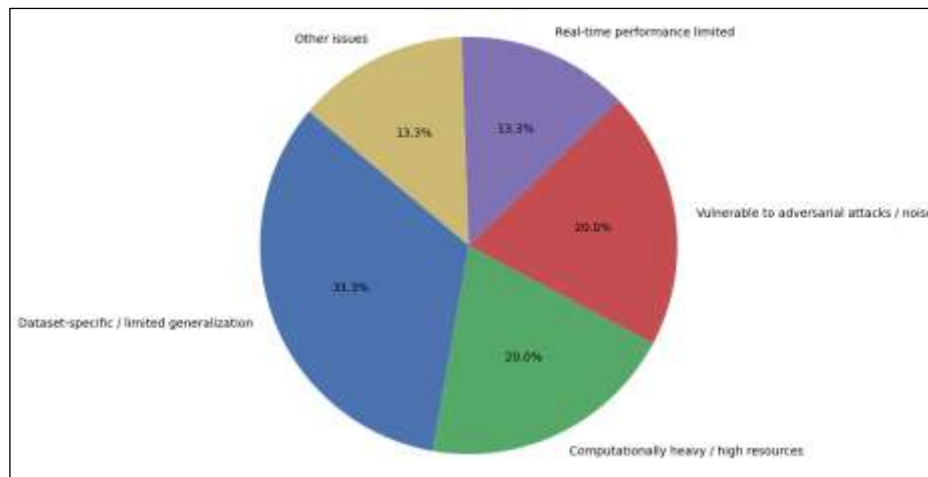


Figure 4: Main Gaps / Limitations in Deepfake Detection Studies.
Source: Authors, (2026).

VII. CONCLUSION

Multimedia deepfakes are not only innovative opportunities, but also a big challenge in terms of security. The biggest challenge is the dynamism of methods that are generated by generative AI such as GANs and cross-modal synthesis that tend to avoid conventional detection tools. One of the crucial problems is generalizability because traditional detectors cannot handle new subjects, different quality, and various environmental circumstances. Existence of only small, homogenous and low-quality data also compounds the effectiveness of current models. Novel multimodal deepfakes incorporate video/ audio/ text/ physiological signals, which further intensify these issues, especially when minuscule manipulation/ watermark removal methods are employed. Effective detection is complicated by high computational demands and other ethical and legal limitations, such as privacy, freedom of expression and regulatory lag. Nevertheless, efficient and generalizable detectors that can withstand unknown types of manipulation should be considered in research in the future. Multimodal cues comprising of visual, audio, text and behavioral features can enhance accuracy in detection. Transformer models and attention mechanisms can be used to capture inter-modal relation better. Minimized or lightweight architectures are necessary to facilitate large-scale and approximate real-time analyzes. Methods like adversarial learning, anomaly detection and content authentication by blockchain can be used to enhance robustness and traceability. Adaptive learning to changing threats will be facilitated by the creation of enormous and heterogeneous datasets. The main drawbacks of CNNs, RNNs, multimodal analysis, and cryptographic verification are still adversarial evasion, cross-domain manipulations, and limitations of datasets. These issues should be solved through sustained innovation, the work of cross-disciplinary teams, and good governance. When addressing privacy, protecting digital trust, and facilitating the responsible development and deployment of synthetic media technologies, it is important to integrate technical, ethical, and legal solutions.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Methodology: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Investigation: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Discussion of results: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Writing – Original Draft: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Writing – Review and Editing: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Resources: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad m.

Supervision: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

Approval of the final text: Nour Thani Mustafa, Omar Ibrahim Al-Saif, Wasan Shakir Awad.

IX. REFERENCES

- [1] M. D. and T. N. Nguyen, "Digital Face Manipulation Creation and Detection : A Systematic Review," *Electron. Artic.*, vol. 12, no. 3407, pp. 1–37, 2023.
- [2] M. R. Oraibi and A. M. Radhi, "Enhancement Digital Forensic Approach for Inter-Frame Video Forgery Detection Using a Deep Learning Technique," *Iraqi J. Sci.*, vol. 63, no. 6, pp. 2686–2701, 2022.
- [3] Y. L. and S. L. Xin Yang, "EXPOSING DEEP FAKES USING INCONSISTENT HEAD POSES Xin," *arXiv:1811.00661v2*.
- [4] A. Nabila, E. Wulandari, A. Ma, and W. A. Salah, "Understanding Generative Adversarial Networks (GANs): A Review," *Control Syst. Optim. Lett.*, vol. 3, no. 1, 2025.
- [5] E. Ward, Z. Wang, and Q. She, "Generative Adversarial Networks in Computer Vision : A Survey and Taxonomy," *arXiv:1906.01529v4*, no. 1, pp. 1–41, 2020.
- [6] M. A. Abdulhamed and A. N. Hashim, "A Survey on Detecting Deep Fakes Using Advanced AI-Based Approaches," *Iraqi J. Sci.*, vol. 65, no. 9, pp. 5254–5269, 2024.
- [7] R. Andreas, D. Cozzolino, L. Verdoliva, J. Thies, M. Nießner, and C. Riess, "FaceForensics ++: Learning to Detect Manipulated Facial Images," *arXiv:1901.08971v3*, 2019.
- [8] P. K. and S. Marcel, "DeepFakes : a New Threat to Face Recognition ? Assessment and Detection," *arXiv:1812.08685v1*, pp. 1–5, 2018.
- [9] B. Zi, M. Chang, J. Chen, and Y. Jiang, "WildDeepfake : A Challenging Real-World Dataset for Deepfake Detection," *arXiv:2101.01456v2*, 2020.
- [10] Mohammed, Naqaa Luqmana;Aziz, Mothanna Sh. Alsaif, Omar Ibrahim. , "Design and implementation of robot control system for multistory buildings", *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(5), pp. 2682–2689, 2020
- [11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," *arXiv:1909.12962v4*, 2020.
- [12] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," *arXiv:2001.03024v2*, 2020.
- [13] L. Guarnera et al., "The Face Deepfake Detection Challenge," *J. Imaging Artic.*, vol. 8, no. 263, pp. 1–21, 2022.
- [14] J. Hao, "Advances in Deepfake Generation and Detection Technologies : Challenges and Opportunities," *Trans. Comput. Sci. Intell. Syst.*, vol. 6, 2024.
- [15] S. Singh and A. Dhumane, "MethodsX Unmasking digital deceptions : An integrative review of deepfake detection , multimedia forensics , and cybersecurity challenges," *MethodsX*, vol. 15, p. 103632, 2025.
- [16] K. J. Patel and M. B. Desai, "Ai-Driven Advances and Challenges in Deepfake Technology : A Comprehensive Review," *J. Electr. Syst.*, pp. 1388–1400, 2024.
- [17] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection : A Systematic Analysis," *J. Sens. Actuator Networks*, pp. 1–38, 2025.
- [18] R. R. Kumar et al., "Report on the Follow-Up To the Regional Implementation Strategy of the Madrid International Plan of Action on Ageing in Lithuania," *Front. Neurosci.*, vol. 14, no. 1, pp. 1–13, 2021.
- [19] A. Ayub, K. Asif, A. Laghari, S. Azeem, I. Sajid, and U. Muhammad, "A survey on multimedia - enabled deepfake detection : state - of - the - art tools and techniques , emerging trends , current challenges & limitations , and future directions," *Discov. Comput.*, 2025.
- [20] V. Livernoche et al., "O PEN F AKE : An Open Dataset and Platform Toward Real-World Deepfake Detection," *arXiv:2509.09495v2*, pp. 1–26, 2025.
- [21] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics : a survey of digital forensic methods for multimodal deepfake identification on social media," *PeerJ Comput. Sci.*, pp. 1–40, 2024.
- [22] M. Porta-lorenzo, M. Vázquez-enríquez, A. Pérez-pérez, J. L. Alba-castro, and L. Docío-fernández, "Facial Motion Analysis beyond Emotional Expressions," *sensors Artic.*, vol. 22, no. 3839, 2022.
- [23] Thanoon, K.H., Hasan, S.Q., Alsaif, O.I., "Biometric information based on distribution of arabic letters according to their outlet",. *International Journal of Computing and Digital Systems*, 90(5), pp. 981–991. 2020.

- [24] M. Alkhatib, "A Multifaceted Deepfake Prevention Framework Integrating Blockchain , Post-Quantum Cryptography , Hybrid Watermarking , Human Oversight , and Policy Governance," Computers, pp. 1–28, 2025.
- [25] Y. A. Hassan and A. M. S. Rahma, "A Comparative Study Between the Fragile and Robust Watermarking Techniques and Proposing New Fragile Watermarking with Embedded Value Technique," Iraqi J. Comput. Commun. Control Syst. Eng., vol. 23, no. 3, pp. 186–196, 2023.
- [26] J. I. Senarathna and G. Scholar, "The Role of Cryptography in Blockchain : Ensuring Immutability , Transparency and Security The Role of Cryptography in Blockchain : Ensuring Immutability , Transparency and Security," pp. 0–14, 2025.
- [27] R. K. Salih and A. H. Kashmar, "Enhancing Blockchain Security by Developing the SHA256 Algorithm SHA256," Iraqi J. Sci., vol. 65, no. 10, pp. 5678–5693, 2024.
- [28] F. Han, P. Yang, H. Du, and X.-Y. Li, "Accelerometer-Based Anti-Spoofing Voice Authentication on Wrist-Worn Wearables," IEEE Trans. Mob. Comput., vol. 23, no. 5, pp. 5571–5588, 2024.
- [29] C. Zuo, X. Zhang, L. Yan, and Z. Zhang, "GUGEN: Global User Graph Enhanced Network for Next POI Recommendation," IEEE Trans. Mob. Comput., vol. 23, no. 12, pp. 14975–14986, 2024.
- [30] K. Yang, "How to prevent deception: A study of digital deception in 'visual poverty' livestream," New Media & Soc., vol. 27, no. 12, pp. 6779–6797, 2025.
- [31] C. Li, A. He, G. Liu, Y. Wen, A. T. Chronopoulos, and A. Giannakos, "RFL-APIA: A Comprehensive Framework for Mitigating Poisoning Attacks and Promoting Model Aggregation in IIoT Federated Learning," IEEE Trans. Ind. Informatics, vol. 20, no. 11, pp. 12935–12944, 2024.
- [32] A. Heidari, N. Jafari, H. Dag, S. Talebi, and M. Unal, "A Novel Blockchain - Based Deepfake Detection Method Using Federated and Deep Learning Models," Cognit. Comput., pp. 1073–1091, 2024.
- [33] I. Haq, K. Malik, and K. Muhammad, "Multimodal Neurosymbolic Approach for Explainable Deepfake Detection," ACM Trans. Multimed. Comput. Commun. Appl., vol. 20, Sep. 2023.
- [34] S. Ullah, X. Qiao, and A. Tariq, "Impact assessment of planned and unplanned urbanization on land surface temperature in Afghanistan using machine learning algorithms : a path toward sustainability," Sci. Rep., pp. 1–18, 2025.
- [35] A. S. H. George, A. Shahul, and A. S. George, "Wearable Sensors : A New Way to Track Health and Wellness Partners Universal International Innovation Journal (PUIIJ) Wearable Sensors : A New Way to Track Health and Wellness," Partners Univers. Int. Innov. J., no. August, pp. 15–34, 2023.
- [36] A. S. George and A. S. H. George, "Partners Universal Innovative Research Publication (PUIRP) Deepfakes : The Evolution of Hyper realistic Media Manipulation Partners Universal Innovative Research Publication (PUIRP)," Partners Univers. Innov. Res. Publ., vol. 1, no. 2, pp. 58–74, 2023.
- [37] P. L. Kharvi, "Understanding the Impact of AI-Generated Deepfakes on Public Opinion , Political Discourse , and Personal Security in Social Media," Futur. issue this J., 2024.
- [38] V. Lotus Company, Hanoi, "Addressing artificial intelligence misuse to perpetrate cybercrime in Indonesia : Deepfake , non-consensual intimate image of women , and child sexual abuse material :," United Nations Build. Rajdamnern Nok Ave., 2025.
- [39] B. Eva and S. Stepney, "Disinformation : sources , spread and impact," no. April, 2024.
- [40] Monetar, "Cyber Risks Associated with Deepfakes," no. September, 2025.
- [41] M. A. Ali, W. Ali, and M. Ali, "Harmonizing Technological Advancement with Ethical and Safe AI Deployment by Bridging the Innovation – Regulation Divide Harmonizing Technological Advancement with Ethical and Safe AI Deployment by Bridging the Innovation – Regulation Divide," vol. 1, no. 1, 2025.
- [42] N. Bonettini, L. Bondi, E. D. Cannas, P. Bestagini, S. Mandelli, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," Proc. - Int. Conf. Pattern Recognit., pp. 5012–5019, 2020.
- [43] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," Adv. Neural Inf. Process. Syst., no. NeurIPS, 2021.
- [44] D. Salvi et al., "A Robust Approach to Multimodal Deepfake Detection," J. Imaging, vol. 9, no. 6, pp. 1–18, 2023.
- [45] Alhashmi, S.Q., Thanoon, K.H., Alsaif, O.I. "A Proposed Face Recognition based on Hybrid Algorithm for Features Extraction"Proceedings of the 6th International Engineering Conference Sustainable Technology and Development Iec pp. 232–236, 9122911.2020.
- [46] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill., vol. 9, no. 2, pp. 22–26, 2018.
- [47] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting Deepfake Videos with Temporal Dropout 3DCNN," IJCAI Int. Jt. Conf. Artif. Intell., pp. 1288–1294, 2021.
- [48] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," 2019.
- [49] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 2307–2311, 2019.