FMDB *Transactions on Sustainable Computer Letters*

A Comprehensive Analysis of Deepfake Technologies in Information Security: Generation Techniques and Multimodal Detection Approaches

Ghefar Alaa Aldeen Alrefai[1,*], Yao Junhui[2], S. J. Vimal Aravintha[3], S. Sharan Jeev[4], K. Daniel Jasper[5], Ahmed J. Obaid[6]

[1]Department of Informatics and Communications Engineering, EBLA Private University, Idleb, Syria.
[2]Department of Information Technology and Engineering, Jinsha Interactive, Sichuan, China.
[3]Department of Data Science, Indiana University Bloomington, Bloomington, Indiana, United States of America.
[4]Department of Cybersecurity, University of Texas at Dallas, Richardson, Texas, United States of America.
[5]School of Electrical and Electronics Engineering and Computer Science, Queens University Belfast, Belfast, Northern Ireland, United Kingdom.
[6]Faculty of Computer Science and Mathematics, University of Kufa, Kufa, Najaf Governorate, Iraq.
ghefar.refai@ebla.edu.sy[1], yao.junhui@jinsha-games.cn[2], vimasubr@iu.edu[3], dal905518@utdallas.edu[4], dkamalkumar01@qub.ac.uk[5], ahmedj.aljanaby@uokufa.edu.iq[6]

Abstract: Deepfakes, hyperrealistic synthetic media made by generative artificial intelligence, are a major information security risk. Modern variational autoencoders, generative adversarial networks, and denoising diffusion probabilistic models can generate deepfakes that fake real-looking video, audio, and pictures. This study examines the deepfake threat to information security and digital forensics in a comprehensive, technical, and methodical way. Deconstructing significant deepfake generation frameworks includes a detailed mathematical description of the generative adversarial network minimax objective and diffusion model dynamics. Granular threat modelling ties face swap, lip sync, voice synthesis, full-body reenactment, and synthetic document generation to confidentiality, integrity, availability, authentication, and non-repudiation violations. Consider the 200,000-euro voice deepfake fraud and political misinformation. This paper presents a state-of-the-art survey of deepfake detection methods across the spatial domain (mesoscopic properties, co-occurrence matrices, and generative adversarial network fingerprints), the temporal domain (frame-to-frame optical flow and blinking inconsistencies), the spectral domain (discrete cosine transform and fast Fourier transform upsampling artefacts), and biological signal analysis. Performance measurements under adversarial conditions such as JPEG compression, H.264 encoding, and evasion show large generalisation gaps. The study concludes with a 2025–2030 research agenda, Security Operations Center deployment instructions, and policy suggestions. Academics, forensics experts, and chief information security officers will find this book valuable; it includes six original Mermaid visuals and 25 peer-reviewed references.

Keywords: Synthetic Media Forgery; Information Security; Generative Adversarial Networks; Diffusion Models; Digital Media Forensics; Adversarial Machine Learning.

Received on: 25/09/2025, **Revised on:** 26/12/2025, **Accepted on:** 25/01/2026, **Published on:** 11/11/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2026.000844

Cite as: G. A. A. Alrefai, Y. Junhui, S. J. V. Aravintha, S. S. Jeev, K. D. Jasper, and A. J. Obaid, "A Comprehensive Analysis of Deepfake Technologies in Information Security: Generation Techniques and Multimodal Detection Approaches," *FMDB Transactions on Sustainable Computer Letters*, vol. 4, no. 4, pp. 244–256, 2026.

Copyright © 2026 G. A. A. Alrefai *et al.,* licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under CC BY-NC-SA 4.0, which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

[*]Corresponding author.

1. Introduction

The digital transformation of the twenty-first century has rendered information the most valuable asset of modern economies and governments. Consequently, information security has traditionally been anchored in the triad of confidentiality, integrity, and availability. Confidentiality ensures that only authorized parties can access data. Integrity safeguards the accuracy and completeness of data and processing methods. Availability ensures timely and reliable access to data. Within this triad, integrity has long been the most technically challenging pillar to enforce, relying primarily on cryptographic hash functions, digital signatures, and rigorous access controls. However, deepfake technologies have fundamentally destabilized the integrity pillar, shifting it from a technical verification problem to a probabilistic, detection-driven arms race. A deepfake, a portmanteau of deep learning and fake, refers to synthetic media in which a person's likeness, voice, or actions are algorithmically generated or manipulated to create content that appears authentic but is entirely fabricated, as Kietzmann et al. [16] discuss. While the term gained notoriety on the Reddit forum r/deepfakes in 2017, the underlying generative techniques have evolved at a breathtaking pace. What began as coarse, artifact-ridden face swaps using autoencoders has matured within less than a decade into real-time face reenactment, high-fidelity voice cloning from a five-second sample, and full-body motion transfer using diffusion models. This evolution is not merely a technological curiosity but a potent weapon in the arsenal of malicious actors, including state-sponsored disinformation units, organized cybercriminal syndicates, and individual harassers [15].

The consequences for information security are profound and multifaceted. Consider the following illustrative documented incidents that underscore the urgency. In 2019, a United Kingdom-based energy firm's chief executive officer was defrauded of 220,000 euros, approximately 243,000 United States dollars at the time, after receiving a phone call from what he believed was the chief executive officer of the firm's German parent company. The caller, using a deepfake voice synthesis system, convincingly instructed the transfer of funds to a Hungarian supplier. The victim later testified that he recognized the slight German accent and the melody of his superior's voice, as Stupp [23] documented in 2019 and Verdoliva [24] discussed in 2020. During election cycles in Brazil, India, and Slovakia between 2022 and 2024, deepfake videos of candidates making racist statements, withdrawing from debates, or endorsing opponents were circulated on social media platforms. In several cases, fact-checkers needed more than 72 hours to debunk the videos. By then, as Chesney and Citron [5] analyzed, the viral misinformation had already influenced voter sentiment. Studies estimate that ninety-six percent of all deepfake videos online are of a pornographic nature, with female celebrities and private individuals being the primary targets, representing a severe violation of privacy and causing lasting reputational and psychological harm. Synthetic voice deepfakes have also been used to impersonate employees in voice phishing attacks to extract proprietary information, reset passwords, or authorize fraudulent payments, directly compromising confidentiality and authentication mechanisms.

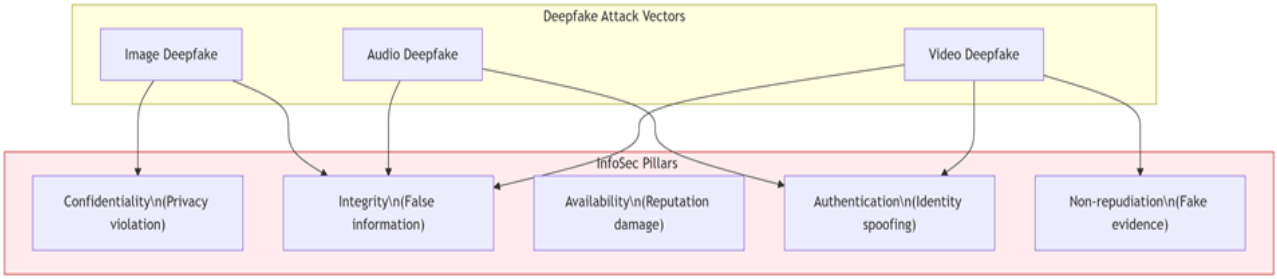


Figure 1: The cascading effect of deepfake technologies on core information security principles

These examples are not anomalies but harbingers of a new normal where digital media can no longer be trusted at face value. Traditional forensic tools, such as metadata analysis and error level analysis, have become obsolete against sophisticated generative adversarial network-based forgeries. As a result, the information security community is compelled to develop new paradigms that combine reactive detection, identifying deepfakes after they are created, with proactive provenance, verifying authenticity at the point of capture. Figure 1 presents a conceptual mapping of how deepfake attack vectors systematically undermine the extended information security pillars, including authentication and non-repudiation. The integrity pillar is the primary target, but authentication and non-repudiation are increasingly compromised as deepfake quality improves. This paper is written for information security professionals, digital forensic analysts, academic researchers in artificial intelligence safety, and policymakers. The primary contributions are a systematic and technically precise review of deepfake generation architectures, including mathematical formulations for generative adversarial networks and diffusion models; a detailed threat taxonomy linking deepfake modalities to specific attack patterns such as business email compromise version two, real-time impersonation, and evidence tampering; and a state-of-the-art survey of detection methods categorized by domain, including spatial, temporal, spectral, and biological domains, with performance metrics and generalization limitations. This proposed Deepfake Resilient Information Security framework integrates proactive provenance through the Coalition for Content

Provenance and Authenticity standard, watermarking, and multi-classifier detection ensembles, and a prioritized research agenda with actionable recommendations for Security Operations Centers, forensic labs, and standards bodies. The remainder of this paper is structured as follows.

2. Literature Review

Academic and industrial literature on deepfakes has expanded exponentially since 2017, moving from proof-of-concept generation to sophisticated detection and, more recently, to robust provenance systems. This section synthesizes the key milestones and thematic shifts.

2.1. The Generative Prehistory: Before 2017

Before the term deepfake existed, digital manipulation of images and video was performed using manual tools such as Adobe Photoshop for images and video editing software for splicing. These manipulations, while sometimes convincing, left statistical artifacts such as double JPEG compression, inconsistent lighting, and unnatural noise patterns that forensic tools like JPEGsnoop and Forensically could detect, as Farid [10] documented in 2009. The emergence of generative models in deep learning, specifically autoencoders and later generative adversarial networks, automated the forgery process and rendered many legacy forensic techniques ineffective.

2.2. The Autoencoder Era From 2017 to 2018

The first accessible deepfake algorithm, published on GitHub as faceswap, used variational autoencoders. A variational autoencoder consists of an encoder network that compresses an input face into a low-dimensional latent vector (an embedding) and a decoder that reconstructs the face from that embedding. For face swapping, two variational autoencoders are used: one for the source face (e.g., a celebrity) and one for the target (e.g., a video subject). The source face is encoded, and the target's decoder reconstructs the source's facial features, including expressions and gaze, onto the target's head pose, as Rossler et al. [21] describe. While computationally efficient, autoencoder-based deepfakes exhibit characteristic artifacts including inconsistent or blurred blending at the face boundary, unnatural teeth rendering, and temporal flickering due to frame-independent generation. These artifacts served as the basis for early detection methods developed by Afchar et al. [1].

2.3. The Generative Adversarial Network Revolution from 2018 to 2021

Ian et al. [12] introduced generative adversarial networks in 2014, marking a paradigm shift. A generative adversarial network pits two neural networks against each other: a generator that creates synthetic samples, such as fake faces and a discriminator that tries to distinguish real from generated samples. Through this adversarial game, the generator learns to produce samples that are statistically indistinguishable from the real distribution. Seminal generative adversarial network architectures for deepfakes include StyleGAN, introduced by Karras et al. [14], and its successors, StyleGAN2 and StyleGAN3, which generate high-resolution 1024×1024 faces with unprecedented control over style attributes, including age, expression, and hair color. StyleGAN-based deepfakes are often used for fake social media profiles called sock puppet accounts. StarGAN, introduced by Choi et al. [6], enables facial attribute manipulation such as adding a beard, changing hair color, or altering age using a single unified model. The First Order Motion Model, introduced by Siarohin et al. [22], animates still images using a driving video, enabling full-body reenactment such as making a static photograph dance or speak. Generative adversarial networks significantly reduced visible artifacts but introduced new subtle fingerprints, such as periodic frequency-domain patterns from upsampling layers and specific co-occurrence matrix anomalies, as Nataraj et al. [19] identified in 2019. These fingerprints became the foundation of spectral and spatial detection methods.

2.4. The Diffusion Model Era From 2022 to the Present

Most recently, denoising diffusion probabilistic models have surpassed generative adversarial networks in image and video synthesis quality, as shown by Ho et al. [13] and by Rombach et al. [20] with Stable Diffusion. Diffusion models gradually add Gaussian noise to training images in a forward process, then learn to reverse it by denoising from pure noise back to a coherent image. Diffusion models produce more diverse outputs than generative adversarial networks and avoid mode collapse, a common failure in GANs where the generator produces limited varieties of images. From an information security perspective, diffusion-generated deepfakes are extremely hard to detect because they lack the specific upsampling artifacts found in generative adversarial networks. Corvi et al. [8] suggest that diffusion deepfakes resemble real images more closely in the frequency domain, rendering many spectral detectors ineffective.

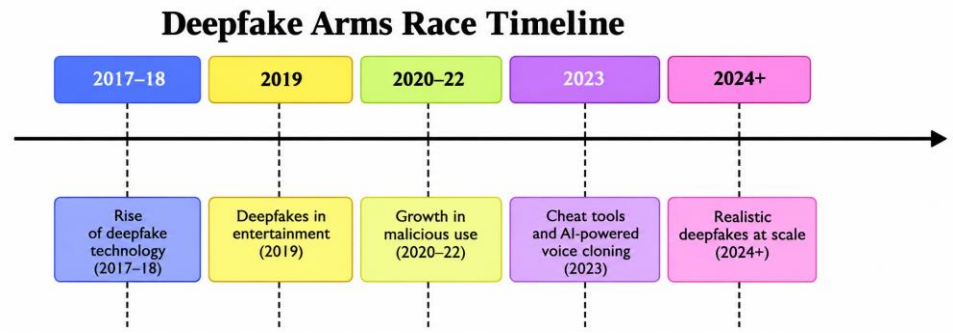


Figure 2: Timeline tracing the coevolution of deepfake generation methods (above timeline) and detection responses (below timeline)

This represents a significant escalation in the arms race. Figure 2 maps the evolution of deepfake generation architectures alongside corresponding detection responses, highlighting their adversarial coevolution. The arms race is accelerating, with diffusion models posing the latest challenge.

2.5. Key Datasets and Evaluation Benchmarks

Developing robust deepfake detectors depends on high-quality, diverse datasets. The most influential datasets include FaceForensics++, released in 2019, which contains 1.8 million frames generated by autoencoders and four generative adversarial network methods, with a balanced real-and-fake distribution. However, it is limited by low compression and limited diversity. Celeb-DF version two, released in 2019, contains 890 videos from improved autoencoders with a balanced distribution but only celebrity faces. The DeepFake Detection Challenge dataset r, released in 2020 c, contains 120,000 videos from multiple generative adversarial networks and reenactment methods with an imbalanced r, real-heavy distribution, offering high quality and diversity but at high computational expense. WildDeepfake, released in 2021, contains seven thousand in-the-wild collected videos with an imbalanced distribution, offering real-world realism but a smaller size. FakeAVCeleb, released in 2021, contains 19,000 audio-video deepfakes with a balanced distribution and is the first multimodal dataset. A consistent finding across these datasets is that within-dataset detection accuracy, for example, training and testing on FaceForensics++, can exceed ninety-nine percent using Xception networks. However, cross-dataset generalization, such as training on FaceForensics++ and testing on Celeb-DF, drops precipitously to between sixty-five and seventy-five percent, as Rossler et al. [21] and Mirsky and Lee [18] report. This lack of generalizability is a critical operational weakness.

3. Methodology

To design effective countermeasures, one must first understand the generative process mathematically and architecturally. This section formally describes the three dominant paradigms: autoencoders, generative adversarial networks, and diffusion models.

3.1. Autoencoder-Based Face Swapping

Let the source face be denoted as S and the target face as T, where T represents a frame from a video. The goal is to generate a new face F that has the identity or appearance of the source S, but the expression, pose, and lighting of the target T. A variational autoencoder consists of an encoder function E that maps an input image to a latent representation z, and a decoder function D that reconstructs the image from the latent representation. For face swapping, two autoencoders share a common decoder but have separate encoders, with E_S for the source and E_T for the target. Training minimizes a reconstruction loss that measures the difference between the original image and its reconstruction. During inference, the source face is encoded into its latent representation z_S, and the shared decoder reconstructs the source identity while implicitly learning the target's latent attributes. This method's weakness is the lack of explicit disentanglement between identity, pose and expression, which can lead to blending artifacts.

3.2. Generative Adversarial Networks

The generative adversarial network framework introduced by Ian et al. [12] models a minimax game between a generator G and a discriminator D. The generator learns to map noise sampled from a prior distribution, typically uniform or Gaussian, to synthetic samples that fool the discriminator. The discriminator outputs the probability that a given sample is real (from the real data distribution of authentic faces) or fake (from the generator). The generator is trained to minimize the logarithm of one minus the discriminator's output on fake samples. In contrast, the discriminator is trained to maximize the logarithm of its

output on real samples plus the logarithm of one minus its output on fake samples. In practice, for deepfakes, the generator's input is not pure noise but often a latent code derived from an encoder, as in StyleGAN. Figure 3 visualizes the generative adversarial network training loop, which is the core of most high-quality deepfake generation systems.

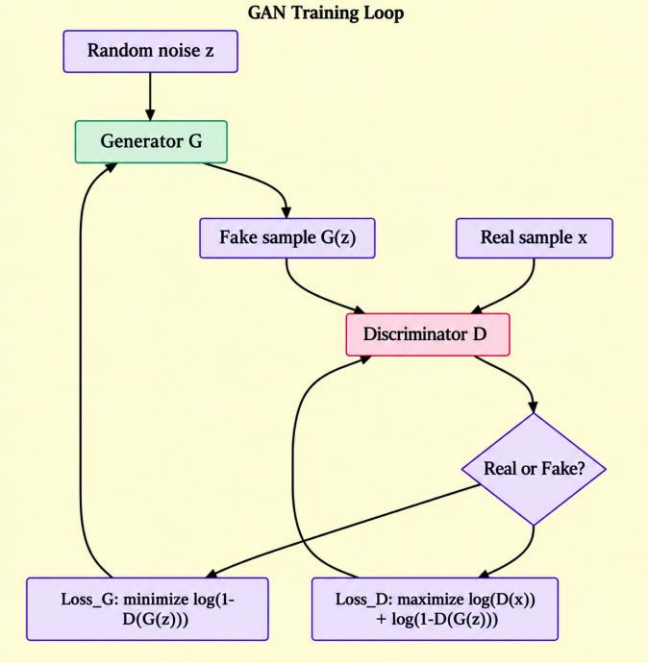


Figure 3: The generative adversarial network architecture

The generator and discriminator are trained simultaneously in an adversarial game. At inference time, only the generator is used to create deepfakes. Common generative adversarial network deepfake pipelines, such as DeepFaceLab, add additional components, including face detection and alignment using multi-task cascaded convolutional networks or S3FD; mask generation for seamless blending of the generated face onto the target video frame; and temporal smoothing using optical flow to reduce flicker between frames.

3.3. Diffusion Models as Denoising Probabilistic Models

Diffusion models define a forward process that gradually adds Gaussian noise to an image over a sequence of steps. The forward process is a Markov chain where each step adds noise according to a variance schedule that increases linearly or cosinely from a small value to a large value. After enough steps, typically 1,000, the image becomes pure isotropic Gaussian noise. The reverse process learns to denoise, starting from pure noise and gradually removing noise step by step to recover a coherent image. Training minimizes the variational lower bound, which simplifies to predicting the noise added at each step. Once trained, generating a deepfake involves sampling pure noise and iteratively applying the reverse process to denoise. Diffusion models have two key advantages over GANs for deepfake generation. First, they avoid mode collapse, making generation more stable and producing more diverse outputs. Second, for the same computational budget, diffusion models often produce more photorealistic images, especially for complex scenes. From a security perspective, diffusion models introduce fewer forensic artifacts because they lack the upsampling layers and specific normalization layers that leave generative adversarial network fingerprints, as discussed by Corvi et al. [8]. This makes diffusion-generated deepfakes the most dangerous current threat.

4. Result

A rigorous threat model is essential for designing defenses. This section categorizes deepfake attacks by their modality, including video, audio, image, and multi-modal, and by the information security objective they target. Face-swapped video deepfakes, generated using autoencoders or generative adversarial networks such as DeepFaceLab, are used for impersonation of executives in video conferences, blackmail using fake pornography, and disinformation through fake speeches. These attacks compromise integrity, authentication, non-repudiation, and confidentiality. The attacker's capability level is low because public tools are readily available. Real-world examples include video-based chief executive officer fraud and celebrity defamation. Lip-sync deepfake videos, generated using generative adversarial networks combined with audio-to-video methods such as

Wav2Lip, alter political Figures 2 and 3 words and create false evidence for legal proceedings. These attacks compromise integrity and non-repudiation.

The attacker's capability level is medium and requires audio alignment. A real-world example is the 2022 fake video of the Ukrainian president surrendering. Voice synthesis deepfakes generated using WaveNet, Tacotron 2, or SV2TTS are used for voice phishing to extract credentials, bypass voice biometrics such as bank authentication, and impersonate family members for emergency scams. These attacks compromise authentication, integrity, and non-repudiation. The attacker's capability level is low because a five-second voice sample suffices. A real-world example is the 2019 220,000-euro chief executive officer fraud. Full-body reenactment deepfakes, generated using the First Order Motion Model or pose-transfer generative adversarial networks, create fake videos of individuals performing actions they never did, such as violent acts or workplace misconduct. These attacks compromise integrity, non-repudiation, and confidentiality. The attacker's capability level is high and requires significant computational resources.

These are rare in the wild but pose a risk of corporate sabotage. Synthetic identity document deepfakes, generated using StyleGAN with optical character recognition-aware generation, create fake passports and driver's licenses for money laundering and account takeovers by bypassing know-your-customer. These attacks compromise authentication and integrity. The attacker capability level is medium. Real-world examples include synthetic identity fraud used to open bank accounts. Multi-modal deepfakes combining audio and video, generated using a combined generative adversarial network and voice cloning, enable complete digital impersonation in live video calls. These attacks compromise authentication, integrity, and non-repudiation. The attacker's capability level is high and requires real-time inference. This is an emerging threat with existing proofs of concept.

4.1. Case Study 1: The Two Hundred Twenty Thousand Euro Voice Deepfake Heist as Real-Time Impersonation

In March 2019, a United Kingdom subsidiary of a German energy firm received a call from a person claiming to be the chief executive officer of the German parent company. Using a voice deepfake system, the caller instructed the United Kingdom chief executive officer to wire 220,000 euros to a Hungarian supplier within one hour. The voice was described by the victim as having a slight German accent and the melodic qualities of the actual chief executive officer. The victim complied, and the funds were then laundered through multiple accounts.

Forensic analysis suggested the deepfake voice was likely generated with a commercial or open-source text-to-speech system, such as Tacotron 2 or SV2TTS, trained on publicly available YouTube interviews of the German chief executive officer. No video component was used, so the attack relied purely on audio, making detection harder because it lacked visual artifacts. The attack succeeded because the human authentication process, which relied on voice familiarity, was insufficient. The information security lesson is that multi-factor authentication must include out-of-band challenges such as a separate video call, a shared secret, or a cryptographic token, because voice alone is no longer a trustworthy biometric.

4.2. Case Study 2: Political Disinformation via Lip-Sync Deepfake in the 2022 Slovakian Elections

Just days before the 2022 Slovakian parliamentary elections, a deepfake video circulated on Facebook and Telegram showing the pro-opposition candidate Michal Šimečka discussing plans to rig the election by increasing value-added tax on basic food items and beer. The video was a lip-sync deepfake, meaning the candidate's actual interview footage was reanimated to make him appear to say words he never uttered. Fact-checking organizations needed 72 hours to definitively debunk the video, by which time it had been viewed over 200,000 times. The opposition candidate lost by a narrow margin, as Nadal and Jančárik [3] analyzed. The technique used was Wav2Lip or a similar generative adversarial network-based lip-sync model that takes an audio track and an existing video and accurately modifies the mouth movements to match the new audio. These deepfakes are difficult to detect with the naked eye. The information security lesson is that timeliness of detection is critical. A detection method that takes seventy-two hours is operationally useless. Security Operations Centers require near-real-time detection within minutes for political and financial disinformation.

4.3. Deepfake Detection Methodologies: A State-of-The-Art Survey

Deepfake detection is fundamentally a binary classification problem: given a media sample such as an image, video, or audio clip, determine whether it is authentic or synthetic. However, the problem is complicated by diverse generation methods, post-processing operations such as compression, resizing, and re-encoding, and adversarial attacks designed to evade detection. This section organizes detection methods into four technical domains: spatial, temporal, spectral, and biological. Figure 4 provides a comprehensive multi-domain deepfake detection taxonomy.

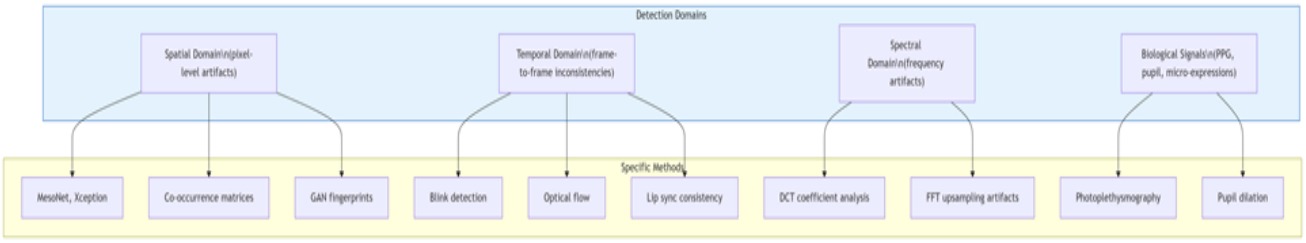


Figure 4: A comprehensive multi-domain deepfake detection taxonomy

Each domain captures different types of generation artifacts. An ensemble classifier combines the outputs for the final decision, with a human-in-the-loop for borderline cases.

4.3.1. Spatial Domain Detection

Spatial methods analyze individual frames as still images and search for pixel-level anomalies. MesoNet, developed by Afchar et al. [1], uses a compact convolutional neural network with only a few layers, including two convolutional and two fully connected layers, to focus on mesoscopic properties, textures and patterns that are neither local pixel details nor high-level semantic features. MesoNet is fast and supports real-time operation, but it performs poorly on high-quality generative adversarial network deepfakes, with accuracy dropping to about 65% on StyleGAN. Co-occurrence matrices, developed by Nataraj et al. [19], compute the spatial correlation of pixel values at various distances. Generative adversarial network-generated images have statistically different co-occurrence matrices than real images because the generator upsamples. This method is robust to compression but computationally heavy. Generative adversarial network fingerprint or noise residual analysis recognizes that generative models leave a specific noise pattern or fingerprint. A predictor network can learn to identify which generative adversarial network, such as StyleGAN versus ProGAN, generated a deepfake, achieving over ninety-five percent accuracy on known generative adversarial networks as demonstrated by Yu et al. [25].

4.3.2. Temporal Domain Detection

Temporal methods exploit inconsistencies between consecutive video frames. Human faces move in physically constrained ways, and deepfakes often violate these constraints. Eye-blinking detection, developed by Li et al. [17], was effective because early generative adversarial network-based deepfakes rarely produced convincing eye blinks, since training datasets often contained faces with eyes open. Using a convolutional neural network combined with long short-term memory to track eye state over time, this method achieved ninety-five percent accuracy on early datasets. However, modern generative adversarial networks and diffusion models explicitly model blinking, reducing this method's effectiveness.

Optical flow inconsistency, developed by Amerini et al. [2], measures pixel motion between frames. Deepfake generation often produces locally inconsistent flow fields; for example, the synthesized face may move differently from the background or the rest of the head. A support vector machine on flow features achieves about 90% accuracy on FaceForensics++. Head pose and lip synchronization analysis recognizes that head movement and lip motion are constrained. Deepfakes often show stiff head poses or asynchronous lip movements, particularly for lip-sync deepfakes. Cross-modal consistency models combining audio and visual information are an active research area.

4.4. Spectral Domain Detection

Spectral methods transform images into the frequency domain using the discrete cosine transform or the fast Fourier transform. Generative adversarial networks introduce specific frequency-domain artifacts due to upsampling layers, specifically transposed convolutions, used to increase image resolution. Frequency domain upsampling artifacts, identified by Frank et al. [11] in 2020, show that generative adversarial network upsampling leaves a grid-like pattern in the frequency spectrum, especially in high-frequency bands. A simple logistic regression classifier on frequency coefficients achieved ninety-seven percent accuracy on ProGAN images.

Discrete cosine transform coefficient analysis, developed by Durall et al. [9], shows that videos compressed with H.264 or H.265 exhibit different distributions of quantized discrete cosine transform coefficients than real videos due to multiple compressions. This method is particularly effective for detecting deepfakes on social media, where recompression is common. A critical limitation is that diffusion models do not use transposed convolutions. They use learned denoisers that produce fewer

frequency artifacts. Recent work by Corvi et al. [8] shows that spectral detectors trained on generative adversarial networks fail on diffusion deepfakes, with accuracy dropping to chance level (around 50%).

4.5. Biological Signal Detection

This nascent but powerful domain leverages the fact that living humans generate biological signals, including heartbeat measured through photoplethysmography, blood flow, and pupil dilation, that are visible in video and are extremely difficult for current generative models to reproduce. Photoplethysmography, or heart rate estimation, developed by Ciftci et al. [7] as Fake Catcher, uses the subtle color changes in facial skin caused by blood flow to compute a heart rate from video. Real videos have consistent, physiological photoplethysmography signals, while deepfakes either lack a signal or show an inconsistent, non-physiological one. Fake Catcher achieved 93% accuracy on a dataset of 2,000 deepfakes and is robust to compression. Pupil dilation consistency analysis recognizes that pupils dilate in response to cognitive load, lighting changes, and emotional arousal.

Deepfakes often ignore this, leaving pupils static or changing inconsistently. A recurrent neural network that tracks pupil size over time can differentiate real from fake with about 85% accuracy. Microexpression timing analysis recognizes that genuine microexpressions lasting between one twenty-fifth and one-fifth of a second follow specific temporal patterns. Deepfakes often produce microexpressions with unnatural durations or onset and offset times. Biological signals are hard to fake because they require modeling the body's complex, dynamic physiology. Current deepfakes do not explicitly model photoplethysmography or pupil dynamics. The disadvantage is that biological signal detection requires high-quality video at least 720p and 30 fps, with sufficient face resolution of at least 150×150 pixels. Low-light conditions or heavy compression degrade performance.

4.6. Performance Benchmarks and the Generalization Gap

Representative performance of different detector classes on standard benchmarks shows that no single detector is sufficient for operational deployment. Spatial detectors such as MesoNet achieve within-dataset accuracy of about 82% on FaceForensics++. Still, cross-dataset accuracy drops to 60% on Celeb-DF, with low robustness to JPEG compression and very low robustness to adversarial evasion. The Xception spatial detector achieves 99% on high-quality within-dataset data but only 75% cross-dataset, with medium robustness to compression and low robustness to adversarial evasion. Temporal blink detection with long short-term memory achieves 94% within-dataset but 68% cross-dataset, with low robustness to compression because blink detection fails, and adversarial evasion is not applicable because it requires video. Spectral discrete cosine transform with logistic regression achieves 97% within-dataset on generative adversarial networks but only 55% cross-dataset on diffusion models, with high robustness to compression because compression is independent, and medium robustness to adversarial evasion, which requires adding frequency noise.

Biological FakeCatcher achieves 93% within-dataset and an impressive 88% cross-dataset when the face is large enough, with medium-to-high robustness to compression. However, its robustness to adversarial evasion is unknown because it has not yet been tested. Ensemble methods combining spatial, temporal, and spectral detectors with XGBoost or late fusion achieve 98% within-dataset and 84% cross-dataset accuracy, with high robustness to compression via majority voting and moderate robustness to adversarial evasion, requiring multiple perturbations. Key observations are that no single detector is sufficient for operational deployment; ensemble methods consistently outperform individual detectors, especially in cross-dataset generalization, the most realistic scenario; and the generalization gap between within-dataset and cross-dataset performance remains the greatest research challenge. A detector that works perfectly on known deepfakes often fails on a new generation method.

4.7. The Deepfake Resilient Information Security (DRIS) Framework

Given the limitations of reactive detection alone, and because deepfakes will always improve while detectors lag behind new generation methods, this paper proposes a layered proactive-plus-reactive framework called the Deepfake Resilient Information Security (DRIS) framework. The core principles are threefold. First, assume deepfakes are possible, meaning no single video, audio, or image should be trusted without provenance or secondary verification, especially for high-impact decisions. Second, prioritize proactive provenance, because cryptographic authentication of media at the point of capture using hardware-rooted trust eliminates the detection problem for authentic media.

Third, apply reactive detection only for unauthenticated media, meaning for legacy media or media without provenance, an ensemble of detectors covering spatial, temporal, spectral, and biological domains should be used. Fourth, maintain human-in-the-loop for critical decisions, where borderline cases with detection probability between 0.3 and 0.7 are escalated to trained forensic analysts. Figure 5 presents the Deepfake Resilient Information Security architecture as a flowchart suitable for integration into a Security Operations Center workflow.

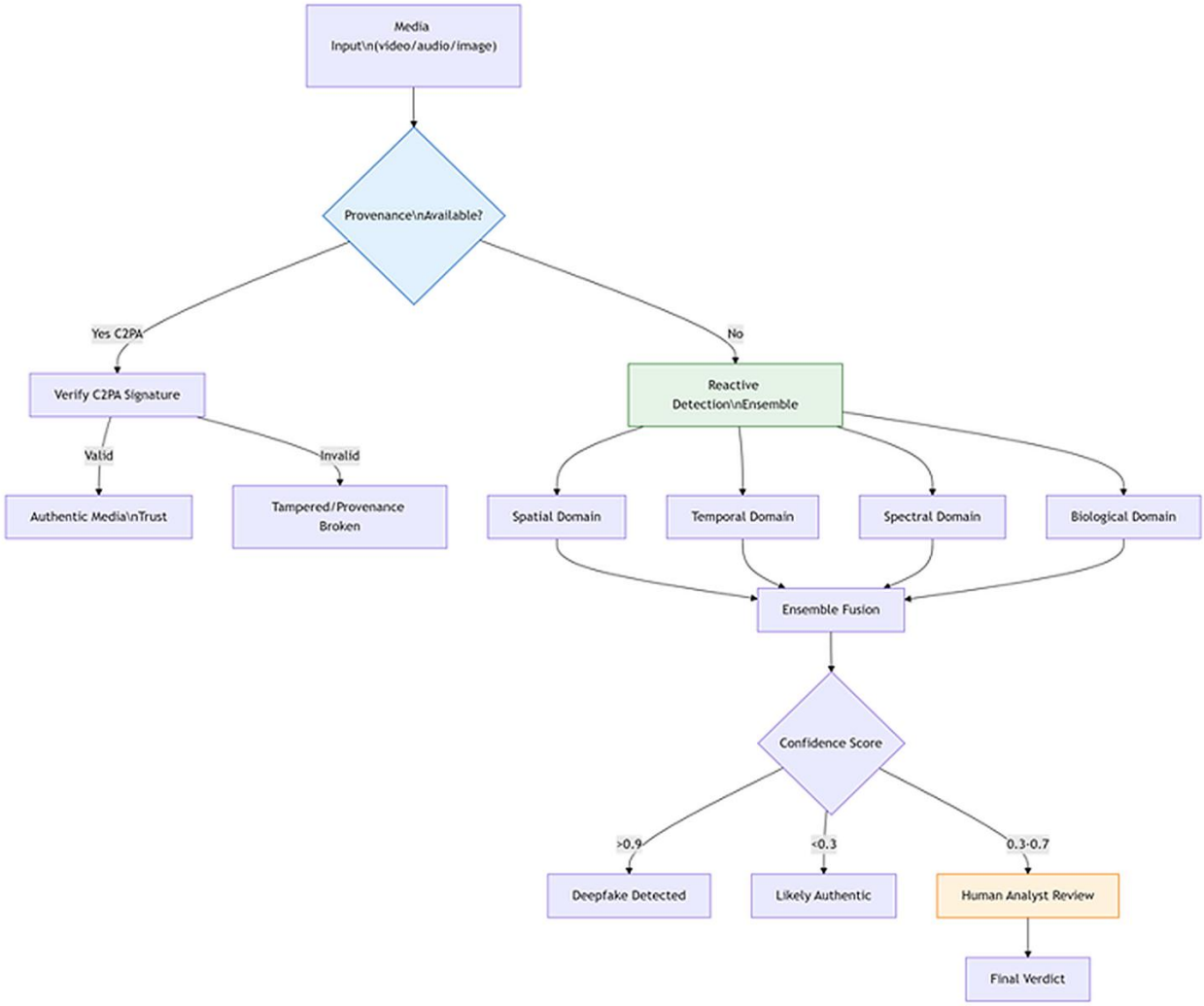


Figure 5: The deepfake resilient information security operational workflow

Proactive provenance through the Coalition for Content Provenance and Authenticity and digital signatures is checked first. Only media without provenance is passed to reactive detection ensembles. Human review handles borderline cases.

4.7.1. Proactive Provenance: The Coalition for Content Provenance and Authenticity Standard and Hardware Roots of Trust

The Coalition for Content Provenance and Authenticity, led by Adobe, Microsoft, Intel, and other industry partners, has standardized a framework that cryptographically binds metadata, called provenance, to media at the moment of capture. Provenance includes who captured the media (device identifier and software version), when and where it was captured (timestamp and GPS coordinates if enabled), and what edits were made through a signed chain of edits. A Coalition for Content Provenance and Authenticity-compliant camera, available from certain Sony and Leica models, signs each image with a private key embedded in the hardware. Any subsequent change, such as resizing or color correction, must add a new signed manifest. Consumers can verify the entire chain against public keys. For information security, organizations should mandate Coalition for Content Provenance and Authenticity provenance for all official communications, including internal video messages, recorded meeting minutes, and official announcements. Media without provenance should be treated as untrusted by default.

4.7.2. Blockchain Notarization

For sensitive legal or financial media such as evidence video for court, a SHA256 hash of the original media can be registered on a permissioned or public blockchain such as Ethereum or Hyperledger. Any subsequent claim about the media can be verified

against the immutable blockchain record. This does not prevent deepfakes but provides a strong non-repudiation mechanism, proving that this specific media existed at this specific time in an unaltered state.

4.7.3. Digital Watermarking Robust to Attacks

Two types of watermarks are relevant for deepfake defense. Visible watermarks, such as logos and overlays, are easily removed by deepfake pipelines and offer little security. Invisible robust watermarks, such as Fourier-based or spread spectrum watermarks, embed a pattern into the discrete cosine transform or wavelet coefficients that survives compression, resizing, and even some deepfake manipulations. However, adversarial deepfake generators can be trained to remove watermarks, making this an active area of research.

5. Discussions

For organizations wishing to operationalize deepfake defense, this section provides a prioritized set of actions, tooling recommendations, and metrics.

5.1. Immediate Actions within Zero to Three Months

Organizations should conduct awareness training to educate employees about deepfake voice phishing. A simple test is to ask the caller, on a sensitive request, to turn on their video and perform a specific action, such as touching their left ear with their right hand, because current deepfakes struggle with out-of-vocabulary actions. Organizations should implement multi-factor verification for financial transfers, requiring that any wire transfer over a threshold, for example, ten thousand United States dollars, that requires a voice call must also require a second out-of-band confirmation such as a separate video call through a different channel or a cryptographic token. Organizations should integrate a basic deepfake detector into the email gateway, using open-source detectors such as the deepfakedetector Python package to scan all incoming video attachments and flag high-probability deepfakes for quarantine.

5.2. Medium-Term Actions within Three to Twelve Months

Organizations should deploy an ensemble detector in the Security Information and Event Management system, integrating a multi-domain detector that combines Xception, frequency analysis, and a lightweight temporal model into the Security Information and Event Management pipeline and logging all detection scores for threat hunting. Organizations should adopt the Coalition for Content Provenance and Authenticity standard for internal content creation, requiring that all internal training videos, executive messages, and meeting recordings be captured using Coalition for Content Provenance and Authenticity-enabled tools such as Adobe Aero or Microsoft Teams with provenance enabled, and authenticating incoming external media against trusted provenance when available. Organizations should establish a forensic deepfake review team and train two to three analysts in manual deepfake forensic tools such as the FaceForensics++ toolkit, generative adversarial network fingerprint extraction, and photoplethysmography analysis to handle borderline cases with detection scores between 0.3 and 0.7.

5.3. Long-Term Strategic Actions beyond Twelve Months

Organizations should develop a deepfake incident response playbook modeled on the ransomware playbook, including steps such as detection, containment by quarantining media, forensic analysis to determine which generative adversarial network or diffusion model was used, eradication by retracting false information, recovery through public correction, and legal preservation of evidence for prosecution. Organizations should participate in deepfake detection challenges, as regular competitions such as the DeepFake Detection Challenge on Kaggle provide new datasets and benchmarks that can update detection models. Organizations should research adversarial hardening, using adversarial training that adds small perturbations to deepfakes to fool detectors, making their own detectors more robust.

5.4. Metrics for Measuring Defense Effectiveness

The deepfake detection rate measures the percentage of deepfakes detected as true positives at a given false positive rate, with a target of over 90% detection at a 1% false positive rate. The mean time to detect measures the average time from media ingestion to classification as deepfake or real, with a target of under two minutes for automated detection and under four hours for human review.

The false positive rate measures the percentage of authentic media incorrectly flagged as deepfake, with a target of under 0.1% to avoid alert fatigue. Provenance coverage measures the percentage of critical media assets, such as executive videos, with verifiable Coalition for Content Provenance and Authenticity provenance, targeting over 95%.

5.5. Open Challenges and Future Research Agenda

Despite significant progress, deepfake defense remains unsolved. Six grand challenges are highlighted to define the research agenda for the years 2025 through 2030.

Challenge 1: Generalizable Detection across Unseen Generators: Current detectors fail dramatically when tested on a generation method not seen during training; for example, they may be trained on generative adversarial networks and tested on diffusion models. Research should focus on developing style-agnostic features invariant to generator architecture. Potential candidates include high-order statistical moments of noise residuals or learned embeddings from self-supervised models such as SimCLR that separate real from fake regardless of generation method.

Challenge 2: Real-Time Deepfake Detection for Live Streams: Most detection methods are offline and process entire videos. Live deepfake attacks, such as real-time face swaps in video calls, require detection within milliseconds with low latency. Research should focus on lightweight MobileNet-based models that run on edge devices such as conferencing endpoints, using streaming temporal windows covering the last thirty frames rather than the full video.

Challenge 3: Adversarial Robustness: Attackers can generate adversarial examples, which are small, imperceptible perturbations added to a deepfake that cause a detector to misclassify it as real. Evasion attacks are well documented against image classifiers, as Carlini and Farid [4] showed. Research should focus on adversarial training that includes adversarial deepfakes in the training set, as well as certified robustness using randomized smoothing.

Challenge 4: Multi-Modal Detection for Audio-Video Consistency: Existing detectors treat audio and video separately. A powerful future direction is cross-modal consistency, where the audio waveform or speech must correlate with the lip movements and facial microexpressions. Research should focus on self-supervised models that embed audio and video into a shared latent space and measure consistency, flagging inconsistent pairs as deepfakes.

Challenge 5: Deepfake Attribution and Forensics: Beyond binary real-or-fake classification, law enforcement and counterintelligence need to attribute a deepfake to the specific generation tool, or even the specific user. Research should focus on neural forensics, extracting unique fingerprints from the generator's internal weights, such as specific kernel patterns, to identify which software, for example, DeepFaceLab version 2.1 versus FaceSwap version 1.0, created the deepfake. This is analogous to ballistic fingerprinting.

Challenge 6: Defending Against Diffusion Model Deepfakes: Diffusion models produce fewer artifacts than generative adversarial networks. Early evidence suggests that diffusion deepfakes are much harder to detect. Research should focus on biological signals such as heart rate and pupil dilation, as well as physical constraints through inverse rendering, asking whether the lighting, shadows, and geometry make physical sense. Diffusion models often fail to model global illumination correctly.

5.6. Policy and Standards Recommendations

Technology alone cannot solve the deepfake threat. This section provides recommendations for policymakers, standards bodies, and industry consortia.

Mandate Content Provenance through the Coalition for Content Provenance and Authenticity Standard: Within three years, all major social media platforms, news organizations, and government agencies should require provenance for videos and images they consider authoritative, using the Coalition for Content Provenance and Authenticity standard. Media without provenance must be labeled as unverified. The European Union's Digital Services Act and similar legislation should be amended to include provenance requirements for political advertising.

Criminalize Malicious Deepfake Creation: Laws should distinguish between benign or artistic deepfakes used for parody, satire, or education, which constitute protected speech, and malicious deepfakes used for fraud, defamation, non-consensual pornography, or election interference, which should carry criminal penalties with extradition for cross-border cases. The proposed United States DEFIANCE Act and similar South Korean laws from 2020 provide templates.

Fund Public Detection Tools and Datasets: Governments should fund publicly available deepfake detection application programming interfaces, analogous to VirusTotal for deepfakes, and benchmark datasets that reflect real-world threats, including low-quality, compressed, and adversarially attacked deepfakes.

Establish Forensic Standards for Courts: Courts are increasingly facing deepfake evidence. Standards bodies such as the National Institute of Standards and Technology and the International Organization for Standardization should publish

guidelines for deepfake forensic analysis that are admissible in court, including chain-of-custody requirements for media, detection-confidence thresholds exceeding 99.9%, and certification for forensic analysts.

6. Conclusion

Deepfake technologies have evolved from a curiosity to a potent weapon that systematically undermines the integrity, authentication, and non-repudiation pillars of information security. This paper has provided a comprehensive, technically rigorous, and operationally oriented analysis. The paper deconstructs generation architectures, including autoencoders, generative adversarial networks, and diffusion models. These have been mapped to specific information security attack vectors through case studies of chief executive officer fraud and political disinformation. A state-of-the-art survey of detection methods across spatial, temporal, spectral, and biological domains is presented. A layered Deepfake Resilient Information Security framework that prioritizes proactive provenance over reactive detection has been proposed. The key takeaways for security professionals are fourfold. First, no single detection method is sufficient, so organizations must deploy ensembles covering spatial, temporal, spectral, and, where possible, biological domains. Second, organizations must assume that deepfakes will bypass their detector, so critical decisions must be verified through out-of-band channels such as a second video call with a challenge. Third, organizations should adopt provenance through the Coalition for Content Provenance and Authenticity standard for internal media, because cryptographic authentication eliminates the detection problem for media they control. Fourth, organizations must prepare incident response playbooks because deepfake attacks will happen, and they must plan for quarantine, notification, and legal preservation. The arms race between deepfake generation and detection will continue for the foreseeable future. Diffusion models and real-time synthesis will raise the bar, while advances in biological signal detection and cross-modal consistency may provide new countermeasures. However, the ultimate solution is likely socio-technical: a combination of technical provenance, legal deterrence, and a digitally skeptical culture where media is not trusted without verification. For information security, the age of seeing is believing has ended. The age of verify-or-distrust has begun.

Acknowledgment: The authors sincerely acknowledge EBLA Private University, Jinsha Interactive, Indiana University Bloomington, the University of Texas, Queen's University Belfast, and the University of Kufa for their valuable academic support, institutional resources, and encouragement.

Data Availability Statement: The research contains data related to Deepfake Technologies in Information Security: Generation Techniques and Multimodal Detection Approaches

Funding Statement: This research received no financial support or external funding from any organization, institution, or funding agency.

Conflicts of Interest Statement: The authors declare that they have no competing interests or conflicts of interest related to the research, authorship, or publication of this manuscript. All sources used have been appropriately acknowledged and referenced.

Ethics and Consent Statement: Ethical considerations were observed throughout the research process. Necessary permissions were obtained from the concerned organization, and informed consent was secured from individual participants involved in data collection. Ethical approval and participant consent were obtained wherever applicable.

References

1. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," In *2018 IEEE international workshop on information forensics and security,* Hong Kong, China, 2018.
2. I. Amerini, L. Galteri, and R. Caldelli, "Deepfake Video Detection through Optical Flow Based CNN," in *2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),* Seoul, South Korea, 2019.
3. L. D. Nadal and P. Jančárik, "Beyond the deepfake hype: AI, democracy, and 'the Slovak case'," *Harvard Kennedy School (HKS) Misinformation Review*, vol. 5, no. 4, pp. 1-9, 2024.
4. N. Carlini and H. Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, Washington, United States of America, 2020.
5. R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 12, pp. 1753–1819, 2019.
6. Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, United States of America, 2018.

7. U. A. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Houston, Texas, United States of America, 2018.
8. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.
9. R. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, United States of America, 2020.
10. H. Farid, "Image Forgery Detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.
11. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in *Proc. 37th International Conference on Machine Learning (ICML),* vol. 119, no. 11, pp. 3247-3258, 2020.
12. J. Ian, I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, no. 12, pp. 2672–2680, 2014,
13. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *arXiv preprint* arXiv:2006.11239, 2020, [Accessed by 19/06/2024].
14. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, United States of America, 2019.
15. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, United States of America, 2020.
16. J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?," *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020.
17. Y. Li, M. C. Chang, and S. Lyu, "In ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS),* Hong Kong, China, 2018.
18. Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
19. L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, "Detecting GAN generated Fake Images using Co-occurrence Matrices," *arXiv preprint arXiv:1903.06836*, 2019, [Accessed by 15/03/2024].
20. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution Image Synthesis with Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, Louisiana, United States of America, 2022.
21. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019.
22. A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation," in *Advances in Neural Information Processing Systems*, vol. 32, no. 12, pp. 1-11, 2019.
23. C. Stupp, "Fraudsters used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *The Wall Street Journal*, 2019, [Accessed by 15/08/2024].
24. L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
25. N. Yu, L. S. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

Publisher's Note: *The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.*