Journal of Computer Science

Review Article

Deepfake Technology: A Comprehensive Review of Trends, Applications, Ethical Concerns, and Challenges

Battula Thirumaleshwari Devi and Rajkumar Rajasekaran

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Article history
Received: 19-01-2025
Revised: 10-06-2025
Accepted: 03-07-2025

Corresponding Author:
Rajkumar Rajasekaran
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
Email: vitrajkumar@gmail.com

Abstract: The digital age has fundamentally transformed how information is created and disseminated, raising critical concerns about the authenticity and trustworthiness of online content. Recent advances in Artificial Intelligence (AI), particularly deep learning, have given rise to deepfakes: highly realistic synthetic media generated by manipulating or replacing faces, voices, and actions in videos. While deepfake technology offers innovative applications across various industries, its rapid proliferation has also enabled malicious uses, including fake news, financial fraud, identity theft, and cyberattacks. Consequently, robust deepfake detection has become essential to preserving digital integrity and mitigating social and security risks. This paper presents a comprehensive review of deepfake technology, examining its creation techniques (e.g., autoencoders, generative adversarial networks), diverse media types (text, image, audio, video), and evolving detection methods. It also surveys publicly available datasets and evaluates the performance of state-of-the-art detection models. Beyond technical aspects, the review critically discusses the ethical, legal, and societal implications of deepfakes, including privacy violations, consent, misinformation, and regulatory challenges. By synthesizing current trends and identifying research gaps, this study aims to provide a balanced understanding of both the potential benefits and threats posed by deepfakes, and to inform future efforts in detection, governance, and responsible use.

Keywords: Manipulation, Deepfake Creation, Deepfake Detection, Autoencoders, GAN, and LSTM

Introduction

The digital era has fundamentally transformed the creation, dissemination, and consumption of information, bringing with it unprecedented challenges to the authenticity of online content. Among recent technological advances, Artificial Intelligence (AI) stands out as one of the most transformative, with applications spanning virtually every sector. A particularly concerning manifestation of AI's power is the emergence of deepfakes, highly realistic synthetic media generated through deep learning techniques. The term "deepfake" itself is a portmanteau of "deep learning" and "fake", and typically refers to the use of neural networks to swap a target individual's face onto another person's body in video, creating convincingly realistic footage of events that never occurred (Aslam & Santhi, 2019).

Deepfake techniques leverage large image and video datasets to generate highly realistic media. Celebrities and politicians, who present a lot of usable content on social networks or other platforms, are primarily at risk of deepfakes. For instance, Rana Ayyub, an Indian journalist, received death threats courtesy of bad actors by impersonating her in a pornographic video that circulated on social media platforms such as Twitter and WhatsApp. At first, deepfakes targeted the switching of faces of different celebrities or political figures with those of other actors in pornography. Deepfake videos, the kind in which a person's image is substituted for another's, initially surfaced in 2017 with one in which a star's head was superimposed on an adult film actor's body. The priority is obvious when Deepfake is used to portray several world leaders giving faux speeches in politically motivated falsification pernicious to global security. Deepfakes pose threats beyond politicians and celebrities, affecting individuals across various domains. For instance, a voice deepfake was used in a scam to con a CEO into releasing $243,000 (Akhtar, 2023): a demonstration of the use of deepfake in other vices and immoralities. To avoid such risks, the field of deepfake detection has received a lot of attention from scholars and practitioners resulting in the emergence of many solutions that seek to detect deepfake content.

SCIENCE Publications

Overview of Deepfake Technology

Deepfake technology refers to the use of artificial intelligence, particularly deep learning algorithms, to generate synthetic content, such as videos, images, or audio, that closely mimics real people. The core technique underpinning most deepfakes is the Generative Adversarial Network (GAN), which comprises two competing neural networks: a generator that creates fake content and a discriminator that attempts to distinguish it from real content. Through this adversarial process, the generator iteratively improves, eventually producing highly convincing forgeries (Cassia et al., 2025). Other AI techniques, such as autoencoders, are also commonly employed, especially for face swapping, superimposing one individual's face onto another's body in video footage.

Although manipulated images and videos predate modern AI, deepfake technology proliferated rapidly during the 2010s with the advent of deep learning. GANs, first introduced by Goodfellow et al. (2014), represented a major breakthrough. Soon after, AI-powered face-swapping applications emerged, enabling more automated and precise manipulations. By 2017, deepfakes captured public attention as manipulated videos of celebrities circulated online, revealing the technology's dual-use potential: while it offers creative opportunities in entertainment, it also enables harmful applications such as misinformation and non-consensual pornography (Barni et al., 2020). As deep learning models and computational power continue to advance, deepfake generation tools have become both more sophisticated and more accessible, allowing even non-experts to produce realistic synthetic media.

This study makes several key contributions to the field of deepfake research. First, it provides a systematic categorization and in-depth examination of various deepfake types, elucidating their underlying creation methods and distinguishing characteristics. By doing so, it offers readers a clear technical foundation for understanding how different forgeries are generated.

Second, the paper presents a balanced analysis of both the beneficial applications and the potential threats posed by deepfake technology. It explores constructive uses in domains such as entertainment, education, and healthcare, while also critically examining risks related to misinformation, fraud, and privacy violations. This dual perspective contributes to a more nuanced understanding of deepfakes as a dual-use technology.

Third, the study surveys current datasets commonly used for training and evaluating deepfake detection systems. It assesses their strengths and limitations, identifies critical gaps, such as lack of diversity, poor ecological validity, and insufficient multimodal content, and underscores the need for more robust and representative benchmarks.

Finally, the paper foregrounds the ethical, legal, and social implications of deepfake technology. It addresses pressing concerns including consent, digital identity manipulation, regulatory responses, and the erosion of public trust. By synthesizing these dimensions, the study aims to foster informed discourse and support the development of responsible innovation, effective governance, and sustainable countermeasures.

Different Types of Deepfakes

AI technology has advanced, and it is now possible to create content that is very believable while it is fake. This manipulation can take formats that can affect all types of media, whether written, audiovisual, or both. All of them come with certain difficulties and dangers: from sharing fake news to phishing or impersonating people. The different types of deepfakes are explained below and illustrated in Fig. 1.

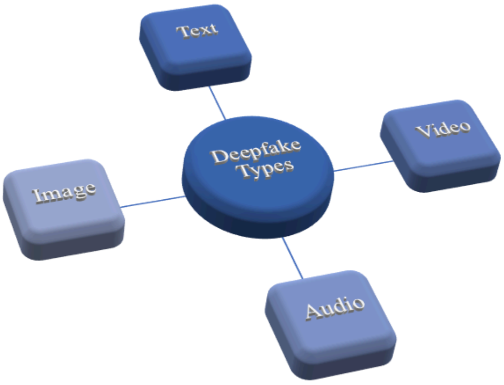

Fig. 1: Different Types of Deepfakes

Text: These are forms of text that mimic people's writing styles because they are created by AI models. They can be used to create false news, impersonate individuals in messages or emails, and spread disinformation. These texts often replicate tone, style, and context to appear authentic. Their realism poses risks to trust, security, and online communication.

Image: Deepfake approaches in images mean either synthesizing another image with the same person to have an entirely new photo or replacing a person's face in an image with a fake one. This is usually done with some ill intention, like opening a fake account or bullying people online.

Video: Commonly used deepfakes known as face swap, where someone's face or behaviour is changed to depict him/her as saying or doing things he/she never had or said. This type has been employed in positive uses, including entertainment, as well as negative uses.

Audio: Audio deepfakes mean producing fake audio based on any human speech, which artificially imitates that individual's voice to say something they never said.

It could be used by scammers, impersonators, or anyone who wants to change public opinion about something.

Importance and Impact of Deepfakes

Deepfakes have profound societal, political, and economic implications, as their ability to create highly realistic, fabricated media poses significant risks across multiple domains. In society, deepfakes can be used to spread misinformation or fake news, potentially undermining public trust in media and institutions. Politically, they can be weaponized to manipulate public opinion, influence elections, or defame individuals through the creation of misleading videos of politicians or public figures. Economically, businesses can be targeted by deepfake fraud, such as impersonating executives to authorize fraudulent transactions, which can lead to financial losses and damaged reputations. These impacts raise significant ethical and legal concerns (Schiff et al., 2023). On an ethical level, deepfakes challenge the integrity of information, creating dilemmas around consent, privacy, and the potential for harm. Legally, the rapid evolution of deepfake technology has outpaced the development of regulations, leading to difficulties in prosecuting malicious uses, protecting individuals from defamation, and ensuring accountability for the creation and distribution of harmful content (Baltrušaitis et al., 2019). The tension between protecting free speech and regulating harmful uses of deepfakes further complicates the legal landscape, requiring updated laws and international cooperation (Mubarak et al., 2023).

Deepfake Fraud Progression From 2020 to 2024

Over the past five years, deepfake fraud has undergone significant evolution, driven by advancements in AI technology. The increasing accessibility of tools for creating synthetic media has amplified the use of deepfakes in cybercrime, misinformation, and fraud (Lee et al., 2025). Table 1 provides a detailed explanation of the progression of deepfake fraud over the past five years.

Table 1: Deepfake Fraud Progression From 2020 to 2024

Year	Key Events	Key Trends
2020	- Launch of user-friendly deepfake creation tools. - Voice-based phishing scams emerge. - Governments begin exploring regulatory responses.	- Democratization of deepfake generation. - Initial awareness of audio-based fraud. - Early legal exploration.
2021	- Surge in non-consensual deepfake pornography (approx. 90% of all deepfake content). - High-profile scams in finance and politics. - Pilot deployment of AI-based detection tools.	- Rising ethical and legal concerns. - Increased public and media attention. - Early corporate interest in countermeasures.
2022	- CEO voice impersonation attacks exposed. - Interdisciplinary collaborations (AI ethics + cybersecurity). - The visual and audio quality of deepfakes significantly improves.	- Biometric systems targeted. - Cross-domain counter-strategy development. - Realism challenges traditional detection.
2023	- Detection systems deployed in defense, banking, and media. - Introduction of the EU AI Act and U.S. DEEPFAKES Accountability Act. - Public awareness campaigns grow; detection remains difficult for users. - Deepfake creators use adversarial techniques.	- Institutional and policy-level responses intensify. - Consumer uncertainty persists. - Detection tools face evasive tactics.
2024	- Deepfakes breach biometric authentication systems. - Global demand for coordinated regulation rises. - Improved AI-based countermeasures developed.	- Systemic cross-sector threat emerges. - Calls for global governance grow. - Regulation and resilience become priorities.

Deepfake Attacks on Biometric Authentication Systems

The use of deepfake generates threats to biometric authentication because it enables imposters to take on the unique identity traits of real people. Because of progress in GANs and neural synthesis, fraudsters now can create realistic synthetic videos and recordings that can trick both smartphone and surveillance security systems (Dani & Mustafa, 2025). AI technology in videos can display realistic reactions and movements, bypassing features like liveness checks in face recognition. In another way, cloned voices can fool voice-recognition systems and allow people to get access or request that fraudulent orders be made (Garg et al., 2025). These threats point out that many biometric systems are unable to tell real data from simulated information. It is necessary to improve detection algorithms that spot minor differences introduced in the manufacturing stage, as well as to use authentication methods that look at both the biometrics and how a person behaves. To guard the integrity of these systems, AI, anomaly detection, and probably instant verification can be employed.

Objectives and Scope of the Study

The primary objective of this review is to provide a comprehensive and critical analysis of deepfake technology by exploring its generative mechanisms, real-world applications, ethical and legal challenges, and the latest advancements in detection techniques. This paper aims to bridge the gap between technical understanding and societal implications by examining both foundational methods (e.g., GANs and autoencoders) and emerging paradigms such as Neural Radiance Fields (NeRFs).

To maintain focus and coherence, this study is organized around a set of clearly defined objectives that span the technical, societal, and forward-looking dimensions of deepfake technology.

The first goal is to systematically categorize and explain the major types of deepfakes, including image, video, audio, and text-based forgeries, alongside the underlying synthesis techniques such as Generative Adversarial Networks (GANs), autoencoders, and emerging methods like Neural Radiance Fields (NeRFs). Building on this foundation, the study further analyzes the expanding influence of deepfakes across critical domains, including media and entertainment, politics, and cybersecurity, illustrating both their innovative potential and their capacity for harm.

In parallel, the research evaluates state-of-the-art deepfake detection strategies, benchmarking their performance across diverse datasets and identifying key limitations related to generalization, robustness, and real-world deployment. Beyond technical assessment, the study also explores the multifaceted ethical, legal, and psychological challenges posed by deepfakes, particularly issues of consent, privacy, digital identity, and the psychological impact of identity manipulation.

Finally, the paper highlights emerging trends and charts future research directions. These include the rise of 3D-aware synthetic media, the development of hybrid and multimodal detection frameworks, and the growing need for policy interventions and governance mechanisms. By integrating these objectives, the study aims to provide a holistic and forward-looking perspective that informs both researchers and practitioners working to mitigate the risks of deepfakes while harnessing their beneficial applications.

Literature Review

Text

Recent advances in large language models (LLMs) have enabled machines to generate text that closely mimics human writing across a wide range of domains, including news articles, storytelling, and even scientific publications. This rapid progress has blurred the distinction between human-authored and machine-generated content, intensifying the need for robust deepfake text detection methods to mitigate risks such as misinformation, plagiarism, and fraudulent content. (Wiseman et al., 2017). The increasing sophistication of large language models (LLMs) has introduced significant societal challenges, most notably the proliferation of fake news and plagiarism. Despite advances in detection techniques, existing methods are typically evaluated in constrained settings, limited to specific domains or generative models, which severely limits their applicability in real-world environments. In practice, detection systems must contend with texts generated by an array of unknown models and drawn from diverse, often unseen domains, all without prior knowledge of their provenance. To address this gap, researchers have proposed "wild testbeds" for deepfake text detection that aggregate human-written content alongside outputs from various LLMs. Even under these more realistic conditions, however, human annotators perform only marginally above chance when distinguishing synthetic from authentic text. Automated detection methods, when evaluated across a broad spectrum of real-world deepfake content, similarly struggle, particularly when confronted with out-of-distribution samples that diverge from their training data (Ackley et al., 1985; Akhtar, 2023). This lack of generalization remains a critical barrier to deploying reliable detection systems in practice. Among the approaches assessed, supervised methods fine-tuned on pre-trained language models (PLMs) have demonstrated the strongest performance. Nevertheless, they continue to falter when faced with texts originating from unseen domains or previously unencountered generative architectures (Tipper et al., 2024). Encouragingly, recent work indicates that optimizing decision boundaries can substantially enhance out-of-distribution robustness, suggesting that highly effective deepfake text detection in real-world contexts remains an attainable goal.

Recent advances in text generation have been driven in part by increasingly sophisticated decoding strategies. Techniques such as Top-k sampling and Top-p (nucleus) sampling modify the token selection process by reshaping the probability distribution over the vocabulary. Top-k restricts the candidate pool to the k most likely tokens, while Top-p selects tokens whose cumulative probability exceeds a predefined threshold, both approaches yield less repetitive and more naturally varied text (Rumelhart et al., 1986). Temperature sampling further influences generation by scaling the logit distribution prior to the softmax function: lower temperatures produce more deterministic, conservative outputs, whereas higher temperatures increase diversity and randomness. Before the advent of transformer architectures, natural language processing (NLP) relied heavily on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. Despite their contributions, these models suffered from fundamental limitations, including difficulty in maintaining coherence over long sequences and an inherent inability to parallelize training due to their sequential nature (Hochreiter, 1998; Bengio et al., 1994; Graves, 2012). The introduction of the Transformer model in 2017 marked a paradigm shift. By leveraging a self-attention mechanism, Transformers process all tokens in parallel, dramatically accelerating both training and inference. This architecture employs bidirectional attention for tasks like text classification and unidirectional (causal) attention for autoregressive generation, ensuring coherent left-to-right flow. Today, Transformers form the backbone of state-of-the-art NLP systems, including BERT, GPT-2, and GPT-3, enabling unprecedented efficiency and power in text generation.

Image

Recent advances in face swapping and AI-generated facial imagery have introduced increasingly sophisticated techniques that raise critical concerns regarding identity protection and unauthorized system access. These synthetic artifacts are often visually indistinguishable from genuine photographs, making it exceedingly difficult for users to differentiate between authentic content and forgeries produced by Generative Adversarial Networks (GANs). Early detection methods framed the problem as a binary classification task, predominantly leveraging Convolutional Neural Networks (CNNs) (Dang et al., 2018). These approaches focused on extracting spatial features such as inconsistencies in facial texture, blurring artifacts, contrast anomalies, and steganalysis-based cues that capture hidden patterns within images. Subsequent developments introduced enhanced CNN architectures, including spatial expectation masking and hierarchical feature analysis within the Xception Network, to improve detection accuracy (Hsu et al., 2020). Despite these advancements, the rapid evolution of GAN-based generation techniques underscored the need for more generalized and robust forensic methods. In response, Paris (2023) proposed the use of preprocessing techniques such as Gaussian blur and Gaussian noise to steer models away from low-level pixel artifacts and toward statistically meaningful image features. Similarly, Zhou et al. (2017) developed a CNN-based framework that first extracts facial features and then fine-tunes the model to discriminate between real and manipulated faces, further advancing the adaptability of detection systems.

The challenge of detecting images generated by increasingly sophisticated GANs has prompted the development of specialized forensic architectures. Chollet (2017) introduced a forensic CNN that incorporates Gaussian preprocessing to facilitate discrimination between real and manipulated images. This method exploits a key statistical distinction: authentic images tend to exhibit low-frequency spectral artifacts, whereas GAN-generated images often contain high-frequency pixel-level noise. Despite its effectiveness, this approach also highlighted persistent issues such as model overfitting and the need for greater architectural robustness. In response, pairwise-learning frameworks have emerged as a promising direction for generalizing to unseen forgery types (Paris, 2023). Do et al. (2018), for instance, proposed a two-phase model that learns feature representations of real and fake images by projecting real-fake pairs through a shared common feature network (CFFN). The integration of contrastive loss enabled the model to detect novel fakes, including those produced by GANs not encountered during training, with higher precision and recall than previous state-of-the-art methods.

A range of hybrid approaches has emerged, combining multiple detection techniques to improve forensic performance (Sabir et al., 2019). One notable framework, proposed by Liu et al. (2019a), integrates two complementary streams: the first classifies images as tampered or non-tampered, while the second leverages steganalysis to extract low-level camera noise and residual artifacts. This dual-stream design enables the model to exploit both perceptible and latent traces of manipulation, significantly enhancing its detection capability. In parallel, Do et al. (2018) developed a neural network tailored to detect GAN-generated videos, emphasizing preprocessing steps that enhance statistical feature extraction for improved face forgery detection. Another hybrid strategy combined GAN-generated imagery with pairwise learning to isolate distinguishing features between authentic and altered content. Collectively, these models have demonstrated a consistent ability to surpass the limitations of earlier detectors and adapt to the evolving challenges posed by successive generations of GANs (Liu et al., 2019b). Despite these advances, a critical limitation persists: many state-of-the-art detectors struggle to generalize across diverse datasets and unseen GAN variants, largely because they are frequently trained and evaluated on data drawn from similar distributions (Barni et al., 2020). Addressing this generalization gap is essential for the development of robust, real-world forensic tools. Future research must therefore prioritize architectural adaptability and cross-domain resilience to enable reliable detection of an ever-expanding landscape synthetic media.

Audio

Audio manipulation encompasses a combination of artificial intelligence techniques used in deepfakes and simpler editing methods, such as adjusting playback speed, trimming, or altering context, collectively referred to as "cheap fakes" (Bengio et al., 1994). One effective tool in detecting deepfake audio is Resemblyzer, an open-source solution that extracts high-level audio representations. These representations allow developers to compare two voice samples to identify inconsistencies at any point. Some other ways of fake audio detection include assessing dissimilarities in spectrograms which are visual representations of particular signals in the sound area to recognize genuine signals or synthetic ones (Oord et al., 2016).

Recent innovations in audio deepfake detection have been bolstered by advanced neural architectures such as WaveNet, a deep convolutional neural network (CNN) originally developed for text-to-speech (TTS) and speech recognition applications (Pfefferkorn, 2019). TTS systems synthesize speech from text or phoneme inputs, while voice conversion (VC) systems transform the vocal characteristics of an audio sample to mimic a different speaker while preserving the original linguistic content. Both technologies, while enabling powerful generative capabilities, also introduce new vectors for synthetic speech manipulation. Deep neural network (DNN)-based approaches have demonstrated marked superiority in extracting dynamic acoustic features and assessing the authenticity of audio signals. Unlike conventional methods such as Gaussian mixture model (GMM) classifiers, which primarily rely on static feature representations, DNNs capture temporal dependencies and subtle spectral variations that are critical for distinguishing genuine from synthetic speech (Reynolds, 2009). Empirical studies have shown that these DNN-based detectors consistently outperform their predecessors, offering greater accuracy and robustness in identifying AI-generated voice content.

Video

Deepfake technologies, which manipulate facial features in video content, pose a significant threat to digital authenticity. Manipulations such as face-swapping or expression synthesis often introduce inconsistencies in lighting, head pose, and geometric alignment. These irregularities leave behind "digital footprints" in the form of residual signals, which serve as critical markers for forensic detection. Notably, Afchar et al. (2018) proposed a method tailored to detecting these artifacts, specifically focusing on face-warping disparities. Unlike traditional approaches that rely on computationally intensive image processing to generate negative datasets, this method significantly reduces algorithmic complexity while maintaining detection efficacy.

Current deepfake algorithms often struggle to simulate physiological signals, most notably natural blinking patterns. Because many models lack sufficient training data for closed-eye states, they frequently produce unnatural or infrequent blinking. Early research by Soukupová & Cech (2016) and Bansal et al. (2018) utilized contour circle fitting for pupil detection and blink-rate estimation with high precision. Building on these foundations, Wu et al. (2020) introduced a hybrid CNN-RNN architecture to capture the temporal dynamics of blinking. This approach outperformed traditional methods reliant on facial landmarks and static classifiers like Support Vector Machines (SVM) or Hidden Markov Models (HMM). Furthermore, Li et al. (2018) refined this detection by distinguishing between complete, partial, and non-blinking states, leading to superior performance on specialized forensic datasets.

Biological signals serve as powerful indicators for deepfake detection because they capture physiological nuances that traditional spatial-based forensic methods often overlook. For instance, Wu et al. (2020) utilized a CNN-RNN architecture to monitor temporal eye movements and blinking patterns, achieving high sensitivity by focusing on the transition between open and closed eye states. Beyond ocular cues, researchers have leveraged cardiac activity. Li et al. (2018) explored a method utilizing Photoplethysmography (PPG) to detect facial color dynamics synchronized with the heartbeat. This model, which identifies subtle "blood flow" signals, attained over 97% accuracy across multiple datasets. Furthermore, recent advancements have shifted toward multimodal analysis. Lima et al. (2020) developed a framework that evaluates the synchronicity between audio and visual streams. By extracting parallel representations and employing a triplet loss function, their model identifies discrepancies between speech and lip movement. This multimodal approach demonstrated high reliability, reaching an average accuracy of 96.6% on the DeepfakeTIMIT dataset and 84.4% on the more challenging DFDC dataset.

Beyond single-frame analysis, examining temporal inconsistencies across video sequences has proven to be a powerful strategy for improving deepfake detection. Li and Lyu (2018) introduced a temporal feature analysis method that employs a convolutional Long Short-Term Memory (LSTM) network to capture sequential irregularities in videos, enabling more effective identification of manipulated content. Building on this concept, the SSTNet framework (Afchar et al., 2018) integrates spatial, temporal, and steganalysis features within an LSTM-based architecture, demonstrating robust performance on benchmark datasets such as FaceForensics++ (Tiwari, 2024). The importance of temporal analysis is further underscored by the limitations of image-based approaches when applied to video. Video compression often introduces frame-level information loss, which can degrade the performance of methods designed for static images (Fogelton & Benesova, 2018). While many detection systems focus on individual frames, integrating sequential analysis significantly enhances their ability to identify subtle manipulations that unfold over time. Graves and Schmidhuber (2005) proposed a hybrid model that first extracts frame-level features using a CNN and then processes the resulting feature sequences through an LSTM network. Evaluated on a dataset of 600 videos, this approach demonstrated strong detection performance, highlighting the value of frame-sequence modeling. Zhu et al. (2017) extended the Cycle-GAN framework to develop Recycle-GAN, which incorporates both spatial and temporal constraints to improve detection outcomes. More recently, Kietzmann et al. (2020) proposed a two-stage pipeline: the first stage performs face cropping and alignment using a Spatial Transformer Network (STN), while the second stage analyzes temporal inconsistencies via recurrent convolutional networks. By combining spatial alignment with temporal modeling, this method effectively detects manipulated content across a range of deepfake videos.

Our literature review identifies several critical research gaps that impede the progress of deepfake detection and limit its real-world applicability.

Domain generalization in text-based deepfake detection remains a significant challenge. Current detection methods are predominantly evaluated within narrow contexts, trained and tested on specific domains or particular language models. This restricted scope undermines their effectiveness in real-world environments, where textual content originates from diverse, often unseen sources and exhibits considerable stylistic and structural variation. There is an urgent need for detection frameworks capable of robust generalization across domains, languages, and out-of-distribution inputs, particularly as large language models continue to proliferate.

A second major gap lies in the detection of imagery generated by evolving GAN architectures. Although considerable progress has been made in identifying GAN-generated faces and synthetic images, many existing detectors exhibit poor generalization when exposed to datasets or generative models not encountered during training. As GAN variants and hybrid generative methods rapidly advance, detection systems must evolve accordingly, not merely to recognize known artifacts, but to adapt to novel, unseen manipulation techniques. Research must therefore prioritize model robustness, domain-agnostic feature learning, and resilience against adversarial evasion.

Multimodal deepfake detection remains significantly underexplored. While unimodal approaches for detecting fake audio, images, or video have each advanced independently, the integration of multiple modalities, such as joint audio-visual analysis, has received comparatively little attention. Deepfakes often introduce subtle cross-modal inconsistencies, such as mismatched lip movements or incongruent emotional tone between speech and facial expression. Frameworks that leverage such discrepancies could substantially improve detection accuracy and robustness, yet the development of scalable, aligned multimodal architectures remains in its infancy.

Finally, the use of temporal and biological signals for video forensics requires further refinement. Features such as irregular eye blinking, inconsistent head movement, and heartbeat-derived photoplethysmographic signals offer promising, hard-to-spoof cues for deepfake detection. However, current approaches are highly sensitive to variations in video quality, compression artifacts, frame rate, and recording conditions. To transition from controlled benchmarks to practical deployment, these methods must be made more reliable and invariant to such distortions. Broader and more diverse datasets capturing naturalistic physiological and behavioral patterns are also urgently needed.

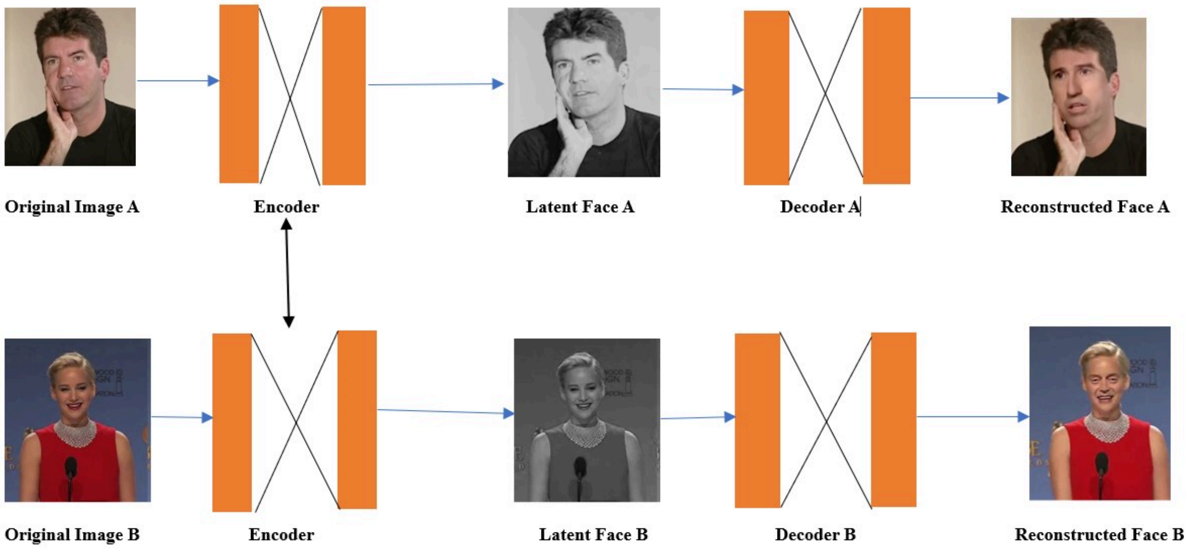

Fig. 2: Autoencoder-Based Face-Swapping Framework

Technical Aspects of Deepfake Creation

Autoencoders

Autoencoders, a specialized type of neural network, were among the first technologies used in the creation of deepfakes. They work by learning efficient representations of input data through unsupervised learning, consisting of two main components: the encoder and the decoder. The encoder compresses the input pixels into a smaller, more manageable size, encoding essential features such as skin texture, color, facial expressions, and head pose. After that, this compressed data is transmitted to the latent space where the model searches for patterns and structural correlations in the input information. In this case, the learning process is most important when the network focuses on the more important features of the input; for example, recognition of more significant facial features - though less important features are also discarded. After this compressed representation has been found, the decoder tries the best it can, to reconstruct as close to the input as it can, all while trying to minimize the differences between a compressed image and the reconstructed image. In this way, the autoencoder is trained to generate realistic outputs to become a suitable application of deepfake generation by replicating the unique features that are requirement to form a realistic fake (Nguyen et al., 2022).

Figure 2 illustrates the use of autoencoders for face-swapping, specifically demonstrating how one face can be replaced with another. In this process, both faces are reconstructed along the paths indicated by the red arrows, with Face B being transformed to resemble Face A. A distinctive feature of this architecture is that both faces are encoded by the same encoder network, enabling it to learn and represent the general facial features common to both subjects. This shared encoding ensures that the latent representations of the two faces are positioned as closely as possible within the lower-dimensional embedding space. Consequently, the decoder associated with Face B can leverage the latent representation of Face A to reconstruct Face B while incorporating the characteristic features of Face A. This mechanism forms the foundation of several prominent deepfake frameworks, including DeepFaceLab, DFaker, and TensorFlow-based implementations, demonstrating the efficacy of autoencoders for high-quality face-swapping applications (Luttrell et al., 2018).

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have emerged as the foundational technology underpinning modern deepfake generation, owing to their unique architecture comprising two competing neural networks. Introduced by Usukhbayar and Homer (2020), GANs consist of a generator that synthesizes new data samples and a discriminator that evaluates whether these samples are real or fake, as illustrated in Figure 3. This adversarial process drives both networks to iteratively improve: the generator strives to produce increasingly realistic outputs capable of fooling the discriminator, while the discriminator becomes more adept at identifying synthetic content. Over time, this competitive dynamic enables the generator to produce highly convincing fake data.

In the context of deepfakes, GANs are central to generating realistic synthetic media, including videos, images, and audio. For example, GANs can synthesize facial expressions, movements, or vocal inflections that, when incorporated into forged videos, create compelling illusions of authenticity (Mirza & Osindero, 2014). Trained on large-scale datasets, the generator learns to produce outputs that closely mimic real data, while the discriminator continuously assesses their quality. The training process reaches equilibrium when the discriminator classifies generated content as authentic approximately 50% of the time, indicating that the generator has successfully produced examples indistinguishable from genuine data.

While GANs enable remarkable applications across entertainment, education, and creative industries, they also raise profound ethical concerns. The same capabilities that make GANs powerful tools for innovation render them susceptible to misuse, including the creation of misinformation, privacy violations, and malicious impersonation.

A notable deepfake-specific architecture is VGGFace, which builds upon the standard GAN framework by incorporating two additional loss functions: adversarial loss and perceptual loss. The adversarial loss encourages the generator to produce outputs that closely resemble real data, enhancing visual realism, while the perceptual loss ensures structural and semantic alignment with the target face. These enhancements enable VGGFace to generate highly convincing face-swapped outputs, making it a prominent method in deepfake creation (Pan et al., 2020).

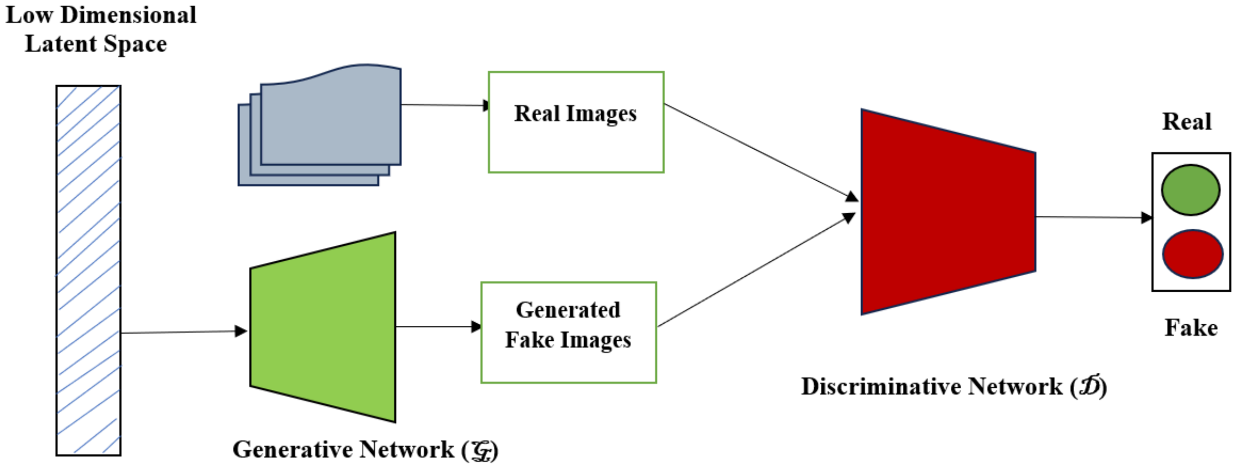

Fig. 3: Creation of Deepfakes using GAN

Neural Radiance Fields and Emerging Deepfake Generation Techniques

The recent emergence of Neural Radiance Fields (NeRFs) marks a significant advancement in generative modeling, offering a powerful approach for creating photorealistic 3D environments and facial reconstructions. Unlike GANs and autoencoders, which primarily operate on 2D image data, NeRFs employ differentiable rendering techniques to model a continuous 5D function that describes light interactions at every point in three-dimensional space. This volumetric representation enables NeRFs to generate highly consistent and lifelike images of a subject from arbitrary viewpoints, preserving accurate lighting, geometry, and depth information.

In the context of deepfake generation, NeRFs facilitate the creation of sophisticated synthetic content with enhanced realism and spatial coherence. They support dynamic manipulations such as altering facial expressions, animating digital avatars, and rendering novel views of a subject within video sequences, capabilities that exceed those of traditional 2D-based methods. Advanced variants including pi-GAN, NeRF-W, and hybrid GAN-NeRF architectures further refine output quality by integrating volumetric rendering with adversarial learning, pushing the boundaries of synthetic media realism.

However, the very characteristics that make NeRFs powerful generative tools also introduce new challenges for deepfake detection. Unlike 2D-generated forgeries, which often leave detectable artifacts such as blending inconsistencies or frequency-domain anomalies, NeRF-generated content is largely free of such telltale signs. Their outputs maintain spatial and temporal coherence across viewpoints, enabling them to evade conventional 2D-based detection models that rely on frame-level analysis.

Consequently, identifying NeRF-based deepfakes necessitates a paradigm shift in forensic methodology. Detection systems must move beyond 2D analysis and adopt approaches that operate natively in three-dimensional space, extracting multi-view features, analyzing volumetric inconsistencies, and examining cross-frame geometric coherence. The rise of NeRFs underscores a critical evolution in deepfake technology and highlights the urgent need for detection frameworks capable of addressing synthetic media beyond the two-dimensional domain.

Technical Aspects of Deepfake Detection

Deepfake detection has gained significant attention due to the rising use of Artificial Intelligence (AI) to create hyper-realistic synthetic content. Deepfakes are created using deep learning techniques, particularly generative models like Generative Adversarial Networks (GANs) and autoencoders, to manipulate text, images, videos, or audio convincingly. The detection of deepfakes requires an understanding of various technical aspects, including machine learning, signal processing, and computer vision (Bahdanau et al., 2014; Tariq et al., 2018; Wang et al., 2017). The deepfake detection pipeline, as illustrated in Fig. 4, consists of six key stages: data collection, face detection, feature extraction, feature selection, model selection, and model validation. These interconnected steps enable the systematic identification of manipulated content by progressively refining data and selecting optimal models for accurate detection.

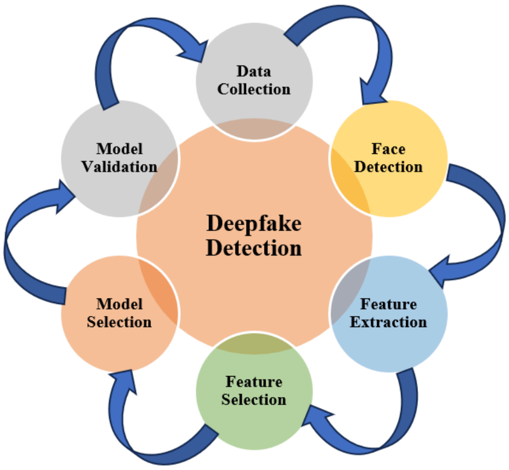

Fig. 4: Steps of Deepfake Detection on Media Files

Deep Learning

Deep learning is an approach similar to neural networks, utilizing multiple hidden units within its architecture (Rossler et al., 2019). Its structure, inspired by artificial neural networks, includes a potentially unlimited number of hidden units of fixed size, designed to extract additional information from input data. The number of hidden layers required depends on the complexity of the data being trained; more complex datasets demand deeper architectures to produce accurate results (Hopfield, 1982; Marra et al., 2019). In recent years, deep learning has been successfully applied across numerous fields and is expected to remain a vital tool in advancing various technologies.

Convolutional Neural Network (CNN)

CNN is a widely used deep neural network architecture consisting of an input layer, an output layer, and one or more hidden layers, similar to other neural networks. In CNNs, the hidden layers process input data from the first layer by performing convolution operations, which involve applying mathematical filters to extract important features from the data (Schuster & Paliwal, 1997). Another important thing regarding CNNs is that they also include matrix multiplication, whereas non-linear activation functions are ReLU and others including pooling layers. Local connections, for example, average pooling, diminish the intricacy of the data by providing a representation of the features, which makes the work of the network lighter and more efficient (Wang & Dantcheva, 2020).

In the context of deepfakes, CNNs are employed to identify subtle artifacts such as blending inconsistencies, unnatural textures, or warping at facial boundaries, signatures that often arise during synthetic media generation. Models like XceptionNet and EfficientNet have shown high accuracy on deepfake datasets like FaceForensics++, demonstrating CNNs' effectiveness in capturing pixel-level anomalies. The architecture of the CNN network that is used for detecting deepfakes is shown in Fig. 5. Here, multiple convolutional and pooling layers are used to find out the structure of images, and afterwards, fully connected layers sort the information as either real or false. It depends on an activation function called Leaky ReLU to add some curves to the model's responses and improve its detection of minor details in changed content.

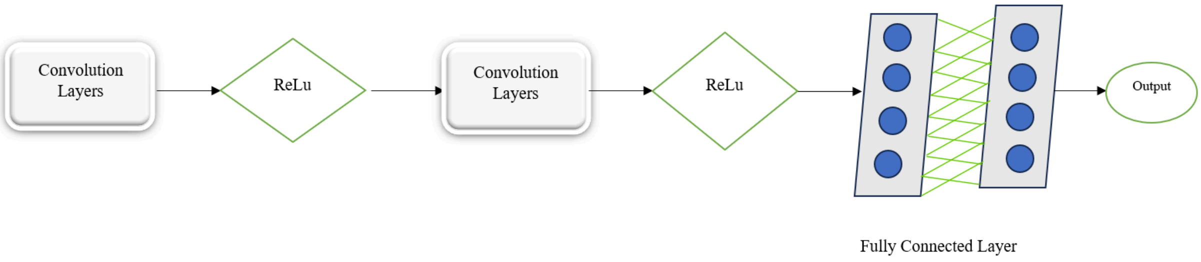

Fig. 5: CNN Architecture

Recurrent Neural Network (RNN)

RNN is an artificial neural network designed to identify patterns from data that is sequential or temporal. It is made of many concealed layers each of which contains its specific bias and weight. The two most important characteristics of RNNs relate to the connections between nodes where there is a direct loop since the input and output of the information being passed are cyclic (Karras et al., 2017). This architecture has a recurrent hidden state a capability that allows RNNs to process from and learn from temporal sequences as or time series and language data.

In deepfake detection, RNNs help identify inconsistencies in facial expressions or blinking patterns that may not align naturally over time. These inconsistencies are often introduced during frame-by-frame manipulation in deepfakes. RNNs can detect unnatural head movements or speech-lip sync mismatches that flag temporal forgery. In RNNs, as shown in Fig. 6, sequential processing occurs along with layers for storing the relationships among frames or segments of the audio data. Because the hidden layers hold intermediate outputs, the model spots changes that seem inconsistent, for example, unrealistic blinking in a video or lip movements that don't relate to the audio.

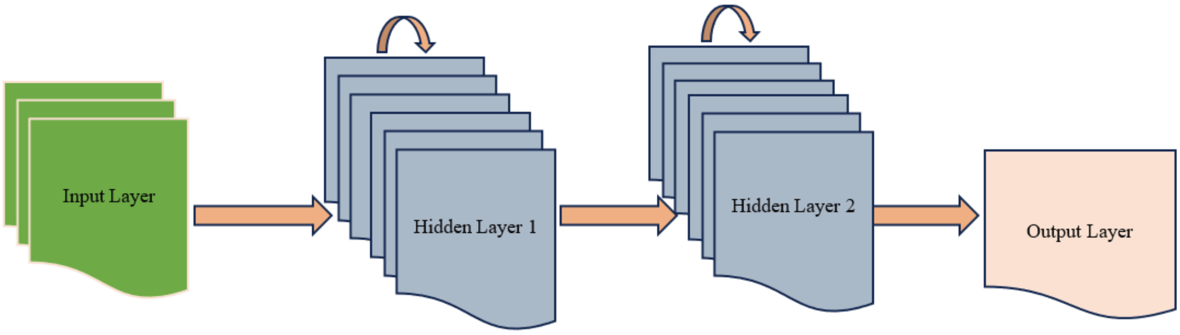

Fig. 6: RNN Architecture

Long Short-Term Memory (LSTM)

LSTM is a particular kind of Recurrent Neural Network (RNN) that's specifically built for long-term dependency problems. In contrast with normal RNNs, LSTMs provide feedback connections to learn from the whole sequence. The core architecture of LSTM consists of three gates: These are normally referred to as the input gate, the forget gate, and the output gate. From the current and the previous states, the cell state retains information from previous time steps. The input gate controls which values will be written to the cell state while the forget gate uses a sigmoid function to decide which parts of that state need to be forgotten. The output gate provides a control signal of what information from the current time step should be passed to the next stage. This structure enables LSTMs to control long-term dependency and pass important information for a very long time (Khatri & Gupta, 2023).

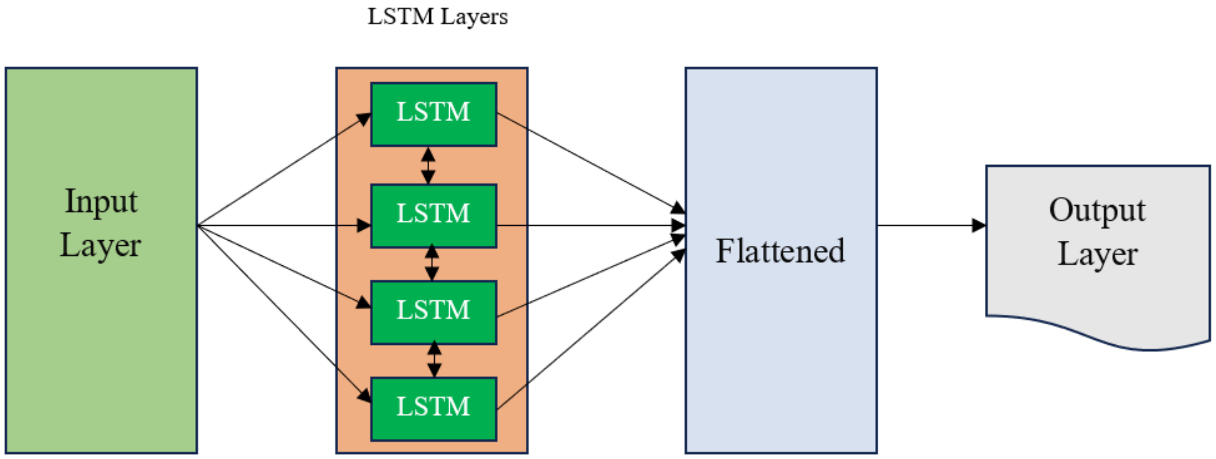

Fig. 7: LSTM Architecture

In deepfake detection, LSTMs are utilized to model longer sequences of video or audio data, enabling detection of contextual abnormalities. For example, LSTMs have been applied to detect mismatched emotional tones between facial expressions and vocal intonation in deepfake videos, an area where simple frame-level detectors may fall short. Fig. 7 represents the LSTM architecture, in which there are memory cells with gating systems and it works with flattened inputs to trace long-term patterns over time. When changing sequential data into flattened vectors, the model understands contextual features such as mismatch in emotions or the way people speak, improving the ability to spot deepfakes.

Traditional Feature-Based Methods

In a similar spirit to the traditional paradigm of feature engineering, traditional methods of deepfake detection aim to detect indications of manipulation through the utilisation of handcrafted features and statistical natures over architectures that are fundamentally neural (Baltrušaitis et al., 2019). These methods use knowledge of the domain to reveal discrepancies and features that were incorporated into the generation process. Major techniques comprise frequency domain analysis, biological signal measurement, head pose and Landmark measurement, and synchronization of Audiovisual sound and vision.

Frequency Domain Analysis

In frequency domain analysis inputs are looking for irregular patterns in the frequency elements of pictures and motion pictures. Using Fourier Transform and wavelet analysis, and the like, it points out compression artifacts, texture inhomogeneities or any other spatial discrepancies originated in the synthesis procedure. Such methods are the most efficient for the detection of distortions that are not educed in the spatial domain (Nguyen et al., 2022; Mittal et al., 2020).

Biological Signal Analysis

Biological signal analysis takes advantage of signals that are somewhat difficult to mimic by deepfake algorithms. For instance, normal eye blinking patterns can hardly be observed or are random in fake videos because generative models hardly learn (Shan et al., 2007). Likewise, the PPG signals that record the variation in the skin color due to the blood flow resulting from heartbeat extensively distort deep fakes. These cues present a noninvasive approach for monitoring the irregularity of activity in the facial structures (Ciftci et al., 2024).

Head Pose and Facial Landmark Analysis

Head pose and facial landmark analysis focus on detecting inconsistencies in the spatial configuration and movement of key facial features, offering a valuable cue for identifying manipulated content (Fogelton & Benesova, 2018). By tracking the position, angle, and trajectory of head movements, these methods can reveal unnatural patterns often present in deepfake videos, such as abrupt changes in orientation, jerky motion, or physically implausible rotations. Additionally, anomalies in eye gaze direction or asymmetries in facial features, such as misaligned or poorly rendered side profiles, are common artifacts in tampered images. These irregularities arise because generative models often struggle to maintain coherent three-dimensional geometry and temporal continuity across frames. Thus, analyzing head pose dynamics and landmark consistency provides a robust, complementary approach to conventional artifact-based detection methods.

Audio-Nisual Synchronization Analysis

Audio-visual synchronization analysis automatically assesses the synchronization between a speaker's voice and lips visible on screen. Some of the common problems identified when creating deepfake videos include temporal synchronization issues identified as temporal discrepancy or wrong pronunciation of words which is articulated as temporal incoherence (Bengesi et al., 2024). This technique is most effective when identifying fake videos where speech is not synchronized with the movements of the mouth and jaw.

Hybrid Approaches

Hybrid approaches that integrate deep learning with traditional feature-based methods have demonstrated significant improvements in the accuracy and robustness of deepfake detection. These solutions leverage the complementary strengths of both paradigms: deep neural networks excel at learning hierarchical representations and identifying complex patterns directly from data, while feature-based techniques contribute domain-specific forensic cues that target known artifacts of synthetic media.

In hybrid architectures, Convolutional Neural Networks (CNNs) are commonly employed for spatial analysis, enabling the extraction of multi-scale feature pyramids from image and video data (Bao et al., 2023). Trained on large datasets, these networks learn to detect subtle distortions in facial regions, irregular textures, and generative artifacts that may elude hand-crafted feature extractors. Moreover, CNNs are capable of identifying non-linear manipulation indicators, such as inconsistent gaze direction, unnatural illumination patterns, or implausible head poses, that are difficult to encode through explicit rules (Graves & Schmidhuber, 2005). By combining the representational power of deep learning with the interpretability and specificity of forensic features, hybrid systems offer a more resilient and generalizable framework for detecting increasingly sophisticated deepfakes.

Recurrent Neural Networks (RNNs), particularly advanced variants such as Long Short-Term Memory (LSTM) networks, play a critical role in analyzing temporal consistency for deepfake detection. By design, RNNs are inherently well-suited for sequential data, making them highly effective for video analysis where the temporal relationships between frames carry essential forensic information (Chollet, 2017). These networks can track how facial shape, motion, and spatial positioning evolve over time, revealing artifacts such as unnatural transitions, jitter, or mismatched facial landmarks that betray the presence of manipulation.

When combined, CNNs for spatial feature extraction and RNNs for temporal modeling form a robust foundation for deepfake detection frameworks. This hybrid architecture enables the system to assess video content not only at the level of individual frames but also across the motion dynamics that connect them. As a result, the model becomes better equipped to identify even intricately manipulated videos where per-frame analysis alone may fall short (Al-Dhabi & Zhang, 2021).

Hybrid models offer significant practical advantages due to their flexibility in integrating conventional feature-based methods during preprocessing or feature extraction stages. Techniques such as frequency domain analysis, biological signal detection, and audio-visual synchronization can be applied upstream to extract forensic cues, which are then fed into deep learning architectures for further refinement. This layered approach helps guide the learning process by focusing the model on features known to be indicative of manipulation, thereby enhancing both efficiency and detection accuracy (Dehghani & Saberi, 2025).

By combining hand-crafted features with the representational power of deep neural networks, hybrid systems address a wide spectrum of deepfake generation techniques more effectively than either paradigm alone. They consistently demonstrate higher detection rates across diverse scenarios, outperforming purely feature-based or purely deep learning-based counterparts. This superiority stems from their ability to capture both explicit forensic traces and subtle, learned patterns of manipulation.

Such approaches are particularly valuable in real-world settings, where deepfakes are often high-quality and exhibit only subtle inconsistencies across spatial and temporal dimensions. Hybrid frameworks, by design, are better equipped to detect these nuanced artifacts, offering improved robustness and generalization in the face of evolving generative methods.

Advances in Forensics

Emerging frontiers in deepfake forensics are characterized by the integration of innovative technologies aimed at developing more robust, scalable, and resilient countermeasures against synthetic media manipulation. Notable among these advances are blockchain-based systems for content provenance and video verification, which provide immutable records of media authenticity, and adversarial training techniques designed to proactively identify and adapt to novel deepfake generation tactics. Together, these approaches seek to fortify forensic pipelines against increasingly sophisticated forgeries, enhancing both detection reliability and operational scalability in real-world contexts.

Blockchain Technology in Deepfake Detection

Blockchain helps address the issue of deepfake detection since it improves the proof and origin of content. Blockchain secures the content origin and makes it resistant to change by saving digital hashes and historical data on a secure ledger, which can be accessed anytime (Chang et al., 2020). Such a single record allows anyone to find any unreal or altered information by comparing it with the authentic one included in the end-of-year report. Blockchain works alongside AI-based systems by providing the reliability and accountability of media items used on different platforms. Still, there are several technical problems that hold it back, such as making the network big enough, keeping metadata private, and hooking up in real-time with current detection systems. Therefore, current systems are becoming stronger by combining blockchain with machine learning, so they can identify and verify the authenticity of data in an efficient and decentralized way (Jbara & Soud, 2024).

Adversarial Training to Preemptively Detect New Deepfake Techniques

As deepfake technology advances, adversarial training has been developed to target brand new forms of deepfakes proactively. Adversarial training is simply the deliberate training of deep learning capability with the exact intention of identifying and avoiding adversarial attacks, these are slight modifications that are introduced to fool machine learning systems. When it comes to deepfake detection, adversarial training is employed to help the model be safe from new deepfake-generating techniques that may not have been used previously in the creation of the training set (Black et al., 2022). This approach operates by training deepfake detection models with synthetic data created by advancing deepfake approaches, a fresh manipulation technique, or even other kinds of video synthesis. The model is then used to detect such new techniques by training the distinction between real and fake news while fakes become even more real or complex. However, this boosts efficiency during the evaluation of unknown or new categories of deep fake since, during the training of the model, various adversarial instances are introduced to the model.

The impact of these advancements, both blockchains for video authentication and adversarial training, are making way for authentic deepfake detection. Blockchain offers means for maintaining enduring protection of digital media, and adversarial training assists detection models in staying cautious of new and sophisticated deepfakes (Chen et al., 2021). Combined, all these developments improve the capacity of forensic solutions to preserve the integrity of the video material and offset the threats inherent in ever-more elaborate deepfakes. They are critical in fields such as media, law enforcement, and security, where the trustworthiness of video evidence is paramount. Table 2 explains the comparative analysis of Deepfake Detection methods and their performances.

Table 2: Comparative Analysis of Deepfake Detection Methods

Method Type	Representative Models	Detection Domain	Dataset(s)	Reported Accuracy	Strengths	Limitations
CNN-based	XceptionNet, MesoNet	Spatial (image/video frames)	FaceForensics++, DeepfakeTIMIT	95-99%	Good feature extraction; widely tested	Sensitive to compression; may overfit
RNN-based	LSTM-CNN hybrids	Temporal (video sequences)	FaceForensics++	~93%	Captures temporal inconsistencies	Computationally expensive
Transformer-based	ViT, Swin Transformer	Spatial + Temporal	Celeb-DF v2, DFDC	95-97%	Global and contextual info	Needs large data; longer training
Frequency-domain	FFT+CNN, Steganalysis	Frequency	FaceForensics++, Celeb-DF v2	90-96%	Robust to visual post-processing	Sensitive to image resolution
Biological signal-based	DeepRhythm, EyeBlinkNet	Physiological signals	Custom + FaceForensics++	85-92%	Exploits natural human cues	Dataset limitations; user-dependent
Multi-modal approaches	Audio-visual models, lip-sync	Visual + Audio	DFDC, TIMIT	88-94%	Effective for talking head videos	Difficult to generalize

Applications of Deepfakes

Media and Entertainment

Deepfake technology has profoundly influenced the media and entertainment industries by enabling innovative storytelling and enhanced audience engagement (Bengio et al., 1994). One major application is in enhanced visual effects and CGI, where deepfakes complement traditional CGI by creating more realistic facial animations and character renderings. This allows filmmakers to map actors' facial expressions onto digital characters, preserving emotional nuances and fostering stronger viewer connections in films and video games (Hopfield, 1982). Another important use is digital de-aging and actor resurrection, where extensive archival footage is analyzed to create convincing younger versions of actors or digitally revive deceased performers. This opens new creative possibilities but also raises ethical questions regarding consent and the legacy of actors.

Social Media and Misinformation

Deepfake technology poses profound challenges to social media integrity and public trust, particularly through the creation of manipulated videos aimed at political and social manipulation. By superimposing the faces of public figures onto fabricated footage, malicious actors can influence public opinion, incite discord, and undermine democratic processes. The 2019 alteration of a video featuring Nancy Pelosi demonstrated how even relatively simple manipulations can mislead viewers and amplify false political narratives. Similarly, the release of a deepfake video depicting Mark Zuckerberg sparked widespread debate on the ethical boundaries of synthetic media and the responsibilities of platforms hosting such content. Beyond the political sphere, deepfakes are increasingly weaponized to exploit public trust for malicious purposes, including identity theft, phishing scams, and the dissemination of divisive falsehoods intended to inflame social tensions. A particularly alarming trend is the rise of non-consensual deepfake pornography, which constitutes a grave violation of individual privacy and has prompted urgent calls for comprehensive legal protections. In the commercial domain, the posthumous use of deepfake technology in advertising, exemplified by the unauthorized digital recreation of Arnold Schwarzenegger, has ignited discussions about dignity, consent, and the respectful treatment of individuals after death (Krishna et al., 2024; Zhang et al., 2023). Collectively, these developments underscore the urgent need for enhanced media literacy, the establishment of robust ethical guidelines, and the continued advancement of deepfake detection technologies to safeguard information integrity and protect individuals from harm.

Education and Training

One of the most promising short-term impacts of deepfake technology lies in its application to education and training. By enabling highly realistic and interactive simulations, deepfakes can significantly enhance learning experiences across a wide range of professional domains, including medical, military, and legal education. Trainees can engage with lifelike scenarios, practice complex procedures, and develop critical decision-making skills in controlled, risk-free environments. This capacity for immersive simulation not only improves knowledge retention and skill acquisition but also offers scalable, cost-effective alternatives to traditional training methods, positioning deepfakes as a transformative tool for pedagogical and professional development.

Medical Education

In terms of medical learning, deepfake technology has a progressive solution through developing artificial and realistic visual scenarios where students and medical practitioners perform various operations and make decisions without harming an actual patient. For example, deepfake can synthesize patient models that realistically simulate wide spectrums of medical illnesses and facial and voice responses in the simulations. This makes it possible for trainees to communicate with such avatars actively, reacting to particular simulation cases in real time, thus, simulating actual patient meetings (Yu et al., 2018). When identifying and managing such virtual clients, medical students sharpen their clinical competencies as well as the communication skills that are critical determinants of patient advocacy. The high-fidelity of these simulations is particularly useful in training hands-on tasks such as compassionate communication skills which are part of the curriculum. They are able, for example, to rehearse how to convey bad news, how to steer a conversation, how to form relationships with patients, and so on, while a trainer can correct trainees in case something goes wrong, although no one's life would be at risk. Apart from that, it creates a better and more empathetic healthcare system since the application of this new deep-fake technology prepares students for the real world before they have to interact with real patients (Rombach et al., 2022).

Military Training

The military stands to benefit significantly from deepfake technology, particularly through the development of highly realistic, immersive simulations for training personnel in combat scenarios. By leveraging deepfake-generated environments, training programs can recreate dynamic situations, such as firefights, airstrikes, ambushes, and sabotage missions, allowing soldiers to practice tactical responses without exposure to the physical dangers of live exercises. These simulations can incorporate deepfake-generated avatars that function as adversaries or allied forces, adapting their behavior in real time based on trainee actions to provide a diverse range of combat experiences (Kirchenbauer et al., 2023). For instance, soldiers can rehearse maneuvers such as navigating complex terrain, coordinating team movements, or responding to sudden ambushes within a controlled yet highly realistic virtual setting. This approach enhances operational readiness by exposing personnel to unpredictable, combat-like situations that develop adaptive decision-making and flexibility (Wang et al., 2023). Moreover, the ability to replay and review training sessions after completion allows for detailed debriefing and performance analysis, fostering improvements in both individual judgment and team coordination. By offering repeatable, scalable, and risk-free exposure to a wide spectrum of tactical challenges, deepfake-based simulations provide military personnel with experiential learning that surpasses traditional mock exercises, ultimately strengthening preparedness for the complexities and uncertainties of real-world operations.

Legal Education

Deepfake technology opens new and innovative possibilities for legal education, particularly through its application in mock trials and courtroom simulations. By generating realistic, AI-driven avatars of witnesses, defendants, or jurors, law students can engage in immersive exercises that closely replicate the complexities of actual court proceedings (Ciftci et al., 2020). These simulations provide a dynamic environment in which students can practice essential professional skills, including cross-examination, witness questioning, argument construction, and courtroom etiquette, within a controlled, risk-free setting. The interactivity of deepfake-generated avatars allows students to respond in real time to virtual testimony, adapting their strategies based on the immediate reactions of simulated participants. This real-time engagement sharpens critical thinking and fosters the ability to pivot effectively during trial advocacy (Ramluckan, 2024). Furthermore, students can experiment with different legal approaches and observe the consequences of their choices, gaining insight into both strategic decision-making and the ethical dilemmas that may arise in practice, all without real-world repercussions. Through iterative practice and feedback, students deepen their understanding of legal procedures, courtroom dynamics, and effective communication with judges and jurors. These experiences not only build confidence but also better prepare future lawyers for the demands of actual legal practice. As such, the integration of deepfake technology into legal pedagogy represents a significant advance in experiential learning, equipping students with the practical competencies essential for professional success (Meskys et al., 2020).

Marketing and Advertising

Deepfake technology is rapidly emerging as a transformative force in marketing and advertising, offering brands the ability to launch highly targeted campaigns through the use of virtual influencers and AI-generated avatars. By creating realistic digital representatives that embody brand values and messaging, companies can engage consumers in novel and personalized ways. These avatars can be tailored to resonate with specific demographic segments, appearing in advertisements, product demonstrations, or interactive video content that aligns with the interests and preferences of target audiences. This level of personalization significantly enhances viewer engagement and fosters brand loyalty, as consumers are more likely to connect with content that feels customized to their needs. Moreover, because deepfake avatars are fully animated and interactive, users can engage in real-time dialogue, asking questions or seeking product recommendations, thereby increasing the shareability of content and providing brands with valuable insights into customer sentiment and behavior. However, the adoption of deepfake technology in marketing also raises important ethical considerations. The use of hyper-realistic avatars blurs the line between authentic and synthetic representation, necessitating clear disclosure to prevent consumer deception. Brands must navigate this terrain carefully, ensuring that audiences are aware when they are interacting with AI-generated rather than human figures (Dvoskin, 2022). Transparency is essential to maintaining trust and avoiding the erosion of credibility that could result from undisclosed synthetic media. While deepfake technology offers unprecedented opportunities for interactive, personalized marketing, its deployment must be preceded by thorough ethical scrutiny. Balancing innovation with honesty will be critical to harnessing its potential without compromising the integrity of consumer relationships.

Security and Surveillance

Deepfake technology is increasingly being explored for applications in security and surveillance, particularly within the domains of defense, counterintelligence, and policing. Proponents argue that deepfake-generated simulations can be leveraged to prepare armed forces personnel for a range of threat scenarios, enabling them to rehearse responses to adversarial tactics without engaging in actual conflict (Bethu et al., 2024). These simulations offer a safe yet realistic environment for training, allowing military units to refine strategies and improve operational readiness. In counterintelligence, deepfake technology presents novel opportunities for developing decoys or disseminating disinformation to mislead potential adversaries, thereby protecting sensitive information and disrupting hostile activities. Law enforcement agencies may also benefit from deepfake applications, such as reconstructing crime scenes, generating synthetic eyewitness accounts for investigative purposes, or creating simulated environments to aid in clue identification and case reconstruction. While these applications illustrate the innovative potential of deepfakes to enhance security and surveillance capabilities, they also introduce significant challenges. Chief among these is the risk of misuse, the same tools that enable beneficial simulations can be exploited for fraudulent or malicious purposes. This dual-use nature underscores the urgent need for robust risk management strategies, clear governance frameworks, and safeguards to prevent the abuse of deepfake technologies in security contexts. In sum, the integration of deepfakes into security and surveillance offers promising avenues for training, deception, and investigation. However, realizing these benefits requires careful consideration of the associated risks and a commitment to ethical oversight to ensure that such powerful tools are used responsibly (Lundberg & Mozelius, 2024).

Datasets and Evaluation Matrices of Deepfakes

Available Datasets

Face Forensics ++

FaceForensics++ is a widely used benchmark dataset for detecting manipulated facial images, comprising over 1,000 original YouTube videos altered using multiple face manipulation techniques. Each video in the dataset is accompanied by ground-truth masks and pixel-level annotations indicating tampered regions, enabling high-quality training and evaluation of deepfake detection models. Its comprehensive design and realistic manipulations have made it a foundational resource for developing and benchmarking forensic methods targeting fraudulent face imagery.

Eye-Blink

The Eye-Blink dataset is a collection of 80 videos of 20 people blinking, each lasting a few seconds, and filmed under different lighting circumstances and camera positions. The Talking Face Dataset includes approximately 5,000 video clips of people speaking, each lasting 25 frames and with an image size of 720x576 pixels. The mEBAL Dataset comprises more than 340,000 eye blink photos from 3,000 people, each with a unique head posture and lighting circumstances. The UDF Dataset comprises 98 movies of genuine and artificial faces, each lasting 11 seconds.

WildDeepfake

WildDeepfake is a deepfake detection dataset that combines real and deepfake internet samples. Unlike prior datasets, which solely comprised synthesised facial pictures, this one contains a variety of body kinds. However, a larger dataset is required to construct full-body deepfakes and enhance deepfake detection algorithms.

DFDC

Facebook's DFDC dataset, a massive collection of face swap films, is the most extensive and accessible, with over 100,000 movies from 3426 paid actors of all genders, ages, and races.

Deeper Forensic-1.0

Deeper Forensic-1.0 is a large dataset, which contains 50,000 authentic and 10,000 counterfeit videos. It is an important tool for spotting deepfakes. DF-VAE, a conditional automatic encoder, produces modified videos that properly replicate real-world settings using a combination of modifications and disturbances such as compression, blurred vision, noise, and visual abnormalities.

DeepfakeTIMIT

The DeepfakeTIMIT dataset, developed by the Idiap Research Institute using an open-source GAN-based approach, contains over 1,000 video samples. It includes both real and fake videos of 16 individuals, generated using two different face-swapping models with output resolutions of 64×64 and 128×128 pixels. The fake video collection comprises 32 subjects, each contributing ten manipulated videos, providing a valuable resource for evaluating deepfake detection algorithms across varying quality levels.

UADFV

The UADFV dataset, created by the University at Albany, is designed to support deepfake detection by leveraging physiological cues such as eye blinking patterns. It comprises 49 authentic YouTube videos and 49 corresponding fake videos generated using the FakeApp smartphone application. In addition to the video content, the dataset includes 17,300 combined real and fake images for evaluation purposes. This resource is particularly valuable for researchers developing facial recognition systems and practitioners seeking robust tools for deepfake video detection. Each video sequence has a duration of approximately 11.14 seconds and a resolution of 294 × 500 pixels.

ASVspoof

The ASVspoof (Automatic Speaker Verification) dataset is meant to evaluate the vulnerability of speaker verification systems to various spoofing attacks. It consists of audio clips with mimic voices, replay attacks, or other tricks that aim to deceive the system. Datasets like these are used by researchers in the development and testing of ASV systems' capacity to enhance their security against fraud.

Wave Fake

The WaveFake dataset, is designed to support research in audio deepfake detection and promote AI safety. It comprises 104,885 synthesized audio clips, generated using multiple speech synthesis architectures. This collection provides a robust foundation for developing and evaluating detection methods targeting synthetic speech across different languages and generative models.

Forgery Net

ForgeryNet hosts a large face forgery dataset with unified annotations for both picture and video streams. It consists of four different tasks, spatial forgery localization, image forgery classification, temporal forgery localization, and finally video forgery classification. With 2.9m photos and 221k videos, it is the largest openly available dataset of deep face forgery. The dataset has been thoroughly benchmarked and analyzed for face forensics algorithms.

Celeb-Deepfake

The Celeb-DF dataset includes real and fake deepfake videos that were obtained from YouTube. It includes 590 raw videos of people of different ages, races, and genders, and 5639 corresponding deepfake videos created from the public domains of 59 famous personalities through YouTube videos. The dataset has grown from its first iteration, which included 795 deepfake videos.

LAV-DF

The Localized Audio Visual Deepfake dataset (LAV-DF) is a huge-scale public dataset for temporal forgery detection and localization. It contains both real and fake audio-visual content.

Table 3: List of available Datasets

Dataset Name	Modality	Strengths	Weaknesses	Best Use Cases
FaceForensic++	Video/Image	High-quality face manipulations; well-annotated	Mostly frontal faces; lacks diversity	General-purpose detection
Celeb Deepfake	Video/Image	Diverse celebrity faces	Biased towards celebrities	Benchmarking
DFDC	Video	Large-scale, diverse, includes audio	Heavily compressed; labeling complex	Real-world simulation
DeeperForensics-1.0	Image	Robustness testing; synthetic perturbations	Controlled settings	Robustness evaluation
Deepfake-TIMIT	Video	Includes audio; lip-sync analysis	Limited size	Audio-visual fusion
UADFV	Video	Early benchmark	Low resolution; few identities	Proof-of-concept
WildDeepfake	Text	In-the-wild conditions	Small scale	Domain adaptation
Wave Fake	Image	Synthetic audio detection	Narrow scope	Audio-only spoof detection
ASVspoof	Audio	Large-scale voice spoofing	No video	Speaker verification
Eye-Blink	Video/Image	Physiological cues	Low sample size	Liveness-based detection
ForgeryNet	Audio/Video	Broad manipulation techniques	Limited use	Technique-specific testing
LAV-DF	Video/Audio	Language-specific; linguistic cues	Not widely used	Multilingual detection

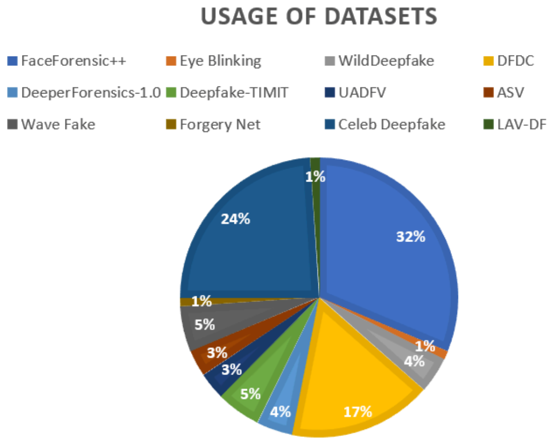

Fig. 8: Usage of Deepfake Datasets

Evaluation Matrices

This section outlines key assessment measures for deepfake detection algorithms in various media formats like text, images, videos, and audio. It stresses the importance of the complex assessment approach, leaving aside not only the global measures but the specific measures associated with the media as well. Evaluating the last aspects of the algorithm's capability for the identification of fake content, we consider such factors as accuracy, rate, and dependability. Key metrics used to assess the performance and reliability of detection algorithms include.

Accuracy

Accuracy corresponds to the average percentage of the correct detection of an object or a face. It was defined as the fraction of instances that were classified correctly, including true positives and true negatives on the total shape.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Recall

Recall measures the number of truly positive items identified as such and is also called true positive or sensitivity

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

Precision

Precision (also known as positive predictive value) measures the proportion of correctly identified positive instances out of the total instances identified as positive.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when the class distribution is imbalanced.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

False Acceptance Rate (FAR)

The False Acceptance Rate (FAR) quantifies the proportion of fake or unauthorized instances that are incorrectly accepted as genuine by a detection or authentication system. It is also commonly referred to as the false positive rate. A high FAR indicates a system's vulnerability to security breaches, as it measures how often impostor attempts are mistakenly granted access.

$$FAR = \frac{FP}{TP+TN} \tag{5}$$

False Rejection Rate (FRR)

FRR measures the percentage of genuine instances that are incorrectly rejected by the detection system. It is also known as the false negative rate.

$$FRR = \frac{FN}{TP+FN} \tag{6}$$

Equal Error Rate (ERR)

The Equal Error Rate (EER) is a performance metric commonly used in biometric verification systems to provide a single summary measure of accuracy. It represents the point at which the False Acceptance Rate (FAR), the proportion of unauthorized users incorrectly accepted, and the False Rejection Rate (FRR), the proportion of authorized users incorrectly denied, are equal. A lower EER indicates higher overall system accuracy and better balance between security and usability.

$$ERR = \frac{FAR+FRR}{2} \tag{7}$$

Error Rate (ER)

The error rate is the total volume of false-negative and false-positive detections in the total number of decisions made.

$$ER = \frac{FP+FN}{TP+TN+FP+FN} \tag{8}$$

Mean Absolute Error (MAE)

MAE is a measure of the average absolute deviations obtained when estimating the errors without reference to their sign. They are the mean of the absolute difference between the prediction and actual observation made over the test sample with each of the individual differences being given an equal importance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

True Class Score (TCS)

TCS is less commonly used but can refer to the proportion of instances correctly classified within a particular class. It is similar to class-specific accuracy.

$$TCS = 1 - \frac{1}{N-1}\sum_{i=1}^{N-1} |C_{i+1} - C_i| \tag{10}$$

These evaluation metrics provide a comprehensive toolkit for assessing the performance and reliability of deepfake detection systems. Each metric captures a distinct aspect of model behavior, ranging from overall accuracy and error rates to more nuanced measures such as precision, recall, and the trade-off between false positives and false negatives. A thorough understanding of these metrics is essential for developing robust detection methods capable of countering the growing threat of realistic synthetic media.

Ethical, Legal, and Societal Implications

Ethical Concerns

Privacy Violations and Consent Issues

Deepfakes raise one of the most urgent ethical questions regarding the violation of personal privacy and the contempt for informed consent. Deepfake technology allows one to use a person's likeness, face-wise, voice-wise, or manneristically, without their knowledge or permission. Given the explosion of non-consensual deepfake pornography, which frequently targets celebrities or people without public profiles, this is particularly troublesome. Significant psychological trauma, reputational harm, and long-term mental health effects including anxiety, depression, and social disengagement can all follow from such exploitation (Ray, 2021).

Real-world examples of the emotional toll of having one's identity altered for public consumption include the ongoing victimizing of female celebrities via AI-generated pornographic deepfakes. Due to antiquated definitions of privacy and consent, victims frequently report feelings of helplessness and shame combined with legal systems that slow down addressing of these offenses. Deepfake victims have little recourse in countries where digital likeness rights are not fully protected, so highlighting a moral and legal void. Moreover, deepfakes blur the boundaries between real and created identities, so undermining confidence in digital communication and posing existential concerns about autonomy over one's digital self (Nnamdi et al., 2023).

Manipulation of Personal Data

Deepfakes generate major questions about the manipulation of personal data outside of illegal image use. More advanced generative models can copy not only appearance but also voice patterns, behavioral signals, and emotional expressions. This realism has made new kinds of identity theft possible, including voice phishing campaigns or impersonation in video conferences whereby victims unwittingly trust and interact with synthetic versions of people they know (Durall et al., 2020). For instance, there have been documented cases where criminals used AI-generated voices to fool business staff members into sending money under the impression they were speaking to their CEOs. These clever frauds take advantage of digital confidence and expose how deeply fakes could compromise organisational integrity as well as personal relationships.

Deepfake creators also frequently obtain training data from public venues such as social media, gathering enormous volumes of personal images and videos without permission. Metadata and unique identifiers can still be obtained even when people try to reduce their digital footprints, so posing questions about surveillance and exploitation. For instance, there have been documented cases where individuals have experienced profound psychological distress upon discovering that their appearance, voice, or likeness has been appropriated and manipulated without consent. This violation underscores a growing concern: the ease with which personal identity can be digitally altered undermines an individual's sense of autonomy and security in the digital age (Verma, 2023). Victims frequently report experiencing "identity dissonance", a disconnection between their true personality and the phony digital version that others see. This adds an additional layer of emotional distress that technical solutions alone cannot address, emphasizing the importance of ethical frameworks based on psychological well-being and human dignity.

Legal Framework

Legal and Regulatory Responses to Deepfakes

Currently, several legislative initiatives have emerged to address the challenges posed by deepfake technology, though these efforts remain fragmented across different jurisdictions. Various countries and regions have enacted laws targeting the malicious use of synthetic media in contexts involving consent, electoral integrity, and defamation. For instance, California and Texas have introduced legislation specifically prohibiting the use of deepfakes in sexually explicit content and political campaigns, particularly during election periods. At the federal level, the proposed Deepfakes Accountability Act is pending approval, aiming to establish a national framework for regulating manipulated video content (Brundage et al., 2018). The bill would impose penalties on individuals who create or distribute deceptive deepfakes with the intent to undermine public trust, influence voting behavior, or manipulate perceptions of individuals in electoral and public affairs contexts.

At the international level, the adoption of the European Union's AI Act in 2024 represents a significant milestone in the governance of deepfake technology. Classifying deepfakes as high-risk applications, the legislation mandates that systems be transparent, accountable, and subject to human oversight. Under this framework, individuals who generate or disseminate synthetic media are required to clearly disclose its artificial origin. Complementing the AI Act, the EU's General Data Protection Regulation (GDPR) provides legal recourse against the unauthorized use of an individual's likeness and grants individuals greater control over their biometric data (Shan et al., 2007).

Meanwhile, China has implemented one of the world's most stringent regulatory regimes for deepfakes, requiring that all AI-generated content, particularly synthetic videos and images, bear clear and visible labels indicating their non-authentic nature. These national and regional initiatives represent critical steps toward addressing the ethical and security challenges posed by deepfakes.

Despite these efforts, a cohesive international legal framework governing deepfakes has yet to emerge. Many countries still lack comprehensive regulations, leaving a patchwork of rules that complicates enforcement and fails to keep pace with the rapid evolution of the technology. Striking an appropriate balance between fostering innovation and ensuring accountability remains a significant challenge (Chesney & Citron, 2018). Addressing the global risks of deepfakes while preserving the beneficial applications of artificial intelligence will require coordinated international cooperation and the development of shared standards.

Landmark Legal Cases And Their Legislative Impact

Several notable legal cases have begun to shape the evolving legal landscape surrounding deepfake technology. A prominent example is the case involving actress Scarlett Johansson, whose face was superimposed onto another individual's body in a pornographic video without her consent. Although Johansson ultimately chose not to pursue legal action, citing the difficulty of identifying and prosecuting the anonymous creators, the incident starkly illustrated the challenges of applying traditional defamation and privacy laws to digitally manipulated content (Nnamdi et al., 2023). The case underscored the inadequacy of existing legal frameworks in addressing the harms caused by deepfakes and highlighted the urgent need for enhanced penalties targeting malicious synthetic media that threaten individual identity and personal security.

A similar incident occurred in the United States in 2020, when a mother from Pennsylvania was charged with creating manipulated images of her daughter's cheerleading competitors, depicting them in compromising positions in an attempt to secure their removal from the team. This case established an important legal precedent by demonstrating that criminal charges could be pursued for the distribution of deepfakes, particularly in contexts involving harassment and defamation. It also illustrated how deepfake technology can be weaponized in interpersonal conflicts, extending its legal and societal implications beyond high-profile celebrities and political figures to ordinary individuals.

Together with emerging legislative efforts, these cases underscore the critical need for future legal frameworks that clearly define deepfake-related offenses, establish proportionate penalties, and provide accessible avenues for victims to seek recourse. As synthetic media technologies continue to evolve, lawmakers will face mounting pressure to enact comprehensive regulations that address the multifaceted harms of deepfakes, from personal privacy violations to the broader erosion of democratic discourse through misinformation (Verma, 2023).

Impact on Public Trust

Deepfake content has become a significant source of public distrust, fueling widespread anxiety over the authenticity of digital media (Schiff et al. 2023). As synthetic manipulations grow increasingly difficult to distinguish from genuine content, individuals and institutions alike face mounting uncertainty about the veracity of news and information. The following areas highlight key public concerns regarding the proliferation of deepfake technology.

Influence on Political Discourse

Deepfakes can be used to create misleading or fabricated content that appears to come from credible sources, potentially distorting political narratives and influencing public opinion. For example, deepfake videos of politicians making controversial statements or engaging in unethical behavior can be used to manipulate voters or incite political unrest. This misuse can distort democracy by spreading and distorting the messages of politics with fake information (Tipper et al., 2024).

Impact on Journalism

In journalism, deepfakes can cause an issue with the authenticity of the news story being produced. Considering the case when news organizations tend to integrate more visual and audio content into their flows, the availability of complex deep fakes increases the risk of fakes passing through the checkpoints. This raises questions on the idea of the job of journalism in presenting good, credible information to the public and questions the trustworthiness of the media in society. (Korshunov & Marcel, 2018).

Effect on Social Media

Moreover, social media groups are susceptible to the circulation of deepfake content. In this regard, deepfakes spread the information pretty fast, and it may take some time until the fake news is detected, so deepfakes have a great potential for virality. The looming problem of deepfakes poses a significant threat of increasing misinformation and politically dividing the community while also undermining the credibility of social media.

Erosion of Trust in Visual and Audio Media

Deepfakes in turn affect the basic trust in video and audio information as such. With deepfakes getting more life-like, people may be put off by what their eyes and ears are telling them, probably a scenario where nobody is guilt-free. Such an approach can cause wider doubts about the reality of all video and audio material, not only deepfakes, which indeed raises the problem of trust in media as a whole and makes it impossible to distinguish between fake news and actual events (Korshunov & Marcel, 2018; Lima et al., 2020).

Recommended Countermeasures and Best Practices

Mitigating the multifaceted risks posed by deepfake technology requires a coordinated strategy that integrates robust policy measures, technological innovation, public education, and responsible design practices. The following countermeasures and best practices are proposed to address the growing challenge of synthetic media manipulation.

Regulatory Frameworks

Governments and regulatory bodies must enact comprehensive legislation to address the misuse of synthetic media. Effective legal frameworks should include provisions that criminalize the creation and distribution of deepfakes intended to deceive individuals or appropriate their likeness without consent. Equally important is the establishment of clear accountability mechanisms, ensuring that those whose deepfake-related actions result in fraud, reputational harm, or other damages can be held legally responsible. Transparency mandates are also essential, requiring content creators to clearly label AI-generated material to inform audiences of its synthetic origin.

Recent legislative developments, such as the European Union's AI Act and the proposed Deepfakes Accountability Act in the United States, represent important initial steps toward formalizing these principles. However, given the borderless nature of digital media, broader international cooperation is necessary to harmonize standards, facilitate cross-border enforcement, and effectively combat the global proliferation of malicious deepfakes.

Digital Watermarking and Provenance

Platforms and content creators should adopt imperceptible watermarking or cryptographic signatures to ensure content authenticity. Tools such as the C2PA standard (Coalition for Content Provenance and Authenticity) provide technical pathways to verify source integrity and detect tampering.

Public Awareness Campaigns

Public awareness campaigns are essential for mitigating the societal impact of deepfakes by equipping individuals with the knowledge and critical thinking skills needed to navigate an increasingly complex media landscape. Governments, media platforms, and non-governmental organizations can play a pivotal role by launching media literacy initiatives that serve multiple objectives. Such campaigns should aim to raise awareness about how deepfakes are created and distributed, helping the public understand the technical capabilities and limitations of synthetic media. They must also teach users how to recognize common signs of manipulation, such as unnatural facial movements, inconsistent lighting, or audio-visual mismatches, that may indicate tampering. Furthermore, these initiatives should encourage healthy skepticism toward sensational or unverified content, fostering a culture of verification before sharing. By empowering citizens with these competencies, societies can build collective resilience against the deceptive use of deepfakes.

Collaborative Detection Platforms

Collaborative detection platforms represent another critical pillar in the fight against synthetic media fraud. Cross-industry collaboration can significantly accelerate both threat detection and response times. Open-source forensic frameworks, such as Deepware and FakeCatcher, can serve as shared foundations upon which researchers and practitioners build and refine detection tools. Equally important is the establishment of intelligence-sharing networks that connect academic institutions, online platforms, and cybersecurity firms. Such networks facilitate the rapid dissemination of information about emerging deepfake techniques and enable coordinated updates to detection systems, ensuring that defenses remain effective against evolving threats.

Ethical AI Development

Ethical AI development must be a guiding principle for both creators of generative models and designers of detection tools. Developers should prioritize fairness by rigorously testing models to ensure they do not disproportionately fail on specific demographic groups, which could introduce bias and exacerbate harm. Accountability mechanisms, including comprehensive audit trails for training data and model outputs, are essential for tracing errors and ensuring responsible use. Finally, transparency must be embedded into system design through explainable outputs that clarify how decisions are made, fostering user trust and enabling meaningful legal and regulatory oversight. Together, these ethical commitments help ensure that AI technologies serve the public interest while minimizing unintended consequences.

Deepfake Detection Challenges and Future Directions

Deepfake Detection Challenges

Figure 9 illustrates the multifaceted nature of deepfake detection challenges, emphasizing that these obstacles collectively compound the difficulty of identifying synthetic media. Technical limitations, such as the rapid evolution of generative models and the absence of universal forensic artifacts, are compounded by data-related issues including insufficient high-quality datasets and inadequate labeling practices. Legal and ethical complexities further complicate the landscape, while the susceptibility of detection systems to adversarial attacks adds an additional layer of vulnerability. Together, these interconnected hurdles underscore the formidable task of preventing deepfake proliferation. Addressing them requires a coordinated effort encompassing advances in AI architectures, development of more diverse and representative datasets, implementation of rigorous annotation standards, and deployment of resilient detection technologies capable of adapting to emerging threats.

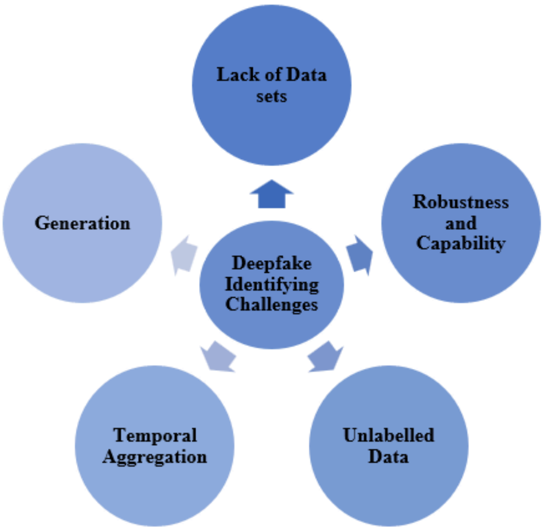

Fig. 9: Different Types of Challenges in Deepfake Detection

Lack of Datasets

Detecting deepfakes remains a formidable challenge, largely due to the absence of consensus on standardized image databases and the inherent difficulty of generating realistic synthetic fake images. While previous research has leveraged Generative Adversarial Networks (GANs) to create fictitious image datasets for training and evaluation, the authenticity of such generated samples often remains questionable, and no universally accepted benchmark dataset currently exists. The availability of publicly accessible, high-quality GAN-generated datasets containing realistic counterfeit photos is essential for enabling consistent and comparable evaluation of detection approaches, ultimately improving both the accuracy and reliability of forensic systems.

Robustness and Capability

To develop deepfake detectors capable of reliable operation in real-world scenarios, it is imperative to enhance their robustness against malicious attacks and their resilience to various modifications of fake content. This requires not only improving resistance to adversarial evasion but also ensuring that detection decisions are interpretable and transparent. To the best of the authors' knowledge, few existing studies have comprehensively evaluated their approaches from both of these critical perspectives, as highlighted by recent research. Consequently, significant attention must be directed toward achieving optimal performance and reliability in deepfake detection systems, balancing accuracy with explainability and robustness against evolving threats.

Unlabelled Data

Developing reliable deepfake detection models is particularly challenging when working with small or unlabeled datasets, a common constraint in domains such as law enforcement and media where data availability is limited. Most state-of-the-art detection systems rely on deep learning architectures trained on large-scale annotated datasets, yet their black-box nature poses significant difficulties for interpretability and trust. The inherently opaque decision-making processes of these models obscure the subtle cues that inform their judgments, making it harder to understand why certain detections succeed or fail. This lack of transparency undermines confidence in their utility for high-stakes applications and highlights the need for more explainable and data-efficient approaches tailored to real-world operational constraints.

Temporal Aggregation

Current deepfake detection algorithms often fail to account for interframe temporal consistency, leaving them vulnerable to temporal anomalies and the presence of mixed real and fake frames across video sequences. Additionally, many approaches require an extra computational step to compute a video quality score for each individual frame, adding complexity and reducing efficiency. These limitations underscore the urgent need for more sophisticated algorithms capable of accurately and consistently assessing temporal dynamics inherent in video data.

Generalization

No single universal feature can reliably identify all fake videos, prompting researchers to pursue more generalized detection approaches. However, a significant limitation of existing work is the reliance on relatively simple datasets such as FaceForensics++. To further enhance the reliability and accuracy of deepfake detection systems, future research must prioritize the use of more complex and diverse datasets that better reflect real-world conditions. Studying challenging and varied data will be essential for developing robust algorithms capable of generalizing across different manipulation techniques and maintaining high performance in practical applications.

Future Directions

As deepfake generation techniques grow increasingly sophisticated, the line between authentic and synthetic content continues to blur, making reliable detection ever more challenging. To keep pace with the rapid evolution of AI-generated media, future detection systems must adopt domain-agnostic models capable of generalizing across diverse and unseen scenarios. Techniques such as transfer learning and training on large-scale, heterogeneous datasets will be essential for building adaptable and robust detectors. Additionally, incorporating continual learning mechanisms will enable models to dynamically adapt to emerging generative architectures and novel manipulation strategies. Multimodal approaches that integrate visual, auditory, and textual cues offer particular promise, as they can uncover subtle inconsistencies, such as mismatches between lip movements and speech or incongruent facial expressions, that unimodal systems may overlook. By embracing these strategies, next-generation detection frameworks can better address the evolving threat landscape and maintain effectiveness in real-world conditions.

Future research will focus on strengthening the analysis of subtle visual signals, such as micro-expressions, eye blink patterns, and heart rate dynamics, across video clips of varying quality. The integration of explainable artificial intelligence (XAI) will be particularly valuable in sensitive domains like law enforcement and media, where transparency in detection decisions can foster greater trust and acceptance among users. Additionally, blockchain-based systems offer a promising avenue for enhancing detection pipelines by providing immutable verification of content origin and authenticity. Ultimately, ensuring that detection tools are deployed responsibly and with due regard for privacy requires the simultaneous development of robust ethical guidelines and regulatory frameworks. Adhering to these principles will enable more reliable and trustworthy identification of deepfakes while safeguarding individual rights.

Conclusion

The pervasive influence of digital technologies has fundamentally reshaped the creation and exchange of information, raising critical questions about the credibility and authenticity of online content. In this landscape, artificial intelligence, and deepfake technology in particular, has emerged as a powerful dual-use tool, offering both significant benefits and substantial risks. While deepfakes enable innovative applications across entertainment, education, and other fields, their potential for misuse poses serious ethical and security challenges that demand urgent attention.

This survey has provided a comprehensive overview of deepfake technology, systematically examining its various types, underlying generation techniques, available datasets, and real-world applications. In parallel, the review has explored the evolving landscape of detection methods and highlighted persistent gaps in generalizability, robustness, and multimodal integration. Beyond technical considerations, the paper has critically discussed the ethical, legal, and societal implications of deepfakes, underscoring the need for responsible innovation, informed policy, and international cooperation.

By synthesizing current knowledge and identifying directions for future research, this work aims to contribute to a growing research agenda focused on mitigating the threats posed by malicious deepfakes while harnessing their constructive potential. Continued advances in detection methodologies, coupled with robust governance frameworks and public awareness, will be essential to preserving trust in digital media and safeguarding the integrity of information in the years ahead.

Authors Contributions

All authors equally contributed to this article.

References

Ackley, D., Hinton, G., & Sejnowski, T. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, *9*(1), 147-169. https://doi.org/10.1016/s0364-0213(85)80012-4

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security*, 1-7. https://doi.org/10.1109/wifs.2018.8630761

Akhtar, Z. (2023). Deepfakes generation and detection: a short survey. *Journal of Imaging*, *9*(1), 18.

Al-Dhabi, Y., & Zhang, S. (2021). Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). *Proceeding of the IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, 236-241. https://doi.org/10.1109/csaiee54046.2021.9543264

Aslam, Y., & Santhi, N. (2019). A Review of Deep Learning Approaches for Image Analysis. *Proceding of the International Conference on Smart Systems and Inventive Technology*, 709-714. https://doi.org/10.1109/icssit46314.2019.8987922

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv*. https://doi.org/10.48550/arXiv.1409.0473

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Bansal, A., Ma, S., Ramanan, D., & Sheikh, Y. (2018). Recycle-GAN: Unsupervised Video Retargeting. *Computer Vision*, *11209*, 122-138. https://doi.org/10.1007/978-3-030-01228-1_8

Bao, G., Zhao, Y., Teng, Z., Yang, Linyi, & Zhang, Y. (2023). Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. *arXiv*. https://doi.org/10.48550/arXiv.2310.05130

Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020). CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis. *Proceeding of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-6. https://doi.org/10.1109/wifs49906.2020.9360905

Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, *12*, 69812-69837. https://doi.org/10.1109/access.2024.3397775

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166. https://doi.org/10.1109/72.279181

Bethu, S., Trupthi, M., Mandala, S. K., Karimunnisa, S., & Banu, A. (2024). AI-IoT Enabled Surveillance Security: DeepFake Detection and Person Re-Identification Strategies. *International Journal of Advanced Computer Science and Applications*, *15*(7). https://doi.org/10.14569/ijacsa.2024.0150799

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An Open-Source Autoregressive Language Model. *Proceedings of BigScience Episode #5 Workshop on Challenges & Perspectives in Creating Large Language Models*, 95-136. https://doi.org/10.18653/v1/2022.bigscience-1.9

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., & Filar, B. (2018). The Malicious Use of Artificial Intelligence. *arXiv*. https://doi.org/10.48550/arXiv.1802.07228

Cassia, M., Guarnera, L., Casu, M., Zangara, I., & Battiato, S. (2025). Deepfake forensic analysis: Source dataset attribution and legal implications of synthetic media manipulation. *ArXiv*. https://doi.org/10.48550/arXiv.2505.11110

Chang, X., Wu, J., Yang, T., & Feng, G. (2020). DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network. *Proceeding of the Chinese Control Conference (CCC)*, 7252-7256. https://doi.org/10.23919/ccc50068.2020.9189596

Chen, H., Takamura, H., & Nakayama, H. (2021). SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation. *Proceeding of the Association for Computational Linguistics: EMNLP 2021*, 1483-1492. https://doi.org/10.18653/v1/2021.findings-emnlp.128

Chesney, R., & Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*, 68. https://doi.org/10.2139/ssrn.3213954

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1800-1807). https://doi.org/10.1109/cvpr.2017.195

Ciftci, U. A., Demir, I., & Yin, L. (2020). How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. *International Joint Conference on Biometrics (IJCB)*, 1-10. https://doi.org/10.1109/ijcb48548.2020.9304909

Ciftci, U. A., Demir, I., & Yin, L. (2024). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1. https://doi.org/10.1109/tpami.2020.3009287

Dang, L. M., Hassan, S. I., Im, S., Lee, J., Lee, S., & Moon, H. (2018). Deep Learning Based Computer Generated Face Identification Using Convolutional Neural Network. *Applied Sciences*, *8*(12), 2610. https://doi.org/10.3390/app8122610

Dani, B., & Mustafa, K. (2025). *The next frontier of cybersecurity: Tackling deepfake threats and adversarial machine learning in smart cities and space with quantum technologies*.

Dehghani, A., & Saberi, H. (2025). Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools. *arXiv*. https://doi.org/10.48550/arXiv.2501.06227

Do, N.-T., Na, I.-S., & Kim, S.-H. (2018). Forensics face detection from gans using convolutional neural network. *2018 International Symposium on Information Technology Convergence (ISITC)*, 376-379.

Durall, R., Keuper, M., & Keuper, J. (2020). Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7887-7896. https://doi.org/10.1109/cvpr42600.2020.00791

Dvoskin, B. (2021). Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance. *SSRN Electronic Journal*, 63. https://doi.org/10.2139/ssrn.3986181

Fogelton, A., & Benesova, W. (2018). Eye blink completeness detection. *Computer Vision and Image Understanding*, *176-177*, 78-85. https://doi.org/10.1016/j.cviu.2018.09.006

Garg, A., Srivastava, G. S., & Masaki, K. (2025). *Beyond the benchmark: Generalization limits of deepfake detectors*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672-2680.

Graves, A. (2012). Long Short-Term Memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, *385*, 37-45. https://doi.org/10.1007/978-3-642-24797-2_4

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *06*(02), 107-116. https://doi.org/10.1142/s0218488598000094

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554

Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, *10*(1), 370. https://doi.org/10.3390/app10010370

Jbara, W. A., & Soud, J. H. (2024). DeepFake Detection Based VGG-16 Model. *Proceeding of the International Conference on Cyber Resilience*, 1-6. https://doi.org/10.1109/iccr61006.2024.10533024

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). *Progressive growing of gans for improved quality, stability, and variation*. https://doi.org/10.48550/arXiv.1710.10196

Khatri, A., & Gupta, N. (2023). A Study on Analyzing Deepfake through various Facial Regions- A Review. *Proceeding of the International Conference on Contemporary Computing and Informatics (IC3I)*, 1133-1138. https://doi.org/10.1109/ic3i59117.2023.10397658

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135-146. https://doi.org/10.1016/j.bushor.2019.11.006

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). T.: A watermark for large language models. *International Conference on Machine Learning*, 17061-17084.

Korshunov, P., & Marcel, S. (2018). DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *arXiv*. https://doi.org/10.48550/arXiv.1812.08685

Krishna, K., Song, Y., Karpinska, Marta, Wieting, John, & Iyyer, Mohit. (2024). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Lee, G., Lee, J., Jung, M., Lee, J., Hong, K., Jung, S., & Han, Y. (2025). Dual-Channel Deepfake Audio Detection: Leveraging Direct and Reverberant Waveforms. *IEEE Access*, *13*, 18040-18052. https://doi.org/10.1109/access.2025.3532775

Li, Y., & Lyu, S. (2018). *Exposing deepfake videos by detecting face warping artifacts*. https://doi.org/10.48550/arXiv.1811.00656

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. *Proceeding of the International Workshop on Information Forensics and Security (WIFS)*, 1-7. https://doi.org/10.1109/wifs.2018.8630787

Lima, O. de, Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). *Deepfake detection using spatiotemporal convolutional networks*. https://doi.org/10.48550/arXiv.2006.14749

Liu, F., Jiao, L., & Tang, X. (2019a). Task-Oriented GAN for PolSAR Image Classification and Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2707-2719. https://doi.org/10.1109/tnnls.2018.2885799

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019b). STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668-3677. https://doi.org/10.1109/cvpr.2019.00379

Lundberg, E., & Mozelius, P. (2025). The potential effects of deepfakes on news media and entertainment. *AI & SOCIETY*, *40*(4), 2159-2170. https://doi.org/10.1007/s00146-024-02072-1

Luttrell, J., Zhou, Z., Zhang, Y., Zhang, C., Gong, P., Yang, B., & Li, R. (2018). A deep transfer learning approach to fine-tuning facial recognition models. *Proceeding of the IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2671-2676. https://doi.org/10.1109/iciea.2018.8398162

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019). Do GANs Leave Artificial Fingerprints? *Proceeding of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 506-511. https://doi.org/10.1109/mipr.2019.00103

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2020). Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, *15*(1), 24-31. https://doi.org/10.1093/jiplp/jpz167

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. *ArXiv:1411.1784*.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions Don't Lie. *Proceedings of the 28th ACM International Conference on Multimedia*. MM'20: The 28th ACM International Conference on Multimedia. https://doi.org/10.1145/3394171.3413570

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*, *11*, 144497-144529. https://doi.org/10.1109/access.2023.3344653

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

Nnamdi, N., Oniyinde, O., & Abegunde, B. (2023). An appraisal of the implications of deep fakes: The need for urgent international legislations. *American Journal of Leadership and Governance*, *8*(1), 43-70.

Oord, A. van den, Dieleman, S., Zen, H., & Simonyan, K. (2016). Wavenet: A generative model for raw audio. *ArXiv Preprint*.

Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R. O. (2020). Deepfake Detection through Deep Learning. *Proceeding of the International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 134-143. https://doi.org/10.1109/bdcat50828.2020.00001

Paris, B. (2023). Seeing Through the Fog of War. *Re-Thinking Mediations of Post-Truth Politics and Trust*, 142-161.

Pfefferkorn, R. (2019). Deepfakes in the courtroom. *Boston University Public Interest Law Journal*, *29*, 245-274.

Ramluckan, T. (2024). Deepfakes: The Legal Implications. *International Conference on Cyber Warfare and Security*, *19*(1), 282-288. https://doi.org/10.34190/iccws.19.1.2099

Ray, A. (2021). Disinformation, Deepfakes and Democracies: The Need for Legislative Reform. *University of New South Wales Law Journal*, *44*(3), 983-1013. https://doi.org/10.53637/dels2700

Reynolds, D. (2009). *Gaussian Mixture Models*. 659-663. https://doi.org/10.1007/978-0-387-73003-5_196

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674-10685. https://doi.org/10.1109/cvpr52688.2022.01042

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceeding of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11. https://doi.org/10.1109/iccv.2019.00009

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning internal representations by error propagation, Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. https://doi.org/10.7551/mitpress/4943.003.0128

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, *3*(1), 80-87.

Schiff, K. J., Schiff, D. S., & Bueno, N. (2023). *The liar's dividend: The impact of deepfakes and fake news on trust in political discourse*.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681. https://doi.org/10.1109/78.650093

Shan, C., Gong, S., & McOwan, P. W. (2007). Beyond Facial Expressions: Learning Human Emotion from Body Gestures. *Procedings of the British Machine Vision Conference 2007*, 1-10. https://doi.org/10.5244/c.21.43

Soukupová, T., & Cech, J. (2016). Eye blink detection using facial landmarks. *Computer Vision Winter Workshop, Rimske Toplice, Slovenia*, 1-8.

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting Both Machine and Human Created Fake Face Images In the Wild. *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, 81-87. https://doi.org/10.1145/3267357.3267367

Tipper, S., Atlam, H. F., & Lallie, H. S. (2024). An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection. *Applied Sciences*, *14*(21), 9754. https://doi.org/10.3390/app14219754

Tiwari, I. (2024). The legal implications of deepfake technology in the entertainment industry. *Dharmashastra National Law University Law Review*.

Usukhbayar, B., & Homer. (2020). *S.: Deepfake videos: The future of entertainment*.

Verma, N. (2023). *Deepfake technology and the future of public trust in video*.

Wang, M., Guo, L., & Chen, W.-Y. (2017). Blink detection using Adaboost and contour circle for fatigue recognition. *Computers & Electrical Engineering*, *58*, 502-512. https://doi.org/10.1016/j.compeleceng.2016.09.008

Wang, Y., & Dantcheva, A. (2020). A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. *Proceeding of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 515-519. https://doi.org/10.1109/fg47880.2020.00089

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., & Li, H. (2023). DIRE for Diffusion-Generated Image Detection. *Proceeding of the International Conference on Computer Vision (ICCV)*, 22388-22398. https://doi.org/10.1109/iccv51070.2023.02051

Wiseman, S., Shieber, S. M., & Rush, A. M. (2017). A.M.: Challenges in data-to-document generation. *ArXiv Preprint*.

Wu, X., Xie, Z., Gao, Y., & Xiao, Y. (2020). SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2952-2956. https://doi.org/10.1109/icassp40776.2020.9053969

Yu, H., Tan, Z.-H., Ma, Z., Martin, R., & Guo, J. (2018). Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4633-4644. https://doi.org/10.1109/tnnls.2017.2771947

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3813-3824. https://doi.org/10.1109/iccv51070.2023.00355

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-Stream Neural Networks for Tampered Face Detection. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1831-1839. https://doi.org/10.1109/cvprw.2017.229

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *International Conference on Computer Vision*, 2242-2251. https://doi.org/10.1109/iccv.2017.244